THE CONSTRUCTION AND EVALUATION OF A DYNAMIC COMPUTERISED ADAPTIVE TEST FOR THE MEASUREMENT OF LEARNING POTENTIAL

by

MARIÉ DE BEER

submitted in accordance with the requirements of
the degree

DOCTOR OF LITERATURE AND PHILOSOPHY

in the subject

PSYCHOLOGY

at the

UNIVERSITY OF SOUTH AFRICA

PROMOTOR: PROF C PLUG

MARCH 2000

DECLARATION

Student number: 463-339-3

I, the undersigned, hereby declare that the thesis titled “The construction and evaluation of a dynamic computerised adaptive test for the measurement of learning potential” is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

_____ _____

Marié de Beer Date

ACKNOWLEDGEMENT

Having been blessed in many ways, I am furthermore grateful in particular to the following individuals and institutions who contributed to make this project possible.

- Prof C Plug, my promotor who guided me with feedback to help clarify both my thinking and my writing. I consider it a privilege to have worked with him and, at the time of his retirement, to have been one of the last students to complete a doctoral thesis under his expert guidance. It has been a learning and developmental experience - both professionally and personally.
- The Human Sciences Research Council (HSRC), for logistic and administrative support during the early phases of the project. In particular, I would like to thank Dr Nicolaas Claassen, in whose division I worked, for his inspiring and enduring enthusiasm for research. A special thank you to Mrs Martie van Gass for her administrative and fieldwork help. The views reflected here are my own and do not necessarily reflect the views of the HSRC.
- The University of South Africa, for financial assistance towards some of the fieldwork in the form of a research award. Colleagues and friends at Unisa for their continued encouragement and interest.
- The National Research Foundation (NRF) (previously the Centre for Science Development (CSD)) for financial assistance towards this research. Opinions expressed and conclusions arrived at, are my own and should not necessarily be attributed to the NRF.
- All institutions and individuals who were involved with the fieldwork for this project and in particular those who participated in the testing.
- Cas Coetzee whom I consulted about statistical analysis and interpretation.
- Moya Joubert for editing and proofreading.
- Friends, family and loved ones - the greatest blessing of all! You never stopped encouraging, believing in or supporting me despite neglect in limited time available for you. A simple thankyou here can never express my gratitude and appreciation - I can but try to live it henceforth.
- In grateful memory to Ma A for fostering enjoyment of life and of work.

** ∞ **

“ … it was my master who taught me not only how very little I knew but also that any wisdom to which I might aspire could consist only in realising more fully the infinity of my ignorance” (Schilpp, 1974, p 3)

** ∞ **

TABLE OF CONTENTS

CHAPTER 3

THE MEASUREMENT OF LEARNING POTENTIAL

CHAPTER 5
CONSTRUCTION OF THE LEARNING POTENTIAL COMPUTERISED ADAPTIVE TEST (LPCAT)

CHAPTER 7

VALIDITY RESULTS FOR THE LPCAT

LIST OF TABLES

LIST OF FIGURES

SUMMARY

THE CONSTRUCTION AND EVALUATION OF A DYNAMIC, COMPUTERISED ADAPTIVE TEST FOR THE MEASUREMENT OF LEARNING POTENTIAL

by M de Beer

Degree: Doctor of Literatrure and Philosophy
Subject: Psychology
Promotor: Prof C. Plug

Recent political and social changes in South Africa have created the need for culture-fair tests for cross-cultural measurement of cognitive ability. This need has been highlighted by the professional, legal and research communities. For cognitive assessment, dynamic assessment is more equitable because it involves a test-train-retest procedure, which shows what performance levels individuals are able to attain when relevant training is provided. Following Binet's thinking, dynamic assessment aims to identify those individuals who are likely to benefit from additional training. The theoretical basis for learning potential assessment is Vygotsky's concept of the zone of proximal development.

This thesis describes the development, standardisation and evaluation of the Learning Potential Computerised Adaptive Test (LPCAT), for measuring learning potential in the culturally diverse South African population by means of nonverbal figural items. In accordance with Vygotsky's view, learning potential is defined as a combination of present performance and the extent to which performance is increased after relevant training. This definition allows for comparison of individuals at different levels of initial performance and with different measures of improvement. Computerised adaptive testing based on item response theory, as used in the LPCAT, is uniquely suitable for increasing both measurement accuracy and testing efficiency of dynamic testing, two aspects that have been identified as problematic. The LPCAT pretest and the post-test are two separate adaptive tests, hence eliminating the role of memory in post-test performance. Several multicultural groups were used for item analysis and

test validation. The results support the LPCAT as a culture-fair measure of learning potential in the nonverbal general reasoning domain. For examinees with a wide range of ability levels, LPCAT scores correlate strongly with academic performance. For African examinees, poor proficiency in English (the language of teaching) hampers academic performance. The LPCAT ensures the equitable measurement of learning potential, independent of language proficiency and prior scholastic learning and can be used to help select candidates for further training or developmental opportunities.

Key terms:
Learning potential; dynamic testing; computerised adaptive testing (CAT); Item response theory (IRT); psychometric testing; psychometric test construction; Vygotsky; Zone of proximal development

CHAPTER 1

BACKGROUND

> "... You cannot take a person who for years has been hobbled by chains, bring him up to the starting line of a race and say - you are free to compete with us - and truly believe that you are treating him fairly."
>
> Lyndon Johnson

1.1 INTRODUCTION

The above quotation could be applied in many settings, but has special significance in the present South African context. Recent political changes have opened up opportunities and competition in many areas in unprecedented ways, but we should acknowledge that at present not everyone approaches the starting line of the “race” with equal or comparable preparation to ensure fair competition.

The 1990s heralded a period of transformation and change in South Africa, with the first democratic election of 1994 as a prominent political symbol. In most areas, however, this event was only a prelude to many phases of change - some of which have already been dealt with, while others are ongoing. An ancient Chinese curse is said to be implied in wishing someone to "live in interesting times". While events in South Africa in the last decade of the 20th century and in the transition to the 21st century have certainly constituted "interesting times", they need not necessarily be regarded as a curse. In fact, in many fields of scientific, business and social endeavour, the adjustments that have been necessary to cope with the many changes in our society, have brought exciting new challenges and opportunities for development. One area that requires new development is psychometric testing. In a policy document (18/9/B) approved by the South African Professional Board for Psychology in November 1998, the need for psychometric tests that have been designed and standardised for all South Africans is emphasised. In the same

document it is noted that few empirical studies have been undertaken to investigate test bias, validity and cultural appropriateness of measures. The need for such studies in South Africa is evident.

Working towards providing equal opportunities and redressing past imbalances by means such as affirmative action policies, has placed a specific focus on training and development. At the same time, it is necessary to take cognisance of the differences in socioeconomic circumstances that still exist between cultural groups in our country, as well as differences in educational standards and access to training opportunities. The scarcity of resources available for training and development, and provision of opportunities to those who have been most disadvantaged, must somehow be aligned without placing the standards or success rate of the training and development opportunities that are provided in jeopardy.

Effective placement of people in educational, training or work positions often means that some are selected, while others are not. In the past, psychometric testing has been considered useful in this regard because the results provide a scientific, objective measure of certain characteristics of individuals. Because resources are scarce and selection and placement are costly, effective and successful outcomes are important in human as well as financial terms. However, much criticism has been levelled against psychometric tests because many of them do not allow for diversity among candidates. In the 1998 policy document of the South African Professional Board for Psychology on the classification of psychometric instruments, an urgent appeal is made for psychologists to address the need for the development and adaptation of culturally appropriate measures. In South Africa the large differences in socioeconomic and educational background with which people come to the assessment situation should be taken into account in the development and use of cognitive ability tests in particular. There should therefore be a change in emphasis from measuring crystallised competencies that are largely the result of educational opportunities, towards the measurement of undeveloped potential, which will allow for redressing of past imbalances.

The Russian psychologist, Lev Vygotsky (1978), first used the term “zone of proximal

development" (ZPD), to indicate the difference between the level of achievement an individual can attain without help and the level of achievement attainable with help. By focusing not only on what the person is presently capable of, but also allowing for future development, allowances can be made for the differences in educational background that individuals bring to the testing situation. In this way potential future performance can also be evaluated, instead of considering only present performance. A focus on learning potential using the dynamic testing approach where training is included in the test administration, will allow for more equitable testing across different cultural and socioeconomic groups.

In terms of psychometric test theory, the development of item response theory (IRT), has brought about improved testing technology in the form of computerised adaptive testing (CAT) which contributes to more effective psychometric measurement. By combining learning potential measurement, the dynamic testing approach and computerised adaptive testing based on IRT, a psychometric instrument that

- makes use of the improved IRT statistical procedures for test development
- contains training as part of the assessment procedure to take diversity in educational backgrounds into account
- focuses on the measurement of learning potential

was developed. Its development, standardisation and evaluation form the core of the present project.

1.2 THE RESEARCH PROBLEM IN HISTORICAL PERSPECTIVE

1.2.1 Psychological testing

Measurement of intelligence was an important focus in psychological testing in the 20th century. Earlier developments included Wundt's laboratory at the end of the 19th century where biopsychological measures were first used in an attempt to distinguish between people. The French psychologist, Binet, and his colleague, Simon, were the forerunners in the development of intelligence tests as we know them today, using tasks that required cognitive reasoning to identify pupils in need of special

education (Binet & Simon, 1905/1916). Although they emphasised development, rather than classification as the focus of their test, later translations and adaptations of their test became widely used in different contexts, leading to the present-day types and uses of intelligence tests. Binet's visionary contribution to the psychometric testing of intelligence is of particular significance to the present project and will be highlighted in further chapters.

In South Africa, the Stanford-Binet, the Binet version that was translated, adapted and standardised by Terman of Stanford University (Terman,1916; Terman & Merrill, 1937), was also used (Claassen, 1997). Cognitive tests developed specifically for South African use were mostly based on tests that were internationally used, such as the Stanford-Binet and the Wechsler-Bellevue. The bulk of intelligence tests currently in use, still resemble the first tests developed in the early 1900s. Although there have been new developments in theories of intelligence in recent years, such as the multiple intelligences theory of Gardner (1983) and Sternberg's triarchic theory (Sternberg, 1985), standardised, commercially available instruments based on these theories are not yet available, although some work has been done with research instruments (Sternberg, 1997a).

In recent years, Vygotsky's (1978) theory of the ZPD has resulted in a different approach to the measurement of cognitive functioning. This has led to the development of dynamic testing using a test-train-test approach for the measurement of learning potential. Learning potential assessment emphasises development and allows for improvement in cognitive performance if relevant training is provided. Feuerstein (Feuerstein, 1979; Feuerstein, Rand, Jensen, Kaniel & Tzuriel, 1987) is the founder of the instrumental enrichment approach to learning potential where the aim and focus are on remediation and changing the level of functioning of the individual concerned. Other dynamic testing approaches have also been developed, such as the test-centred coaching approach, graduated prompting methods, a psychometrically oriented learning test approach, and a testing-the-limits approach. However, although dynamic testing has shown advantages compared with traditional static testing, in an extensive review of the field, Grigorenko and Sternberg (1998) concluded that more research involving larger populations and especially more

validity information using educational or professional criteria is needed to establish the contribution of dynamic testing to the psychological and testing communities.

In South Africa, there was some early emphasis on adaptability, and learning in testing could be seen in, for instance, the General Adaptability Battery (GAB), which was developed in the late 1940s and early 1950s (Biesheuvel, 1952; Claassen, 1997). Biesheuvel (1972b) emphasised learning during the test session for the General Adaptability Battery in an attempt to identify the occupational suitability of large numbers of blacks who had received very little, if any, formal education. More research focusing on dynamic testing and learning potential followed in time. Some experimental instruments for the measurement of learning potential were constructed solely for research purposes (Boeyens, 1989a, 1989b, 1989c), while commercial learning potential instruments were also developed (Taylor, 1994a, 1994b). Unfortunately, although these latter instruments are used in industry, research publications on their results could not be found. Van Niekerk (1991) used the Feuerstein instrumental enrichment programme in a cross-cultural study, but found few significant results. Some researchers made use of existing standard cognitive tests administered dynamically in the characteristic test-train-test manner typically associated with dynamic learning potential assessment (Shochet, 1992, 1994; Zolezzi, 1995).

The history of psychometric tests in South Africa has largely followed international trends but also reflects the sociopolitical history of our country (Claassen, 1997). Initially, separate tests were developed for the separate population groups. In time, this was followed by attempts to construct tests that could be used for people from different cultural groups but who share English or Afrikaans as first language (Claassen, De Beer, Hugo & Meyer, 1991). These attempts at constructing cross-cultural instruments, however, still excluded the largest percentage - approximately 76 percent - of the population made up of African language speakers. Given the recent changes in South Africa and the increasing integration in schools, universities, the workplace and society in general, there is an urgent need for culture-fair instruments that can be used for all our cultural and language groups. A dynamic computerised adaptive test developed specifically for South Africa's

multicultural context will address the need for the psychometrically sound, untimed yet time-efficient measurement of learning potential by means of a test specifically designed for that purpose.

1.2.2 Problems to consider in test construction in South Africa

While the need for the construction of culture-fair and unbiased instruments can be easily understood, the operationalisation and practical implementation of such an endeavour is hampered by many problems. Firstly, the fact that we have 11 official languages has to be taken into account. When a person who is not fluent in a language is tested in that language, the resulting score may well be more an indication of the person's language proficiency, than of the skill that is supposedly measured. The numerous difficulties and biases that can result from cultural, language and dialect differences lead to a preference for instruments with nonverbal, figural content for testing cognitive ability cross-culturally. Although, at a practical level, such a nonverbal instrument should not be used in isolation since it cannot reflect all characteristics of importance, it can nevertheless make a useful contribution as a culture-fair measure of general cognitive ability.

Secondly, there are still major socioeconomic and educational differences between cultural groups in South Africa. There are also differences in the educational standards of schools, and many schools have to cope with poorly educated teachers and/or a lack of teaching materials and resources. A result of these differences is that many people from disadvantaged backgrounds have not had the opportunities to develop their potential fully. Many of these individuals score poorly when assessed with standard psychometric instruments, but these poor scores largely reflect their lack of educational opportunities and not necessarily a lack of potential. Focusing on the measurement of learning potential, or potential for future achievement over and above current levels of achievement, would help to address this problem.

A number of tertiary training institutions have attempted to address the problem they face with students from diverse educational environments applying for entry. At the University of Natal, a test-teach-test (TTT) alternative selection route was initiated to

identify students who have the potential academic ability for successful degree study, even though this potential may not be reflected in the individual's matric results. Instead of testing students on their knowledge and skills, the focus is on the students' ability to engage with new materials and problem-solving tasks typical of degree study. The TTT Programme is intended to prepare students for the University's entrance examination and is mainly targeted at disadvantaged students (Miller, 1992).

At the University of the North, the UNIFY Selection Research Project, a science foundation year between the student's matric (grade 12) year and first year at the University of the North (Unin) is aimed at students who were eductionally disadvantaged but have the potential for university study in mathematics and science. The aim of the project is to improve the numbers and quality of students who enrol for mathematics and science subjects at Unin (Zaaiman, 1995). The ML Sultan Technikon has also expressed a need to investigate the use of a battery of tests for the selection of disadvantaged students for engineering and other science and technology courses in particular. In a study that investigated possible selection criteria, it was found that the level of English proficiency is crucial for academic performance, even in science-related subjects (De Beer & Van Eeden, 1997). Results indicated that performance at school is the best predictor of average first-year performance. However, using only these results would perpetuate the existing situation, and not allow disadvantaged individuals to show their potential. Providing training opportunities for those whose initial low performance improves when relevant training is provided can benefit such disadvantaged students. While they may not meet selection criteria on the basis of their present performance only, when their learning potential is taken into account, they may be provided with opportunities to develop to their full potential.

Van Eeden (1993) also underscores the importance of language proficiency, and found that results of a standardised cognitive ability test administered in a language in which candidates were reasonably proficient, but which was not their mother tongue, were less valid than the results for participants who were tested in their mother tongue. This indicates that when similar standard tests are used for selection purposes, achievement may be underestimated for those not tested in their mother tongue.

Since it has been reported that language also affects academic performance, this points to the need for language proficiency training in educational settings where students or pupils receive training in a language of tuition other than their first language.

These factors emphasise the need to look beyond what has already been achieved, such as present academic performance, present language proficiency, or present performance in standard tests, in the selection of people for training and development. Potential future achievement that could be attained if limiting factors such as a poor educational environment or lack of language proficiency could be addressed, also need to be evaluated and taken into consideration. One way of doing this is to measure learning potential by combining nonverbal, figural content with a test-train-test dynamic testing approach, where some training is incorporated in the test itself. The training involves providing hints and strategies that are helpful in answering the questions by indicating ways and means whereby the correct answer to the questions can be determined. The initial test, also known as the pretest, indicates the current level of performance without help. After training, the person is retested with a post-test. This second score is taken as an indication of the potential future level of performance. The difference score reflects the extent to which present performance may improve after receiving relevant training.

A third problem that needs to be addressed is the efficiency of test procedures, with a specific focus on test construction, bias analysis - also referred to as differential item functioning or DIF - test administration and ease of obtaining the results. IRT has advantages over classical test theory (CTT) on all the factors mentioned. In addition, IRT, together with the availability of computer technology, has made CAT possible. These developments improve both test construction and test administration to produce an extremely time-effective and efficient CAT product. In CAT, items are calibrated beforehand and are stored in an item bank from where suitable items are interactively selected during the testing process to match the examinee's estimated ability level. Because the items are selected commensurate with the examinee's estimated ability level, fewer items are necessary to achieve the same accuracy of measurement as in much longer conventional tests (Weiss, 1983a).

Furthermore, although different items are administered to different individuals, the scale of measurement is the same for all, and all scores can be directly compared.

Sijtsma (1993a, 1993b) suggested that the measurement of change, which is a key concept in dynamic testing and the measurement of learning potential, can be addressed by using IRT models and CAT. These recent developments provide a viable psychometric foundation for assessing learning potential by means of dynamic testing. To date, only one computerised adaptive cognitive ability test has been developed in South Africa (De Beer, 1991; Van Tonder & Claassen, 1992).

1.2.3 Future trends and possible solutions

Although in recent years negative sentiments have been expressed about psychometric testing and the perceived role it played in South Africa's apartheid history, there is still a need for scientific, culture-fair and unbiased instruments that can be used in multicultural settings in schools, tertiary educational institutions and industry. At some level, the choices that have to be made about the use or nonuse of psychometric instruments will reflect social and political values. Decisions about the desirability of test use cannot be made on the basis of psychometric research alone, since it also depends on ideological and sociopolitical value judgments (Goldstein, 1989; Visser, 1996). Test development is a lengthy process, and for cross-cultural applications it is further complicated by various cultural, political and socioeconomic factors (Claassen, 1997). Tests that can be used for all cultural groups without discriminating against any person or group, are needed in education and industry, both of which have become so integrated that the use of separate tests for separate cultural or language groups is generally not a viable alternative. As mentioned before, the Psychometrics Committee of the South African Professional Board for Psychology has made an urgent appeal for the development of such culturally appropriate measures.

Psychometric testing received special mention in the new Employment Equity Act, which was tabled late in 1998 (Employment Equity Act, 1998). The Act states that "Psychological testing and other similar assessments of an employee are prohibited

unless the test or assessment being used:

(a) has been scientifically shown to be valid and reliable;

(b) can be applied fairly to all employees; and

(c) is not biased against any employee or group".

Although the popular press initially interpreted and reported this section of the Act as a blanket ban on all psychological tests, careful reading indicates that it reflects and underscores sound psychometric principles. The main thrust of the Act is its affirmative action component with the aim of deracialising South Africa's labour market by applying the principles of equity to advance previously disadvantaged groups. Implementation of the Act should help decrease racial discrimination (Vapi, 1998).

A test for the measurement of cognitive ability, which takes the ever-changing socioeconomic and educational diversity of our multicultural population into account, will provide much-needed information. A research project involving the construction, standardisation and evaluation of a computerised adaptive test that uses the dynamic test-train-test approach to measure learning potential was conceptualised to address this need. The construction of such an instrument would address several urgent questions in psychometric testing in South Africa, namely:

- Will a measure that includes some training benefit examinees by allowing them to improve on their initial performance?
- Will a measure of learning potential be more culture-fair than standard (static) tests of intelligence and indicate smaller differences in mean scores between the cultural groups?
- Will the distribution of scores of the culture groups indicate learning potential measures to be more equitable measures of general reasoning?
- Will the measurement of learning potential provide a better indication of future academic or other training performance than standard tests, especially for disadvantaged examinees?

1.3 AIMS OF THE STUDY

The overall aim of this research project is to construct, standardise and evaluate a computerised adaptive test for the measurement of learning potential that makes use

of the dynamic testing strategy and is aimed at a target population of people from all culture groups in South Africa with at least five years of education. In attempting this, particular attention will be given to DIF analysis to improve cross-cultural acceptability of the test.

The following specific aims will be addressed in this study:

- to evaluate differential item functioning (DIF) between language and culture groups as well as between the two gender groups for item selection
- to evaluate the reliability and validity of the instrument according to the American Psychological Association (APA) standards for psychometric test development (APA, 1985)
- to assess the predictive validity of the instrument by using academic and other relevant results, so that its practical utility can be evaluated
- to compare the results obtained with this instrument with those of conventional psychometric instruments, specifically to evaluate its construct validity
- to assess the usefulness of this instrument in cross-cultural settings, with specific reference to its predictive validity compared with other standard tests and previous academic performance

1.4 DIVISION OF CHAPTERS

Chapter 2 provides a historical overview of psychological testing of intelligence. The history of the measurement of intelligence is discussed with specific reference to Binet's contribution. Important theoretical and practical issues are also highlighted.

In chapter 3, the theory and measurement of learning potential by means of dynamic testing are discussed. The focus is on Vygotsky's theory and its relevance to a dynamic computerised adaptive test of learning potential. The development of international and national measuring instruments that reflect this theoretical approach is reviewed.

Chapter 4 covers computerised adaptive testing based on IRT. Theoretical principles and the practical benefits of CAT procedures are dealt with and the combination of the dynamic, learning potential approach with the CAT strategy based on IRT is explained and justified. The construction of the specific instrument on which this research is based is discussed in chapter 5.

In chapter 6 the procedure for the evaluation of the test is set out with detailed information on the samples, the measures used and the test procedures followed.

In chapter 7 the results obtained with the Learning Potential Computerised Adaptive Test (LPCAT) and other measures, are reported and analysed.

In the last chapter, chapter 8, the results are discussed. Special attention is paid to the aims of the project to see whether they were met. The limitations and shortcomings of the research are discussed and recommendations made for further research. The chapter concludes with an overview of the findings of the research and the possible contribution thereof to the theory, practice and methods of the psychometric testing of learning potential.

CHAPTER 2

MEASUREMENT OF INTELLIGENCE

2.1 INTRODUCTION

The psychometric measurement of intelligence is nearing its centenary, but in many ways the tests used today still resemble the first intelligence test constructed in the early 1900s. Although this has led people such as Sternberg (1997a) to consider the contemporary cognitive testing industry to be a glaring exception to the rapid general rate of technological development, it will be shown that some of the ideas of the early theorists and practitioners can only now, with the availability of recent theoretical, psychometric and technological developments, be fully utilised. As in many other sciences, early ideas often provide the base for further research and development many years later. One such example in psychology concerning the measurement of intelligence is the work of Alfred Binet, who, together with his colleague, Theodore Simon[1], developed the first psychometric intelligence test in 1905.

2.1.1 Binet's legacy

The Binet-Simon test (Binet & Simon, 1915) was an international breakthrough and Binet became the founder figure in intelligence testing. However, the fulfilment of the promise of Binet's early work on the measurement of intelligence has only now been made possible by

- the development of item response theory (IRT)

[1] For the sake of convenience, reference will be made to only Binet throughout this text, although it is acknowledged that Simon was also involved in most of the research discussed.

- the availability of computer technology to run interactive computerised adaptive tests (CATs) based on IRT
- the use of Vygotsky's theory of the zone of proximal development (ZPD) as theoretical base
- recent trends in cognitive ability assessment to develop dynamic tests with a view to measuring learning potential

Binet's first attempts to measure intelligence psychometrically, were aimed at distinguishing, in a group of retarded school children, those who seemed likely to benefit from further instruction or training from those who would probably not. This initial 1905 test, the first of its kind, evolved into the measurement of intelligence of normal children with the first revision of the test in 1908. The Binet-Simon test formed the basis of most intelligence tests developed since (Binet & Simon, 1915; Wolf, 1973).

Although standard intelligence tests have been widely used in the selection and placement of people and for prediction of academic and work performance for many years, they have not been without problems. For instance, they are not always suitable for cross-cultural testing, and language proficiency as well as socioeconomic and educational disadvantage affect test results. Considering the difficulty involved in constructing cognitive ability tests that are culture fair, an approach that combines the features mentioned above can contribute to the solution of many of the problems encountered today in the cross-cultural testing of cognitive ability.

2.1.2 Definition of terms

Confusion between the concepts of intelligence, cognitive ability and tested intelligence have contributed to the debate on the nature of intelligence. What constitutes intelligent behaviour may differ from one context to another (Sternberg, 1997b). The following terms are defined to clarify their use for this particular research project:

- **Intelligence**

  Intelligence refers to the construct that is measured by standardised psychological tests that provide numerical values to summarise present performance. Psychometric intelligence refers - somewhat narrowly - to that which is measured by intelligence tests. It is usually associated with numerical value(s), and is generally seen to represent a static, unchangeable measure of intelligence.

In layperson's terms, the word "intelligence" is often used to refer to a more general ability which can manifest itself in different ways. What is considered to be indicative of intelligence as defined in this general way, may differ across cultures and contexts and is probably too broadly defined to be empirically or scientifically useful. This view is more descriptive of the term "cognitive ability", as defined here.

- **Cognitive ability**

  Cognitive ability is a more broadly defined term that is not restricted to scores on psychometric instruments but can be used to refer to other manifestations of adaptive behaviour or cognitive performance in broader terms.

Measures of intelligence (psychometric test performance) are used to predict individual differences in cognitive ability, such as effective functioning and adaptive contextual behaviour. In this way intelligence is related to real-world intelligent performance or cognitive ability.

- **Learning potential**

  Learning potential refers to an overall cognitive capacity and includes both present and improved future performance. Implied in the use of the term is the assumption that intelligence - that which is measured with psychometric tests - is changeable, as indicated by changes in intelligence scores obtained with standard tests.

Up to now, the measurement of learning potential has mostly been dealt with by

standard intelligence tests applied in a dynamic way with training simulated in a test-train-retest approach. Present performance (pretest score) and potential future performance (post-test score) after relevant training, are measured. Improvement in test performance which reflects the ZPD is taken as the difference between post-test and pretest scores. This ZPD or difference score is used together with the pretest score to assess learning potential.

2.1.3 Subjectivity and limited measurement accuracy of social science research

Social science, which by its very nature, involves humans studying the behaviour and characteristics of their own kind, is unlikely to be objective. The particular paradigmatical perspective as well as individual context is likely to influence most aspects of human scientific endeavour. From conception of the research problem, through the design of the study and measurement to the interpretation of the research results, subjective interest and expectation can affect the research process and consequent findings. Value-neutral social research is not possible because social research is conceived by humans who are unable to disengage themselves from their own context and are the products of their background, training and social position - all of which will influence the research that is conducted and the way in which the results are interpreted (Hubbard, 1996). The implications of subjectivity in the field of cognitive assessment, where issues such as social standing and societal rewards are at stake, are even more far reaching.

In addition to the possible influence of subjectivity in social science research, the accuracy of measures obtained is not comparable to that found in the physical sciences. There are two main reasons for this. Firstly, the psychological constructs that are measured are not directly visible or measurable and have to be measured indirectly. Consequently the measures obtained cannot be as accurate as those attained with direct measurement. Secondly, the constructs of interest change over time and are subject to the

influence of a myriad of other psychological factors such as emotions, motivation and concentration, and can therefore even change in one individual over a short time span.

According to Cziko (1989), the behavioural sciences enthusiastically adopted the Newtonian model of physics towards the end of the 19th century. According to this model, it is assumed that all relevant variables can be measured objectively and that all physical events are determined completely by - and are therefore predictable by knowledge of - preceding events. However, "while the physical sciences have discarded this view of the physical universe as a giant, predetermined clock, this perspective still dominates mainstream 'scientific' educational (and psychological) research" (Cziko, 1989, p 18). This also raises the issue of consciousness and free will, both of which influence human behaviour in unpredictable ways. Returning to the theorists of the physical sciences, Albert Einstein had a strong belief in a deterministic universe and was convinced that "any 'randomness' observed in quantum phenomena was due to 'hidden variables', which, if discovered and understood, would in principle allow for perfect prediction of all physical phenomena. ... In contrast, Bohr, the leading figure of the Copenhagen School of quantum mechanics, maintained that the uncertainties and probabilities observed in quantum phenomena are intrinsic to the phenomena themselves and not the result of incomplete knowledge" (Cziko, 1989, p 23). It seems much more appropriate for researchers in the human sciences to adopt the latter view.

Human subjects are complex, and it is generally impossible to predict human behaviour perfectly. One human characteristic that has been the focus of intensive scientific investigation and debate is intelligence. Attempts to define, understand, measure and predict intelligence go back a long way in history.

2.2 THE HISTORY OF INTELLIGENCE AND ITS MEASUREMENT

2.2.1 Introduction

A review of the history of any scientific field often brings new insight that contributes to the more general understanding of important elements. Intelligence has been the subject of interest and philosophical debate for many centuries. According to Richardson and Bynner (1984), philosophers from Aristotle to Ockham - who is regarded as the last philosopher of the Middle Ages - have discussed, defined and debated the concept of intelligence. The period following the Middle Ages, was marked by a trend towards individualism. During this phase, the strength model of intelligence, which is characterised by concepts such as power, capacity, level, energy or other euphemisms of strength, and which is also marked by competitive individualism, was the focus of attention. It was used to serve important sociopolitical functions since the hierarchy of cognitive abilities that results from this view, reflects the hierarchical nature of society itself (Richardson & Bynner, 1984).

Early measures of humans by the experimental psychologists involved mainly physical (biological) measures which focused on reaction time and other sensory measures. These measures of perception and discrimination were used as indices of intelligence and still receive support in present-day reaction-time studies (Eysenck, 1994). The history of reaction-time measures dates back to 1796 in the field of astronomy where assistants had to record the passage of stars across the meridian (Garret, 1941). An assistant named Kinnebrook lost his job at the Royal Observatory in England, because his slow reaction time caused too large an error in his recording of star movements (Garret, 1941; Gregory, 1996). In 1822, the German astronomer, Bessel, was the first to compare the results of such reaction-time measures, showing large and enduring individual differences in recording the transit of stars (Garrett, 1941).

Wundt's laboratory (founded in 1879), where thousands of biological and physical measures were taken, was the first example of large-scale testing (Gregory, 1996; Thorndike & Lohman, 1990). The emphasis on physiological

measures in attempts to measure human intelligence continued until the early 1900s. Binet, who developed the first psychometric intelligence test, also initially used physiological and biological measures in his first attempts to develop a test for the measurement of intelligence (Wolf, 1973). Although most people these days associate the measurement of intelligence with psychometric tests similar to the one eventually developed by Binet, physiological measures of intelligence and reaction-time measures are also still used today, although mostly experimentally.

In South Africa, the measurement of intelligence closely followed international trends. This includes aspects of test construction, the use of test results and the technical issues of bias and fairness. Many overseas tests were also adapted and standardised for South African use. In South Africa, the use of tests and test results cannot be separated from the country's history, especially with the racial divisions that were part of the social and work environments. European tests were adapted for use in South Africa, and the suitability of tests for different cultural groups was raised at an early date (Claassen, 1997). Differences between cultural groups were found, but with societal changes, these differences also changed over time and will continue to do so for the foreseeable future. During the years of racial segregation, separate tests were developed for the different cultural groups. Bias and fairness were not such urgent issues, because the target groups were extremely homogeneous. However, the integration of different groups in a multicultural society has meant that existing separate tests have become less appropriate. The present need is for instruments that can be used fairly for all cultural groups in the South African population.

2.2.2 The nature of intelligence and the nature-nurture debate

The early philosophers, Plato and Aristotle, drew a distinction between

cognitive and hormic aspects of human behaviour. The first refers to factors such as thinking, problem solving, meditating and reasoning, while the latter concerns emotions, feelings, passions and the will. Cicero is credited for having coined the term “intelligence”, which is still used today to refer to a person’s cognitive powers and intellectual abilities (Eysenck & Kamin, 1981).

Possibly because, as Linn (1989, p 1) puts it, “intelligence is both a scientific and a folk concept”, even laypersons with little understanding of psychometric principles often do not hesitate to voice their opinions on issues of intelligence and its measurement. There are many differences within and between different cultures regarding physical features such as height, hair colour and texture, eye colour, et cetera. These differences are accepted without controversy, possibly because they are easy to verify empirically by observation. Furthermore, these differences mostly do not affect one’s “standing” in society or chances of “success”. Whereas both “more of” or “less of” certain personality characteristics such as dominance can be seen as a positive attribute, in terms of intelligence this is not the case. Value judgments about intelligence are furthermore not made overtly, which makes resistance to them difficult.

Western society is structured in such a way that many of the privileges and advantages that it offers, are generally more easily accessible to those who function in a particular way in the cognitive domain. “Success” as defined in terms of income, position, education, et cetera, is related to better performance on typical tests of intelligence. Although other characteristics such as social skills and personality also contribute, somehow the specific aspect of cognitive ability has been loaded with importance as a way of valuing and “ranking” people.

The way that IQ has been interpreted as inherent, fixed or immutable has contributed to the debate surrounding and negative attitudes towards its measurement. This static view of intelligence is associated with the hereditary view, while the changeable view of intelligence is associated with the environmental or nurture view. The nature-nurture debate regarding the

nature and development of intelligence is as old as human beings' interest in human intelligence.

More than 100 years ago, when intelligence measurement was still in its infancy, there were already differing views on the subject. John Stuart Mill (1806-1874) and Francis Galton (1822-1911) were two of the prominent early characters and representatives of the "nurture" and "nature" arguments respectively (Fancher, 1985). Mill, who was tutored by his father from a very young age and who did not think of himself as superiorly endowed, believed in the power of the environment and circumstance in cognitive development, and was a staunch supporter of educational programmes. Galton, on the other hand, who first coined the phrase "nature-nurture", was a member of an eminent family. He noted that certain families (including his own), had a disproportionally large number of eminent persons. He checked the family trees of talented individuals and found patterns of eminence in many of them (Fancher, 1985). Galton believed so strongly in the role of genetics in cognitive performance and other preferred characteristics, that he founded the "eugenics" movement aimed at selective human breeding. His purpose with the eugenics movement was to improve the hereditary quality of the human race by special "breeding programmes" of selected "qualified" individuals. This led him to foresee the use of some measure to select the most able men and women for this purpose - hence the development of the idea of an intelligence test (Fancher, 1985). Galton had a keen interest in all kinds of measurement and later pioneered the rating scale and questionnaire methods.

In 1884, Galton collected the first systematic data on individual differences in basic abilities in the hope that these abilities would predict differences in intelligence (Locurto, 1991). He used measures of reaction time, colour vision, hearing acuity, height and weight and obtained the measures of more than 10 000 people at, among other places, the 1884 International Health Exhibition in South Kensington (Locurto, 1991). Galton was one of the first people to notice the occurrence of "regression to mediocrity" whereby parents with an extreme characteristic, tend to produce children who are less extreme

in that characteristic - a pattern that he confirmed in his physical measures. Disappointingly for Galton, however, differences in these basic physiological processes did not relate to real-world measures of intelligent performance, such as academic grades or teacher ratings. Nevertheless, researchers and early experimental psychologists continued in their attempts to obtain measures that would reflect cognitive ability and which could be used to predict future performance with some accuracy.

According to Thorndike and Lohman (1990), efforts to measure human intelligence were a major objective of research in American psychology, before the publication of Binet's 1905 scale. Binet, after abandoning his own initial attempts at using physical measures, eventually used a far different approach in his endeavours to develop an intelligence test. He basically viewed intelligence as modifiable, and regarded views of intelligence as an immutable characteristic as "brutally pessimistic" (Wolf, 1973). This view clearly places him with supporters on the "nurture" side. Ironically though, translations and adaptations of his own instrument came to be used in a way quite contrary to this basic belief of his.

The "nature-nurture" debate about the heritable (unchangeable) versus the environmental (modifiable) views of intelligence has been the focus of numerous books, articles and public debates. Twin studies in support of the hereditary view and social investigations of training programmes in support of the environmental view have continued over many years (Eysenck & Kamin, 1981; Jensen, 1981; Locurto, 1991; Richardson & Bynner, 1984) in repeated attempts to investigate the social versus the genetic influences on cognitive development and performance. Twin and sibling studies have failed to provide unambiguous evidence for the heritability of IQ. Many confounding variables further complicate such research. As an example, even in the same home, siblings do not experience exactly the same environment. Factors such as motivation, nutrition, health, birth order, family relations, school experience, peer influence and educational pressures from the family can also affect performance differentially. Furthermore, parents provide their children with

both genes and a specific environment. The high-IQ parent is likely to provide his or her child with intellectual stimulation in the home and emphasise performance at school. It is therefore extremely difficult to separate the effects of genes from those of the environment (Eysenck & Kamin, 1981). Both heritable factors and the environment influence IQ to some extent, and a large proportion of individual differences in IQ is not accounted for by the direct effects of either heredity or the environment (Locurto, 1991).

Recently the publication of *The bell curve* by Herrnstein and Murray (1994) once again led to heated debate on the nature of intelligence. What was being published in the popular press did not represent the scientific data available on the topic. This led Gottfredson (1997b) to organise 52 prominent scientists in the field of intelligence research to sign a document addressing, from the scientific viewpoint, the most common claims and misconceptions which had resulted from the debate about the book. The intensity of feelings about this topic and the prominence and vehemence of the academic and public debates have continued to fuel the ongoing “IQ controversy”.

In the light of continuing debate surrounding the various views on the nature and development of intelligence, it could rightfully be asked why intelligence generates such controversy. One possible answer to this question may be that society places a high premium on intelligence and it therefore becomes an important form of barter to attain various forms of success in society. Some of these “rewards” include good education, employment and financial rewards (Sternberg, 1997b). Richardson and Bynner (1984, p 512), in referring to research on cognitive ability, are of the view that “the only research programme in this area worth pursuing is the ‘optimistic’ one of investigating what kinds of environmental obstacles impede cognitive functioning more than others and considering the appropriate educational strategies that are needed to overcome them”. This view is in agreement with that of Binet, who was primarily interested in the cognitive assessment of children to assist them in their development with appropriate further training. This interest of Binet is also the basic premise of learning potential assessment.

2.2.3 Theories of intelligence

A review of the different theories of intelligence provides a framework for the different approaches to its measurement. Many different theories of and measurement approaches to intelligence have been developed, each contributing to our understanding of it. The choice of theory affects the way in which concepts are defined, impacts on the measurement devices used and also influences the way in which results are interpreted. Only certain theories will be discussed here in the light of their relevance to the present project.

Binet's concept of the nature of intelligence influenced his choice of tasks used in his intelligence scale. Binet only wrote in French, but a few important papers have been translated into English (Binet & Simon, 1905/1916; 1915; Wolf, 1973). Binet's view of intelligence was that "in intelligence there is a fundamental faculty, the alteration or lack of which, is of the utmost importance for practical life. This faculty is judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances. To judge well, to comprehend well, to reason well, these are the essential activities of intelligence." (Binet & Simon, 1905, pp 42-43, in Thorndike & Lohman, 1990). Although Binet discusses many different aspects of intelligence, he essentially viewed it as a single entity which could be measured by means of a combination of specific tasks. He set out to devise a test that used performance in higher mental functions, rather than physiological measures, to measure intelligence.

Around the same time that Binet was working on the first intelligence test, Charles Spearman was devising his theory of intelligence. According to Spearman's theory, a single "general" factor (*g*) forms the basis of all intellectual activities, while excellence in particular areas depends on specific factors (Spearman, 1904). The high positive correlations found between different intellectual tasks were ascribed to the presence of the "*g*" factor in the

instruments. Consequently, tests for the measurement of intelligence are largely attempts to measure the amount of this factor. According to Schepers (1998), the "*g*" factor can be manipulated by focusing on particular content. The common variance would then be attributed to that particular aspect and not necessarily to what is commonly understood to be the general factor or "*g*" which is assumed to underlie all other abilities. Even when multiple factors are identified, second-order factor analysis usually indicates some underlying general factor. The content of a particular instrument will therefore determine how "*g*" is measured. It should be remembered that the factors extracted by means of factor analysis are mathematical abstractions and are "neither things nor causes" (Gould, 1981, p 255). According to Owen (1998), the tests that measure mainly "*g*" are concerned with abstract relationships such as Raven's progressive matrices (RPM) and Cattell's culture-fair intelligence test (CFIT). The test developed for the present research used the same type of nonverbal, figural items used in Raven's and Cattell's instruments incorporating general reasoning skills, and can therefore be considered to measure "*g*".

Cattell (1963) later found that Spearman's "*g*" split into two distinct general factors, which came to be known as fluid intelligence (g_f) and crystallised intelligence (g_c). Fluid intelligence is involved in tests that have little cultural or educational content, whereas crystallised intelligence involves abilities that have been acquired mostly through education, such as verbal and numerical ability. This theory is still used today, because there is support for the existence of these two general factors (Carroll, 1997a, 1997b; Jensen, 1994). Since fluid general ability is a prerequisite for the acquisition of crystallised general ability, the former is considered to be more fundamental than the latter (Schepers, 1972). According to Cattell (1963), fluid ability reaches an early maximum between the ages of 14 and 15, while crystallised ability increases to between the ages of 18 and 20 or beyond, depending on the opportunities for learning.

The relevance of this theoretical view to the present project is that an attempt was made to construct a test that would measure mainly fluid intelligence, in the form of basic reasoning ability. Measurement of learning potential generally

focuses on disadvantaged groups. The inclusion of content related to crystallised intelligence that measures the result of learning opportunities would therefore only perpetuate differences stemming from a socioeconomic or educational disadvantage, and is therefore unsuitable. Figural items using lines, circles and other geometric shapes which are generally considered to have the least cultural content (Jensen, 1981) were used for this instrument. The norm group consisted of a multicultural group of grade 9 and grade 11 pupils with a mean age of approximately 15 years - the age group in which fluid ability would be close to reaching its maximum.

Multifactor theories are not considered here, since the focus of the present project is on an instrument with one-dimensional content to facilitate the training which forms a crucial part of learning potential assessment. Furthermore, most of the multifactor theories rely largely on crystallised intelligence. For example, Thurstone's (1938) research led to his seven primary abilities, namely verbal comprehension, word fluency, numeric (arithmetic computations), spacial or geometric relations, associative memory, perceptual speed and general reasoning (Anastasi & Urbina, 1997) of which only perceptual speed and spacial or geometric relations are not crystallised abilities.

Vygotsky's (1978) theory of the zone of proximal development (ZPD) which portrays intelligence as changeable, is the most important theory for the present research project. Vygotsky distinguishes between the level of functioning that a person can reach without help and the level of functioning a person can reach with help. This allows one to estimate the potential future level of functioning. The test developed for the present project uses the test-train-test framework involved in the measurement of learning potential, based on Vygotsky's theory.

Contextual theories of intelligence, as represented by the work of Sternberg's (1985) triarchic theory of intelligence, take into account how intelligence is applied to the internal and external world in a variety of contexts. Other factors that also influence performance, such as personality and motivation, are also

acknowledged. The drawback of the contextual theories of intelligence is that every new context requires a different instrument, which is practically unfeasible. However, considering features that are universal in intelligent performance, Sternberg (1984, p 318) argues that there are certain shared elements of human behaviour that are "likely to be a part of intelligent functioning in virtually any human environment" and identifies the following metacomponents:

- recognising the existence and nature of a problem
- deciding upon the processes needed to solve the problem
- deciding upon a strategy into which to combine these processes
- deciding upon a mental representation upon which the processes and strategy will act
- allocating processing resources efficaciously
- monitoring problem solving
- being sensitive to the existence and nature of feedback
- knowing what to do in response to this feedback
- acting upon the feedback

Some of these features, such as identifying the type of problem, decision process for solution, choice of strategy, and the use of feedback, can be related to the development of the test for the present project, which is discussed in more detail in chapter 5.

To further emphasise the generic elements involved in human reasoning in terms of the very basic reasoning skills required, a strategy of reasoning set out by Aristotle almost 24 centuries ago can be used (Neman, 1989). The following are four of the main elements in thinking which Aristotle emphasised:

(1) definition - distinguishing particular features that make distinction and grouping possible
(2) comparison - looking for similarities and differences
(3) causal relations (the effects or consequences) - related to completion of

patterns

(4) authority (the view of the experts), which can be related to feedback or training

By breaking general thinking down into such basic elements, it becomes possible to recognise their application in real-life intelligent behaviour as well as in typical tasks of cognitive assessment instruments. In terms of the instrument for the present research project, geometric shapes and patterns need to be recognised, comparisons need to be made to determine the pattern or causal relation, and authority needs to be included in the initial explanation of practice examples with feedback, in the training in the test as well as in the scoring of answers as correct or incorrect. Such an instrument measures fluid intelligence in a culturally fair manner which makes possible the measurement of a general reasoning ability that can be assumed to apply in most contexts where human reasoning is required.

2.2.4 Measurement of intelligence

The history of the measurement of intelligence has shown specific patterns with some original ideas being revived in later years with the advantage of incorporating new and improved theoretical and technological developments. Psychological testing refers to the use of psychometric instruments to obtain information about individuals or groups so that understanding of the people concerned and decision making that affects them can be improved. Psychological testing owes its existence to the fact that people are different from one another. If there were no differences between individuals, testing would be redundant (Owen, 1998).

In the measurement of intelligence, one's theoretical view of the nature of intelligence influences both the content of the test and the criterion against which the measures obtained will be evaluated. Standard tests can tap only certain aspects of cognitive functioning and consequently cannot claim to

measure intelligence in its widest sense. Despite the fact that some psychologists believe that it is impossible to have a single notion of intelligence suitable in all cultures, the aim of intelligence tests should be to tap into those general cognitive skills and abilities that are generic and which can be found in most cultures and in most activities and behaviour requiring cognition. Such general cognitive abilities could be equated with Sternberg's metacognitive components or the Aristotelian reasoning skills discussed in the previous section. In terms of test content, nonverbal, figural content that measures fluid intelligence is considered to be most culturally fair.

Ability testing goes back 4 000 years, if one takes into account the early civil service examinations set by the Chinese in 2200 BC (Anastasi & Urbina, 1997; Garrett, 1941; Gregory, 1996; Matarazzo, 1990; Thorndike & Lohman, 1990). This practice was discontinued only in 1906 when, in response to widespread discontent, "the examination system was abolished by royal decree" (Francke, in Gregory, 1996, p 4).

Experimental psychology gained widespread prominence in the late 1800s in Europe and Great Britain (Gregory, 1996) with a new emphasis on objective methods and measurable quantities. Some of the measures obtained were interpreted as manifestations of intelligence. In 1879, Wundt, who had been involved in various experiments of measurement from the early 1860s, founded the first experimental laboratory in psychology in Leipzig, Germany (Gregory, 1996). Many prominent scientists from all over the world went to study there to gain practical experience and improve their qualifications (Wolf, 1973). This era of psychological testing is sometimes referred to as the Brass Instruments era of psychological testing because of the use of assorted brass instruments to measure sensory thresholds and reaction times.

The era that followed can be represented by Galton, who believed measurement to be the primary criterion of scientific study and who meticulously obtained thousands of physiological measures in his attempt to investigate the differences between individuals and groups. Galton believed that individual differences are objectively measurable by means of standardised procedures and adapted some

of the earlier psychophysical measures used by Wundt and others to use as quick sensorimotor measures. Galton's efforts in devising practicable measures of individual differences led to him being regarded as the father of mental testing (Gregory, 1996).

James McKeen Cattell completed his doctoral dissertation on reaction time under Wundt's direction (Anastasi & Urbina, 1997). Cattell also had contact with Galton, and on his return to America in the 1890s, established laboratories for experimental psychology. He imported the Brass Instruments to the USA and helped to promote testing. Cattell invented the term "mental test" (Gregory, 1996) and shared Galton's view that sensory measures could be used to measure intellectual functions. Cattell regarded the accuracy with which such measures could be obtained as an advantage, and was especially interested in the measurement of differences in reaction time. In 1901, Wissler, a student of Cattell, obtained both mental test scores and academic grades from more than 300 students at Columbia University, but found virtually no correlation between the mental test scores and academic achievement (Fancher, 1985; Gregory, 1996). The very modest correlations between the mental tests themselves were furthermore damaging to this approach to mental testing.

In Europe in 1881, a decision to enforce universal education in France resulted in retarded children also attending normal schools. At the time, the diagnostic methods used to determine the degree of retardation were intuitive and crude.

In France, an organisation known as La Société had been founded to give teachers and school administrators the opportunity to meet to discuss problems of education and to be active participants in research investigations (Wolf, 1973). When, in the early 1900s, members of La Société spearheaded movements to engage the Ministry of Public Instruction in doing something for retarded school children, Binet, who had become involved in La Société, was appointed to lead the study. The aim was to group those who were considered to be educable retarded children in special classes annexed to the regular school where they could receive special attention. Binet was convinced that these educable retarded children need not be condemned to useless and

barren lives, and that at least some of them could benefit from further training. He was also convinced that intelligence could be deliberately improved by "mental orthopaedics" to strengthen mental ability in the same way that physical exercises can improve physical strength. Binet saw a compelling need "to find a way to differentiate those children who could learn from those who could not" (Wolf, 1973, p 22). What resulted from Binet's collaborative work with Theodore Simon was the instruments that internationally became the forerunners of the metric intelligence scale. This was a fundamental breakthrough which had an important international influence on the subsequent development of measures of intelligence.

Although, on the surface, Binet's first scale appears to have been developed within a year, it was the culmination of over 15 years of development (Wolf, 1973). Binet had succeeded where Galton, Cattell and many others had failed, namely in developing a test which bore a significant relationship to real-life manifestations of intelligent behaviour - the first successful intelligence test (Fancher, 1985). Binet's research was pragmatic, using a large series of short tasks related to everyday problems of life which involve basic processes of reasoning (Gould, 1981).

Binet saw the measures obtained as at best tentative, because further development and learning could lead to different diagnoses in future. Inherent in this view was the changeability of intelligence - that in certain cases it could be further developed than would appear from initial measures. Binet was particularly concerned about possible misdiagnosis, especially in borderline cases. Whereas truly subnormal children could waste much of their own time and the time of the children in the class if placed in an ordinary school, he considered the bigger tragedy to be if a truly normal child was unfairly stigmatised for life by being misdiagnosed as retarded (Fancher, 1985).

By 1904, Binet had turned his attention to the more complex tasks of reasoning and thinking, convinced that "the only way to study the nature of intelligence was to use complex mental tasks that manifestly required the application of

intelligence for their completion" (Thorndike & Lohman, 1990, p 6). Binet's ideas on intelligence influenced his choice of material for his test development. He furthermore assumed that intelligence should be measured by a variety of tasks and that intelligence increased with age in children.

The first (1905) version of the Binet-Simon test consisted of 30 tasks arranged in approximate order of difficulty. In the 1908 scale, tasks were modified and grouped according to the ages at which normal children passed them. This later (revised) scale included 54 tasks and was the first to yield a mental level score (Thorndike & Lohman, 1990). The child began with tasks for the youngest age and proceeded in sequence until he or she could no longer complete the tasks. The age associated with the last task he or she could perform, became his or her "mental level". Wolf (1973) emphasises that Binet used the term "mental level" and not "mental age", which was how it was later interpreted. Binet worked alone on the 1911 revision which involved some technical changes to the 1908 version.

Goddard brought the 1908 Binet-Simon scale to the USA and standardised it for use there. According to Gould (1981), Goddard was the first person to popularise the Binet scale in America. He translated Binet's articles into English, applied his tests and tried to further their general use. However, in contrast to Binet's intention, Goddard regarded the scores obtained as measures of a single, innate ability.

In 1911, the German psychologist Stern proposed the popular concept of mental age which led to the relation between mental age and chronological age, expressed as a single number, the intelligence quotient (IQ) (Thorndike & Lohman, 1990). Children who tested at a mental age higher than their chronological age, would have an IQ greater than 1, while those who tested below their chronological age, would obtain an IQ less than 1. Wolf (1973, p 203) from personal interviews with Theodore Simon reports that Simon "continued to think of the use of IQ as a betrayal of the scale's objective". According to Thorndike and Lohman (1990, p 35), the fact that the interpretation

of IQ scores are the same, regardless of the child's age

> *may be responsible for a notion that has caused untold havoc in mental testing ever since, because it can be misinterpreted to mean that an individual's intelligence is constant. IQ values tended to be stable over time, however this tendency to maintain the same relative position in a group does not imply that the intelligence of any individual is constant and could not be altered by environmental changes.*

Unfortunately, the fact that the intelligence quotient remains essentially constant over the years of the child's development tended to be incorrectly interpreted as meaning that the IQ measured a relatively innate general ability.

Terman (Terman, 1916; Terman & Merrill, 1937) developed his own revision of the Binet-Simon scale for use in the USA with levels of performance expressed as IQs in an attempt to standardise results. He increased the number of items to 90 and introduced standardised scores with a mean of 100 and standard deviation of 15. It was Terman who became the primary architect of the popularity of the Binet scale (Gould, 1981). The revision of the test by Terman - a professor at Stanford at the time - became known as the Stanford-Binet. The Stanford-Binet probably brought the term "IQ" into common use, unfortunately with the aforementioned "static" interpretation as a consequence. According to Wolf (1973), Binet had given permission to Burt in England to translate the test, but not to Terman and other researchers in the USA. Whereas Binet had used only 50 subjects to standardise his scales, Terman used 1 400 subjects and his 1916 revision of the Binet-Simon test came to serve as the benchmark against which other intelligence tests were measured (Locurto, 1991). The following changes occurred during Terman's revision (Thorndike & Lohman, 1990):

- The item arrangement was changed.
- The amount of credit given for a correct answer was different.

- The manner in which the norms were developed was also different.

In the process, Binet's test had changed from a practical developmental tool to identify schoolchildren who needed help to an index interpreted far more rigidly with an altered purpose of ranking, sorting, classification and labelling. In the process, Binet and Simon's claims of precision for their instrument were greatly transcended (Wolf, 1973). Binet did not view the score obtained as an entity unto itself - he saw it as something scalable, like height. He greatly feared the misuse of the scores, which were intended as a guide for identifying children who needed help. Gould (1981) highlights three cardinal principles which Binet insisted upon for the use of his tests. These were later completely disregarded when his tests were translated and standardised in the USA. The principles are as follows:

(1) The scores do not define anything innate or permanent and should be used only as a practical device.
(2) The scale is a rough empirical guide to help identify children who need special help and not a device for ranking normal children.
(3) Emphasis should be placed upon improvement through training, and low scores should not be used to label individuals as innately incapable.

It is obvious that during the translation and altered interpretation of test scores in the USA, Binet's intentions were not acknowledged and adhered to. Being aware and mindful of the possible consequences of self-fulfilling prophesy, Binet's intention was "to identify in order to help and improve, not to label in order to limit" (Gould, 1981, p 152). Developments of learning potential measures over the last two decades have moved closer again to Binet's original intended use and interpretation of intelligence test results.

World War I provided great impetus to cognitive test development. A group of psychologists, including Goddard and Terman, were brought together to develop a test that could be used on all army recruits, with the express idea "to identify recruits with respect to intellectual functioning for placement purposes,

and in particular to keep the feebleminded out of the army" (Locurto, 1991, p 20). These efforts were to have far-reaching consequences for test development and test use in the years to follow. The demands of having to test nearly two million soldiers resulted in the development of the first group test similar the Stanford-Binet which was called form Alpha. A second, nonverbal form Beta, was also developed to be administered to non-English-speaking men (Locurto, 1991). Based on these test results, men were graded and suggestions offered for their proper military placement (Gould, 1981). When World War I ended, these tests became popularised in educational and work contexts, where they are still widely used today.

2.2.5 Use of intelligence test results

Although tests are mainly intended to make information available to improve decision making, the use of intelligence tests has always been controversial. Most of this controversy can be related to the different conceptions of intelligence which were discussed earlier. "As tests are merely samples of behaviour, the generalization of results to the behaviour outside the test situation implies statements of probability rather than certainty" (Owen, 1998, p 5). However, while tests are intended to assist in making appropriate decisions to the benefit of people, unfortunately, in practice, this does not always happen.

In terms of the use of test results, Binet's interest was in identifying present ability with a view to providing developmental opportunities to improve functioning, at whatever level the child functioned. The use of test scores during World War I illustrates a different use of intelligence test results and how the same findings can be interpreted differently. Average test scores for foreign-born recruits rose consistently with years of residence in the USA, and clear indications were found of correlations between intelligence and schooling - with a correlation coefficient of 0,75 between test scores and years of education (Gould, 1981). Low initial performance of new immigrants was interpreted, however, as an indication of the poor quality of newer immigrants,

compared to those who had been in the country for longer. This negative and "fixed" or "immutable" interpretation of cognitive test results led to the 1924 Restriction Act, which limited the number of immigrants allowed (Gould, 1981). Another interpretation might have been that exposure to American society and improved living conditions resulted in improved scores of people who had been in the country for longer. Instead of providing an impetus for social reform, however, the information was used to implement restrictive measures, based on a static, unchangeable view of mental ability.

In South Africa, similar opposing views of test results were held by Fick (1939) who supported the hereditarian view, and Biesheuvel (1943) who took the contextual view which recognises the importance of environmental factors in the development of cognitive ability. Fick (1939) uses the term "inferiority" in describing the differences between African and European examinees and interprets these differences as permanent. Biesheuvel (1943), on the other hand, provides information on a variety of environmental and educational factors that influence the development of cognitive ability and views intelligence test scores "not as a direct measure of innate ability, but as a measure of hereditary potentiality as it happens to have been realized by specific environmental circumstances" (Biesheuvel, 1943, p 18). This again indicates how the same test results can be interpreted and used in quite contrasting ways - to label and discard, on the one hand, or to try to understand and make provision for further development, on the other. The multitude of factors - from nutrition to motivation and the more general socioeconomic and educational environments - that can influence intelligence, precludes giving simplistic answers to this extremely complex problem.

The contextualist view that IQ differences are to some extent a reflection of differences in life circumstances, forms the basis of learning potential assessment where a direct attempt is made to indicate how cognitive measures can be altered, even within the span of a single testing session, when relevant instruction and training (ie improved conditions) are provided. This approach reflects a return to the view of Alfred Binet almost 100 years ago.

2.2.6 Addressing the issue of bias and culture-fairness in tests

Some of the controversy surrounding the use of psychometric tests can be attributed to the inability to distinguish between fairness and bias in testing. Fairness is a concept that relates to the fair and equitable use of tests, and is based on social values and philosophies of test use. Test bias on the other hand, is an objective and technical issue (Visser, 1996). Recently, the term bias has largely been replaced by differential item functioning (DIF). For the present discussion both terms will be used interchangeably.

In the South African context, where cross-cultural testing is of particular interest, it is of crucial importance that DIF should be investigated for tests that are constructed or for existing tests that are generally used. When developing tests for different cultural groups, it is essential to do DIF analysis to ensure the construction of fair and equitable measures which will measure the same construct for the various groups.

According to Cleary (1968, p 115), “a test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent non-zero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup”. Osterlind (1983) defines bias as a systematic error in the measurement process which leads to the consistent distortion of a statistic. A biased item is therefore one for which the probability of success is not the same for equally able test takers of the same population, regardless of their subgroup membership. When membership of a subgroup influences performance in an item, the item indicates bias. A model which is generally accepted by psychometricians and the legal fraternity (Schmitt & Noe, 1986) proposes that a test is considered biased if the criterion score predicted from the common regression line is consistently too high or too low for members of a subgroup.

Various techniques based on both classical test theory and item response theory exist with which bias of items can be analysed. Osterlind (1983) describes IRT-based item characteristic curve methods as the most elegant for teasing out item bias. Ertubey and Russell (1996), suggest that because of their greater sophistication, IRT procedures provide the best results for detecting cultural differences on particular items. Whichever technique is used, a crucial issue is the clarity and theoretical as well as practical justification of the definition of the groups that will be compared. In South Africa, the Society for Industrial Psychology has published guidelines on the validation and use of assessment procedures in the workplace (PSYSSA, 1998a) as well as a code of practice to promote and ensure fair psychological assessment in the workplace (PSYSSA, 1998b).

According to Schepers (1972), measures of fluid general ability can be profitably used in cross-cultural research. He also proposes careful consideration of timing of tests, since speeded tests pose particular problems for developing (disadvantaged) groups. In cross-cultural research, the time limits of tests may add a speeded element for one cultural group but not for another, depending on the importance of time in a particular culture (Schepers, 1972). In the construction of the test for the present project, the test content is aimed at the measurement of fluid ability, and no time limits are used.

Two of the internationally best-known examples of culture-fair tests are Cattell's (1950) culture-fair intelligence test and Raven's progressive matrices (Raven, 1958; Raven, Court & Raven, 1977, 1985). Cattell's test was developed to reduce the influence of verbal fluency, culture and education and was originally developed to aid in the fair measurement of the intelligence of persons who differ in terms of language, culture, education or socioeconomic status (Catttell, 1950). It attempts to reduce the influence of background in the measurement of intelligence, or what Cattell describes as the "general mental capacity factor, '*g*'" (Cattell, 1950, p 3). Raven's progressive matrices (Raven et al., 1985, p 3) were designed as a means to assess a person's ability to "think clearly,

irrespective of past experiences or present ability for verbal communication". Raven's matrices were first published in 1938, and have been used in countless cross-cultural studies over the years. Both Cattell's culture-fair intelligence test and Raven's progressive matrices use the type of nonverbal, figural item content also recommended by Jensen (1981). Similar items were used in the present project.

In the educational field in South Africa, cognitive test results have shown predictive validity for different groups. Claassen et al. (1991) distinguished between environmentally disadvantaged and nonenvironmentally disadvantaged groups. They found that correlations between the scores on the General Scholastic Aptitude Test (GSAT) Senior and school subject percentages generally ranged between 0,20 and 0,50 with correlations generally being approximately 0,10 lower for the environmentally disadvantaged group. In a separate study, Hugo and Claassen (1991) investigated the functioning of the GSAT for black students and found that although the correlations between the GSAT scores and school subjects were significant in most cases, the correlations for the black students were generally considerably lower than those reported for the other groups of students - ranging from close to zero in the case of some subjects to 0,56 for the average score for grade 11 (standard 9) pupils. The verbal score generally underpredicted the school achievement of the African pupils when the regression line for the English-speaking group was used. Van Eeden (1993) concluded from her study with the SSAIS-R (Senior South African Individual Scale - Revised) on children whose mother tongue is an African language, that although the SSAIS-R can be regarded as valid for predicting future school achievement, the predictive value of a specific score for this group may not necessarily be the same as for the norm group.

According to Jensen (1981, p 132)

> *"culture-reduced" tests try to minimize cultural loading by not using words, letters, numbers, or even pictures of familiar common objects. They consist of only simple elements - lines,*

> *curves, circles, and squares - and involve such universal concepts as up/down, right/left, opened/closed, whole/half, larger/smaller, many/few, full/empty, and the like. Quite complex problems involving relational reasoning can be made up of such elements - for example figural analogies, figure series completion, and matrices. Such tests are near the opposite extreme on the culture-loading continuum as compared with tests involving specific factual knowledge or scholastic content.*

It is clear that when instruments of cognitive ability are designed to be used fairly in multicultural groups, the content of the items used should avoid language, numbers, letters and other culturally loaded material. International tests of this nature use items with figural content. Hugo and Claassen (1991) also suggested that items such as figure analogies and pattern completion should be used in the development of an intelligence test for cross-cultural use in South Africa. The item types used for test development of the present project are Figure Series, Figure Analogies and Pattern Completion. These three item types adhere to the international and national recommendations and practices for cross-cultural general cognitive ability assessment.

The emphasis on the measurement of cognitive potential instead of the measurement of a purportedly “fixed amount” of intelligence, underlines the fact that allowance should be made for improved learning conditions to lead to altered levels of performance. According to Armour-Thomas (1992, p 562), the prevailing social and political climate in society is marked by an increased awareness of and respect for cultural diversity among all its people, and that “such awareness and respect have placed an additional responsibility on those connected with the intellectual assessment enterprise to develop and use measures of intelligence that are more culturally sensitive”. The purpose of assessment should be to provide conditions and strategies to uncover potential through dynamic assessment to foster optimal development of intellectual competence for all individuals, regardless of their cultural background. This view is a modern echo of Binet’s original intention. The approach known as

dynamic testing focuses on the identification of potential and provision of learning opportunities. It aims to help individuals improve their level of functioning in a psychometric test with the assumption that they would show similar improvement in real-life training situations if relevant training were to be provided. In recent years there has been an increased interest in this topic as a major field of study in psychological testing (Grigorenko & Sternberg, 1998; Lidz, 1987a, 1987b), representing a move back to the original Binet approach. The influence of Binet, father of the psychometric measurement of intelligence, is clearly still evident in present-day developments in this field.

2.2.7 Binet's view coming full circle

From the beginning, Binet intended his test to be used as a practical device, independent of theory of intellect. He thought that it should be used as an empirical guide to identify children who need special help and that the emphasis should be on improvement through special training (Gould, 1981). Of particular importance for the present project is that Binet had little interest in predictive validity - his greater concern was to identify pupils who needed help and would benefit from remediation (Hilliard, 1990). The same can be said of initial developments of modern-day measurement of learning potential.

Binet did not view the results of his test as a measure of innate intelligence. He regarded his test as a diagnostic instrument which could be used to help identify children whose intelligence was not developing to their full potential. By means of mental exercises, Binet believed that these individuals could be helped to increase their performance. These views are also held by people attempting to measure learning potential.

However, just as Binet's test had originally been developed with a focus on the possible training and improved performance of retarded school children, but was later extended to measure intelligence over the entire spectrum of ability, the present project is aimed at achieving this same extension for the measurement

of learning potential. Whereas initial interest in the measurement of learning potential had focused mostly on mentally retarded or low-ability examinees, the aim of the present project is to extend the measurement and interpretation of learning potential measures to the entire ability range.

The extraordinary contribution of Binet can also be seen in other ways. Even in very recent technological development, such as computerised adaptive testing - one of the most powerful recent developments in psychometrics - elements of Binet's original test can be found (Reckase, 1988). This is dealt with in more detail in chapter 4.

One factor which contributed to the development of dynamic test-train-retest learning potential measures, is the evidence that exists regarding the changes in intelligence test scores over time.

2.3 CHANGES IN INTELLIGENCE TEST SCORES OVER TIME

2.3.1 Introduction

In addition to measurement error which is an inherent part of any psychological measurement, mean group scores on ability tests change over time. These changes can be attributed to general development as well as changes in the socioeconomic and educational level of the population. For instance, the average educational level attained has changed dramatically in the last 50 years. One consequence has also been that mean score differences between culture groups have been shown to decrease as the differences in socioeconomic background and educational opportunities between the groups decrease.

Binet was the first person to focus attention on the possibility of changes in intelligence test scores over time. When he warned against the possible misclassification of people on the basis of test scores, he underscored the problem of inaccurate measurement and also allowed for possible changes of test scores over time. He viewed intelligence as changeable, and his interest in

mental orthopaedics by means of which intelligence scores could be improved, confirms this view.

Internationally, IQ scores are standardised on a scale with a mean of 100 and a standard deviation of 15. A raw score is transformed to this standard IQ score by means of norm tables. Societal changes continually provide improved education, training and access to information which affect the general ability of the population. A case in point is the recent explosion of information that has become available through the Internet. Such changes bring about increases in mean group performance on standard IQ tests over time. Since the population mean IQ is per definition equal to 100, it follows that IQ tests have to be restandardised and renormed from time to time to accommodate such mean group changes in test scores and to update the test content if necessary.

A test can be regarded as outdated as soon as the population mean and standard deviation deviate from 100 and 15 respectively. Once a test has been standardised, its continued use over time could show large mean increases over time ranging up to 25 IQ points in the span of a single generation (Spitz, 1989). This means that the scores of contemporary groups are higher than those of comparable groups who took the same test many years before. Spitz (1989) collated results from studies in which groups were given tests that had been standardised at different times, and found that in a single generation in many nations, large increases in IQ scores were evident. When tests that were standardised many years before are used, this would mean that the mean score might be higher than the theoretical 100 and would therefore not reflect the "correct" interpretation of the scores. In South Africa, when the General Scholastic Aptitude Test (GSAT) was constructed in the mid-1980s, the same was found when GSAT results compared less favourably with those obtained in the New South African Group Test (NSAGT) which had been normed over 20 years before (Claassen et al., 1991). In a way, the IQ mean of 100 is somewhat arbitrary, but its use is fairly ingrained and the standard IQ scale will probably be used for some years to come.

Changes in test scores can also be noted in and between subgroups of the population. For instance, differential changes in scores have been found between cultural groups. Where certain culture groups are disadvantaged, an improvement in the socioeconomic and educational opportunities of the disadvantaged group results in increases in the mean group score which are beyond the normal population increases over time. This in turn leads to smaller differences between the mean scores of culture groups.

2.3.2 Evidence of changing test scores

Vincent (1991) reports shrinking differences between black and white groups over time with differences between younger people becoming smaller while adults still maintain larger difference scores. In a meta-analytic study aimed at detecting patterns in cross-cultural studies on cognition reported between 1973 and 1994, Van de Vijver (1997) also found that performance differences between groups increased with chronological and educational age. In South Africa, Verster and Prinsloo (1988) reviewed the changes found between different cultural and language groups over time. There used to be distinct differences in socioeconomic and educational background between English-speaking and Afrikaans-speaking groups which were reflected in differences in mean group cognitive test results. These socioeconomic and educational differences diminished over time. On reanalysing the data of Biesheuvel and Liddicoat (1959), Verster and Prinsloo (1988) compared the results of different generations and found larger differences between the English-speaking and Afrikaans-speaking adults (12,5 and 13,9 IQ points) than for the younger group (7,0 and 8,2 IQ points). These results indicate that an improvement in socioeconomic and educational circumstances can affect performance in one generation, and is similar to what has been reported internationally (Plomin, 1997; Vincent, 1991).

In terms of overall group performance, an improvement in the socioeconomic and educational conditions of the Afrikaans-speaking group over time resulted in

decreases in the mean score differences between the two language groups. Langenhoven (1957) reported differences of 10 IQ points between English-speaking and Afrikaans-speaking pupils on nonverbal measures. In 1965, a difference of 7,4 IQ points in favour of English-speaking over the Afrikaans-speaking pupils was reported (Verster & Prinsloo, 1988); and in 1981, the difference between these two groups had further decreased to 4,5 IQ points. This latter difference was also later reported by Claassen (1990). Most of these studies focused on the white English-speaking and Afrikaans-speaking community, since at that stage, separate tests for separate cultural groups were common practice, which hampered comparison between cultural groups. Claassen (1997) ascribes the diminishing differences in scores of the English-speaking and Afrikaans-speaking groups to changing socioeconomic and educational conditions which resulted because of the urbanisation of Afrikaans-speaking whites.

2.3.3 Group mean score differences and the interpretation thereof

In addition to mean population test score changes over time, differences between cultural groups and also between age groups in the same culture group have been reported. Various research studies have investigated national and ethnic differences in intelligence, providing data to suit virtually any argument in the spectrum of possible explanations. However, according to Aiken (1996), the question of black-white differences in intelligence remains unresolved. Differences of approximately one standard deviation between Blacks and Whites have been reported in various studies. These findings are usually interpreted by the hereditarians as evidence of the inherent inferiority of the African group (Herrnstein & Murray, 1994). However, if this argument is extended, the approximately one standard deviation also found between White and Asian groups - in favour of the Asian group - should be similarly interpreted, which has not been done. Research shows that differences in intelligence test scores are related to the socioeconomic and educational opportunities of groups, and that changes in the latter bring about commensurate changes in

mean group scores on typical intelligence tests (Claassen, 1997).

Van Eeden (1993) indicated that smaller differences are found between pupils from different cultural, but similar socioeconomic backgrounds. The effect of language proficiency is also an important factor when tests contain verbal material. It is clear that socioeconomic and educational factors influence cognitive test results. South African history serves as an example of how improved conditions for a disadvantaged group (Afrikaans-speaking whites) led to better mean group performance in cognitive tests in a generation or two. Only in the last four to five years has South Africa embarked on the road to improving socioeconomic conditions and educational opportunities of the disadvantaged groups. These changes are certain to impact on future test results. Differences in mean scores between groups should decrease as conditions for the disadvantaged groups improve. However, in the meantime, existing group differences emphasise the need for assessing learning potential rather than crystallised abilities.

Claassen (1996) recently developed the Paper and Pencil Games, as a screening test for all South African pupils in the second to fourth school year to measure developed general scholastic ability. He found that “the mean of the English-speaking group was about one standard deviation above that of the Afrikaans-speaking group which in turn was about one standard deviation above that of the group speaking an African language” (Claassen, 1997, p 304). These differences held for both the verbal and nonverbal scores. The larger difference once again reported between the English-speaking and Afrikaans-speaking groups can be attributed to the cultural composition of the two groups. In the White and Indian communities, the majority of pupils are of a higher socioeconomic status (SES). The English-speaking group, consisting largely of White and Indian pupils, therefore represents a higher socioeconomic group. The Afrikaans-speaking group consists of White and Coloured pupils. Although most of the White group is of higher SES, in the Coloured community, more than 80 percent of the pupils are environmentally disadvantaged or of low SES (Claassen et al., 1991). This could explain the slightly larger differences

of about one standard deviation again found between English-speaking and Afrikaans-speaking groups compared to differences of only about a third standard deviation in earlier studies (Claassen, 1983, 1990, 1996, 1997).

It is clear that improvements in socioeconomic and educational opportunities are reflected in changes in the mean scores of groups. These findings provide support for Plomin's (1997) view that cognitive instruments are a barometer of social and educational standing. At present there are large differences between the cultural groups in South Africa in terms of socioeconomic and educational opportunities as well as general living conditions, with the African group in particular being the poorest off in all respects (Central Statistical Service of South Africa [CSS], 1996c). Political changes have brought hope of addressing these differences, and in time the disparities should decrease. Shuttleworth-Jordan (1996, p 97) notes "signs of a narrowing and possibly disappearing gap across race groups on cognitive test results in association with a reduction in socio-cultural differences" and regards this as being indicative of similarities between people in terms of cognitive processes. Similar to what happened between the English-speaking and Afrikaans-speaking groups, smaller differences between the mean group scores of the cultural groups should follow in time. However, at present, with our society in an ongoing process of change, these differences still impact on test performance and need to be taken into account.

Although socioeconomic and educational factors have been emphasised with regard to their influence on test scores, many other factors also impact on cognitive development and consequently also on intelligence test performance. In a study which investigated the environmental influences of the community environment, Coon, Carey and Fulker (1992), found several aspects of communities that show environmental relationships with the IQ of children over and above the genetic and environmental effects of parental IQ. They used aggregate measures of many sources such as income, education and occupation in the community as an index of the environmental influences affecting the child's intelligence instead of one or two single measures of the

child's household. Social scientists often assume that environments affect intelligence, but they do not always consider the fact that the environment that people experience can also be shaped by intelligence . For example, brighter parents tend to create different environments in terms of linguistic and economic factors and stimulation, than less bright parents, and in that way parental genes for higher IQ are experienced by their children both in an environmental and genetic manner (Gottfredson, 1997a, 1997b).

Jensen (1981) mentions prenatal, perinatal and neonatal factors, nutrition, birth order, family size, home and family environment, environmental deprivation and schooling as several factors that influence IQ. In terms of social and cultural differences, factors such as gender, socioeconomic status, occupational status, rural or urban environment, culture-biased tests, motivation, educational inequity, verbal deprivation, teacher expectancy, malnutrition, styles of child rearing and general environmental factors are put forward as possible reasons for differential cognitive development and IQ results. Situational factors such as the race of the tester, the language and dialect of the tester, tester attitudes and expectations, bias in test scoring or miscellaneous situational effects can furthermore also affect test performance (Jensen, 1981; Aiken, 1996).

According to Gottfredson (1997a), research on intelligence reveals group differences in intelligence which impose choices or dilemmas that people would prefer not to face. This is particularly relevant as far as differences in test scores between cultural groups are concerned. The often opposing ways in which these differences are interpreted can be related back to the basic nature-nurture differences in the view of the nature of intelligence. It is important to realise that differences are not created by tests - they merely reflect differences that exist, provided off course, that bias in items or tests is eliminated before differences are interpreted. Intelligence tests have often been criticised as being biased or unfair because they show differences between groups. One needs to acknowledge that such differences indicated between groups are often mirrored by real-life differences in various criterion measures.

The focus therefore needs to be on the identification of areas for development so that test results can be used to develop and improve areas of concern. In this regard, after questioning the desirability of research on racial differences, Loehlin (1992) concludes that the emphasis should be on the use of results to improve the status of the disadvantaged. This would again reflect Binet's original intention, namely that test scores should be used by educationists and people involved in training and development to identify those who can benefit from help and to use the results to plan further training and development.

According to Pyryt (1996), the most productive way to view differences is as indicators of equity in society, and rather than blaming the tests for revealing differences, they should be used to promote greater awareness of the need to invest in people through social and educational programmes. Measurement of learning potential by means of unbiased, culture-fair tests can also contribute to more meaningful interpretation of group differences.

2.3.4 The current South African context

Retief (1988) proposes that Southern Africa offers a "natural laboratory" for cross-cultural research and that in spite of methodological problems that exist in cross-cultural research, many opportunities for researching and solving cross-cultural problems exist here.

In South Africa, psychological tests were historically developed for industry and education. Both of these sectors used to be segregated along racial lines, which resulted in relatively homogeneous race groups. The tests that were developed, were therefore constructed for such homogeneous groups, which simplified test construction (Claassen, 1997). However, after recent political and social changes, the work and educational settings have become integrated and there are very few areas where homogeneous groups can still be found. Obviously, when multicultural groups are the target groups, the issue of test bias has to be addressed on scientific and psychometric principles, while the social,

political and philosophical issues that affect the use of tests and decisions that result from test use also need to be clarified.

In South Africa, with its multicultural society and changed political dispensation, issues such as affirmative action, cross-cultural assessment, language proficiency and socioeconomic differences are all societal issues that impact on psychological assessment and the construction of psychological tests. Some of these issues can be addressed at a psychometric level, whereas others will need to be discussed, clarified and decided upon on the basis of values and philosophical arguments.

According to Suzuki and Valencia (1997), different ethnic groups show different levels and profiles of scores on conventional tests of intelligence and related abilities. The explanations for this phenomenon include cultural bias in tests, linguistic requirements of tests and the effect of socioeconomic and educational opportunities. Much criticism has been levelled against standardised tests of intelligence in terms of their inappropriateness for the assessment of intellectual functioning of children from diverse cultural backgrounds (Armour-Thomas, 1992) with critics saying that these tests do not accurately measure the intelligence of cultural groups who are accorded a low status in society. Some of the most volatile issues associated with intellectual assessment are the interpretation of differences of test scores between cultural groups with hereditarians, on the one hand, explaining the differences in terms of inherited intellectual inferiority, while environmentalists, on the other, attribute differences to inadequate socialisation experiences (Armour-Thomas, 1992). In South Africa, Biesheuvel (1943) broke with the colonial research tradition which emphasised the inferiority of Africans (Fick, 1939) and emphasised the importance of home environment, schooling, nutrition and other factors on test performance (Retief, 1988; Claassen, 1996). He stressed the importance of understanding the skills and knowledge base of the target population when setting appropriate items.

It is clear, that there is presently an urgent need for culture-fair instruments that

focus on the measurement of learning potential to allow for current differences between culture groups, but with the focus on training and development. Hugo and Claassen (1991) found the figure analogies and pattern completion item types of the GSAT best for cross-cultural use, and suggest that a test with similar items be developed for all cultural groups in South Africa. These item types, together with figure series items, were included in the dynamic test aimed at measuring learning potential which was developed for the present project.

2.4 CONCLUSION

Internationally there is a renewed focus on cross-cultural testing and the measurement of learning potential is a growing field. In the next few decades in South Africa, the focus will definitely be on the construction and equitable use of unbiased and culture-fair tests in multicultural contexts. The dynamic nature of sociocultural processes and influences in South Africa as well as the different and shifting positions that people occupy along a continuum of levels of literacy, urbanisation and Westernisation should also be taken into account in the ethical use of psychological tests (Foxcroft, 1997a; Shuttleworth-Jordan, 1996). This brings about new challenges to test constructors and test users.

The history of intelligence testing shows a clear cyclical pattern with many of the original ideas of Binet becoming relevant again in present-day intelligence testing. Evidence that socioeconomic and educational opportunities affect intelligence test scores is of particular importance to South Africa in the present time of transformation with large differences still existing between the cultural groups on those indices. The measurement of learning potential, which allows for changes in measured intelligence following improvement in learning conditions, provides new opportunities for effective and culture-fair cross-cultural measurement of intelligence.

CHAPTER 3

THE MEASUREMENT OF LEARNING POTENTIAL

3.1 INTRODUCTION

Psychological testing of cognitive ability is generally used to make decisions about individuals and comparisons between people. As indicated in the previous chapter, this field has been beset with controversy and the nature (heritability) versus nurture (modifiability) debate continues. Standard tests of cognitive ability measure the products of prior learning. Hence they rely heavily on the assumption that all examinees have had comparable opportunities to acquire the skills and abilities being measured. However, this assumption is not true when people from different socioeconomic and cultural backgrounds are compared. In the case of people from disadvantaged backgrounds, for example, their abilities are likely to be underestimated, thus jeopardising the goal of fair evaluation (Campione & Brown, 1987; Hugo & Claassen, 1991).

The concept of "general intelligence" or "*g*" has been used for many years. While many researchers and theorists agree that it is problematic to try and represent intelligence by means of a single score, it is also not practically feasible to attempt to measure all possible kinds of intelligence(s) that can be defined. What exactly should be measured presents a very real dilemma for the development of psychometric instruments for the measurement of cognitive ability. The "*g*" factor can be extracted from the correlations between any large and diverse collection of mental ability tests. Thus typical IQ scores obtained from a combination of subtest scores reflect a measure of "*g*". Compared with other factors, it explains a large proportion of the total variance in the test scores. IQ scores and "*g*" have been regarded as useful to explain differences between individuals in terms of cognitive ability and performance. Standard cognitive tests address the need for measuring instruments that allow the evaluation of as well as comparison between individuals in terms of cognitive ability. However, growing dissatisfaction with standard tests of cognitive ability in multicultural

contexts has led practitioners to question the use of these tests, especially for people from disadvantaged backgrounds (Boeyens, 1989a, 1989b; Brown & Campione , 1986; Feuerstein, 1979; Lidz, 1991; Tzuriel & Haywood, 1992). There is concern that standard tests do not adequately allow for the educational and socioeconomic differences which people bring to the testing situation. It has increasingly been recognised that these differences, over and above real cognitive differences, also affect test results. Consequently there have been changes in the way in which cognitive ability is viewed with resulting changes in the way in which it is measured and increasing attempts to accommodate the measurement of people from disadvantaged backgrounds or different cultural groups. The dynamic assessment approach or learning test concept which has evolved as a result, has been described as an innovative new direction in the measurement of intelligence (Grigorenko & Sternberg, 1998). As Tzuriel and Haywood (1992) note, psychometric practices to some extent mirror the social movements and circumstances of their time. Furthermore, the context in which the measures are required and the way in which "ability" is defined, determine the methods we use to measure it.

In the previous chapter, the history of intelligence tests revealed how Binet's initial attempts were aimed at finding an instrument that could be used to distinguish educable mentally retarded subjects who could benefit from training from those who could not. What started as a practical measure to identify low-ability people who could benefit from further training and education developed into a measure of intelligence for people over the entire ability spectrum. However, Binet's view of intelligence as something changeable - something that could be improved by utilising "mental gymnastics" - was lost in the American translation and altered use of his test. Despite Binet's view of intelligence as changeable, which made provision for measuring the ability to learn, and although the concept of learning ability was mentioned repeatedly in later years, it was not operationalised in general psychometric measures of intelligence. In recent years, however, Binet's original views of intelligence as changeable appear to have surfaced again, this time in the form of dynamic assessment or the measurement of learning potential. This field is still considered to be in its infancy, but is receiving widespread attention in research (Grigorenko & Sternberg, 1998).

In the early 1920s, Dearborn (1921) commented that most ability tests were not tests of the capacity to learn, but tests of what has been learned. He indicated the need for tests that involve actual learning, proposing that "theoretically, it would follow that measurement of the actual progress of representative learning would furnish the best test of intelligence" (Dearborn, 1921, p 211). De Weerdt (1927) proposed the use of a dynamic test as a measure of the ability to improve after specific training. In 1951, Ombredane (in Dague, 1972) also stressed the need to use tests of adaptability or educability. Results indicated that although educability is partly a function of previous schooling, the measurement of learning in tests can open up richer perspectives of mental development (Dague, 1972).

The concept of the zone of proximal development (ZPD) proposed by Vygotsky (1978) has provided the theoretical base for a kind of measurement approach, known as dynamic assessment, which incorporates training in test administration in an attempt to measure learning potential or the ability to learn. The development of dynamic assessment can be directly attributed to growing dissatisfaction with the traditional "static" method of measuring intelligence - particularly where disadvantaged or low-ability examinees are concerned. Similar to what happened with Binet's test of intelligence, the dynamic assessment procedures were at first aimed at measuring the learning potential of mentally retarded or low-ability examinees. The present project in particular, is aimed at extending such measures to the wider ability spectrum.

3.2 THE HISTORY OF DYNAMIC ASSESSMENT

Some of the earliest definitions of intelligence refer to the "capacity to learn" or to "profit from experience" as a key element of intelligence (Binet & Simon, 1905/1916; Dearborn, 1921; Thorndike, 1922). Binet and Simon, the developers of the first general test of intelligence, intended their instrument to be used in a way that allows for improvement in test scores, acknowledging the modifiability of test performance.

Standard intelligence tests are known to predict academic success or failure fairly

well, mostly because the content reflects the kinds of problems often found in school curricula (Brown & French, 1979). These tests typically measure developed abilities, reflecting the products of education and experience and not innate capacity or potential (Lohman, 1993). As measures of prior learning, they are based on the assumption that all the individuals who are tested have had equal opportunities for learning.

Whereas the general definition and understanding of ability refers to ability that is available on demand, potential is concerned with what could be, and is based upon the possibility of change (Von Hirschfeld, 1992b). Dynamic testing is one way in which this possibility of change is addressed in psychometric terms. In dynamic testing, training is incorporated in the assessment to allow for differences in prior learning experiences and background. It helps to maximise performance, thereby providing better estimates of intelligence (Carlson, 1989; Tzuriel, 1997). It also offers the possibility of addressing some of the problems of assessing individuals from disadvantaged backgrounds.

Attempts to promote fairness in testing included assessment of bias and fairness in testing, and in recent years, have moved towards the use of new testing strategies, such as dynamic testing. Dynamic assessment supporters criticise traditional static tests of ability for their emphasis on products of prior learning (Tzuriel, 1992). At a conceptual level, proponents of dynamic assessment view cognitive ability as modifiable over time as environmental conditions improve. Measurement of the capacity to learn or the ability to adapt to change is therefore the focus.

Lidz (1987a) provides a historical overview of the development of dynamic assessment, which can be summarised as follows:

- In the 1920s and 1930s, many definitions of intelligence specifically referred to the ability to learn or the rate at which learning takes place.
- In the 1940s, learning ability was investigated in terms of the effect of practice and the relation between intelligence and learning. The general finding was that intelligence tests do not measure learning ability.

- In the 1950s, the effects of teaching or coaching on assessment results were the focus of attention because of the increasing commercialisation of test coaching. Results showed that children with higher initial scores tend to profit more from practice while those with lower initial scores respond more to coaching. It therefore seems that prior learning directly influences future learning and cannot be discounted by focusing on improvement scores only.
- In the 1960s, efforts were focused on devising practical measures of learning in attempts to assess educability. Experiments with the test-teach-test model were conducted - mostly with retarded children. Most authors continued to accept the traditional (static) views of intelligence and concluded only that learning ability was different from intelligence. Differences between ethnic groups were also investigated and proposals for dynamic alternatives suggested. The influence of environmental experiences on higher-level cognitive functions and the inadequacy of IQ measures to estimate the abilities of low socioeconomic status (SES) children were noted.
- In the 1970s, a substantial amount of pioneer research on dynamic testing ensued by people such as Feuerstein, Budoff, Campione and Brown, and Carlson and Wiedl. In 1978, Vygotsky's theory of the zone of proximal development was published in English. In the USA, legislation on psychological assessment, and in particular cognitive assessment, contributed to the need for finding alternative measures that would not be discriminatory. In Israel, the influx of thousands of children from different cultural groups necessitated the investigation of alternatives to the static cognitive measures, which did not provide satisfactory results. General dissatisfaction with static assessment practices, in particular for low-ability and disadvantaged examinees, led to the development of alternative assessment procedures with dynamic qualities. Initial research in support of dynamic assessment was published.
- In the 1980s, use of Vygotsky's theory of the ZPD gained more prominence and many researchers built on the work of the 1970s. The application of dynamic assessment was also extended beyond disadvantaged and educable mentally retarded (EMR) populations to deaf and reading-disabled students.
- Interest and research in dynamic assessment continued in the 1990s (Grigorenko & Sternberg, 1998; Lidz, 1997) with the development of measuring

instruments which use the basic concept of dynamic testing but can also claim adequate psychometric properties - a major concern (Guthke, 1992, 1998). According to Hegarty (1988), there are relatively few tests of learning ability available commercially, which limits their use. In the late 1990s, Grigorenko and Sternberg (1998) reviewed the work in this field and proposed that more research, especially validity investigations, is needed to allow this approach to make its promised contribution.

The need for dynamic measures has been noted by most researchers in this field, but providing the kind of evidence that will support their use has not proven easy. According to Reschly and Wilson (1990), persistent scepticism about measures that have not demonstrated their technical adequacy can be regarded as a strength in the field of psychological assessment. They propose that while healthy scepticism about dynamic assessment appears to be justified, dynamic assessment is moving towards overcoming some of the barriers by improving the technical adequacy of the procedures.

In South Africa, Biesheuvel (1943, 1952, 1972a, 1972b) made important early contributions to adaptability testing. His General Adaptability Battery (GAB) used measurement dynamically and took into account the influence of environmental variables such as culture and education on intellectual development. According to Blake (1972), pioneer work was conducted by Biesheuvel on the construction of the GAB between 1948 and 1952. Intended for administration to persons with primary-school education only, these group tests were administered by silent 16 mm film. The tests had their validity demonstrated in a variety of situations, mainly in the mining and manufacturing industries. According to Biesheuvel (1972b), the theme of adjustment is of crucial importance, also in cross-cultural psychology, and he emphasised that efforts should be directed towards the construction of “tests of adaptability” to measure potentiality to meet educational and vocational demands. Such tests differ from traditional intelligence tests in that they involve a learning component.

In a review of cross-cultural psychology and research in South Africa, Biesheuvel

(1972b) noted that the construct of “adaptability” played a significant role in cross-cultural theorising for a number of years in the early 1950s. Biesheuvel’s (1943) *African intelligence* represented the first comprehensive ecological approach in a study of Black/White differences, indicating various factors that affect cognitive development and performance. This approach emphasised the need for a closer look at the influence of environmental context variables on intellectual development. He suggested that the concept of “adaptability” provides scope for interpretations that include both genetic and cultural influences.

Owen (1998) emphasises the need in present-day South Africa for training or learning tests that allow some component of training in test administration. In our multicultural society, especially where people in the test situation often come from diverse backgrounds, dynamic assessment allows for equalising opportunities to perform optimally.

3.3 THE NEED FOR FOCUSING ON LEARNING POTENTIAL

Standard tests predict academic performance well (Fraser, Walberg, Welch & Hattie, 1987), which means that differences in test performance generally also manifest themselves in differences in academic criterion performance. This is partly due to the fact that many standard tests consist of subtests whose content resembles aspects of the material learnt in schools (eg Number Series, Word Analogies or Word Problems). These tests therefore largely tap scholastic learning, and people from disadvantaged educational backgrounds are more likely to perform at lower levels. Differences in test results may therefore reflect real differences, although these may be largely attributable to differences in educational and socioeconomic opportunities. Focusing on existing abilities and skills will therefore maintain the status quo and underestimate the potential performance of disadvantaged persons. When people from different cultural, educational or socioeconomic backgrounds are tested together, equitable and fair interpretation of test scores becomes extremely difficult. Based solely on current performance results, many disadvantaged examinees will be regarded as unable to cope with the demands of educational and training opportunities.

Given unaltered conditions, standard tests will probably continue to predict future performance reasonably well for all groups, since they are closely related to educational achievement. According to Owen (1998), the same cultural factors that affect test performance are also likely to have an impact on the wider behaviour domain that the test is designed to sample. Using psychometric tests which largely assess the effects of previous educational exposure, however, results in interpretations which are little different from those based on previous academic performance (Boeyens, 1989a; Hamers, Hessels & Pennings, 1996).

At this time in the history of South Africa, there is a need to specifically address the imbalance that resulted from unequal opportunities - hence the importance of anticipating the changes in performance that might occur when learning opportunities and conditions are improved. Thus measures that do not *only* reflect current ability and performance based on what has been learnt, will provide useful additional information. More appropriate selection procedures need to be developed for disadvantaged students with a reconsideration of some of the fundamental assumptions underlying the concepts of ability (Boeyens, 1989a). Procedures are needed which take into account that prior learning experiences and socioeconomic factors affect test performance.

Claassen (1997) proposes that a realistic objective in cross-cultural testing would be to construct tests that presuppose only experiences that are common to the different cultures. This would preclude any verbal material, as well as any material that relates directly to scholastic content such as number series, number problems and possibly even the use of alphabetical characters. Most cross-cultural tests make use of nonverbal content in order to obtain a more culture-fair measure of intellectual abilities. There is some evidence that nonverbal content involving pictures of cultural artifacts such as vehicles, furniture, musical instruments or household appliances do involve cultural loading. Items that are considered to be more culture-reduced include geometrical figures involving lines, circles, triangles and rectangles (Jensen, 1980, p 133).

One practical problem is that the more dissimilar test content is from educational content, the less accurate results are in terms of predicting future academic behaviour. Consequently nonverbal figural tests will be less effective in predicting academic performance than tests with verbal (or numerical) content. De Beer and Van Eeden (1997) found that language performance at school was the best predictor of academic performance - even for nonverbal subjects such as Mathematics. However, since the aim is to provide opportunities to disadvantaged examinees, culture-fair nonverbal material should be given preference. Apart from test content, another factor that needs to be taken into consideration is that "Africans lack the European's concept of competition in speed, and one cannot use speed of execution as a criterion of success" (Dague, 1972, p 66). Preference should therefore be given to power tests rather than timed tests in situations where the examinees are from diverse backgrounds.

The growing interest in learning potential assessment and dynamic testing indicates a shift in the way that cognitive ability is viewed for the sake of taking practical realities into consideration. By looking beyond current performance and acknowledging the possible influence of other factors on performance, more realistic measures and descriptions of cognitive development can be obtained. Despite a vast body of research, there is still no consensus about the extent to which cognitive ability is heritable, fixed or modifiable through experience (Haywood & Switzky, 1986). It is generally acknowledged that genetic factors make an important contribution to cognitive ability (Eysenck, 1971, 1988; Jensen, 1969a, 1969b, 1974, 1980; Plomin, 1997). At the same time it is acknowledged that differences in culture, socioeconomic background, parental education, educational opportunity, language, values, customs and child rearing need to be taken into account in assessment as factors that account for differences in performance (Haywood & Switzky, 1986; Sternberg, 1985). In this section, specific reasons for focusing on the measurement of learning potential will be discussed.

3.3.1 Dissatisfaction with traditional assessment

Many psychologists have expressed dissatisfaction with traditional models of assessment, especially for cross-cultural testing or testing of persons from

disadvantaged backgrounds. Problems with the cognitive assessment of disadvantaged people are related to general problems of the language proficiency, education and schooling of these groups. There is a significant difference in educative resources available in schools that serve disadvantaged populations, and present selection and evaluation procedures are inappropriate for these students (Passow & Frasier, 1996). Children who are in some way different from the norm group against which comparisons are made, may not have acquired the information or skills being measured, but may be able to do so if given the opportunity. When children from culturally different (or disadvantaged) backgrounds are assessed by means of static tests, the results may underestimate their potential level of performance under more favourable circumstances. This means that their future performance can be expected to be better than would be anticipated on the basis of their standard test performance. Earlier studies also indicated substantial changes in the test scores of African subjects following training or coaching (Verster, 1987). While at the time this was interpreted in a way that questioned the meaning of test scores for non-Western subjects, it actually provides support for the use of dynamic assessment procedures. The basic premise of dynamic assessment and the measurement of learning potential is that students from disadvantaged backgrounds can learn and profit from relevant experiences more successfully than one would anticipate when taking account only of present and proven academic and standard psychometric test results.

Seventy years ago, Fick (1929) applied the Stanford-Binet to White schoolchildren and also applied the Army Beta Test to a large sample of Black schoolchildren. He found the mean score of the Black pupils to be much lower than that of the White pupils, and attributed the differences to inferior teaching in the Black schools and the lack of opportunities for Black children. This reflects early acknowledgement that test scores could not be interpreted in isolation and that provision somehow needed to be made to account for differences in prior learning opportunities, although he later interpreted differences between Black and White pupils from the hereditarian viewpoint (Fick, 1939). Biesheuvel (1972b) acknowledged the influence of environmental variables such as culture and education on intellectual development.

Separate tests for different cultural groups were for many years the norm in South Africa, but the political and social changes of the last decade have brought about a need for the development of new tests that can be used for all the cultural groups (Claassen, 1997). Claassen (1983) investigated the functioning of the NSAGT for various cultural groups and found the verbal part unsuitable for Blacks, mainly because of their lack of proficiency in English. This emphasises the need to focus on nonverbal measures to identify undeveloped potential. Furthermore, Claassen's (1983) results pointed to SES as an important variable affecting IQ scores. Verster (1987) predicted a major challenge in the field of psychometric testing for employment selection in South Africa, with cognitive assessments needed to define more precisely the educational needs of different subgroups.

At the 1992 Psychometrics Conference, a number of authors focused on the need for changes in psychometric test procedures in South Africa, with several specifically mentioning dynamic assessment or the measurement of learning potential (Shirley, 1992; Von Hirschfeld, 1992a, 1992b; Taylor, 1992). Shirley (1992) and Taylor (1994b) emphasised the need for measures that can identify potential to enable identification of those who are most likely to benefit from the scarce opportunities to achieve success and compensate for deficient test performance resulting from disadvantage and deprivation.

Claassen (1997, p 297) emphasises that "testing in South Africa cannot be divorced from the country's political, economic and social history". In the South African context, traditional testing procedures can be unsatisfactory for use with people from culturally different backgrounds. The multicultural South African context necessitates the use of procedures and tests that take the diversity of examinees into account. The measurement of learning potential originated in attempts to make provision for people from disadvantaged backgrounds and because these measures make provision for the differences with which examinees come to the assessment situation, it is a particularly suitable approach to use for multicultural or educationally diverse groups

3.3.2 Present differences between groups in South Africa

For many years, South Africa was subjected to race-based policies which segregated communities and which resulted in a large percentage of the population being socio-economically and educationally disadvantaged. Although many changes have taken place since the first democratic election of 1994, the effect of the history of segregated and unequal living and educational conditions will affect people for many years to come. At present large differences in the life circumstances of the different cultural groups still exist and will continue to exist in the foreseeable future. Differences in scores on cognitive tests are one measurable result.

Some of the results of the October 1995 Household Survey (CSS, 1996a), the 1996 Census results (CSS, 1998) and the 1999 Reality Check Survey which have been shown to impact directly on cognitive development will be reported in this section in order to highlight the differences between groups. The October 1995 Household Survey results (CSS, 1996a) are based on information obtained from 30 000 households in October 1995 representing all households in South Africa, while the census results (CSS, 1998) are from the general census of October 1996. The 1999 Reality Check Survey is based on a representative national household survey of 3 000 adults (Reality Check, 1999).

Of the total population of 40,5 million, Africans make up 76,7%, Whites, 10,9%, Coloureds 8,9% and Indians 2,6%. It remains necessary to classify people into cultural groups so that progress and development over time as well as possible continued differences in life circumstances can be monitored (CSS, 1998).

Differences between cultural groups will be reported for socioeconomic and educational variables in particular because these differences are known to affect performance on cognitive ability tests.

- **Age distribution**

  There are different tendencies among the cultural groups of South Africa regarding age distribution. Among Africans, the typical age pyramid of

developing countries is found. Among Coloureds and Indians, the age distribution depicts a situation somewhere between developing and developed countries, while among Whites, the age distribution is typical of industrialised countries (CSS, 1996a, 1996c). These tendencies also have socioeconomic implications. There is a definite need for reprioritisation, and investment in human resources should increasingly be focused on young Africans, living in rural areas - particularly with regard to education and training.

- **Educational attainment**

  Educational attainment among South Africans varies, not only according to race, but also gender. African females have the lowest educational attainment in the country, followed by African males. White females and males have the highest educational attainment. What is noteworthy, however, is that although there are large differences in education between Africans, Coloureds, Whites and Indians, the situation has shown a steady improvement over time (CSS, 1996a, p 13).

  According to the 1995 Household Survey (CSS, 1996a), the mean number of years of schooling of the different race groups in 1991 was as follows: Africans (5,53), Coloureds (6,94), Indians (8,87) and Whites (11,02). The 1999 Reality Check Survey indicated that for the overall South African population, 23 percent of adults have only primary schooling or less, 67 percent have some senior schooling - including 22 percent with matric or equivalent - while only one in ten people has a postmatric qualification, including three percent with a degree. The latter survey also indicated that almost all Whites have passed at least grade 8, against 89 percent of Indians, 72 percent of Africans and 66 percent of Coloureds. An important aspect to note from the 1999 survey is that 33 percent of Africans aged 25 to 29 years have at least passed matric, while the figure is 18 percent for those between the ages of 40 and 44, and two percent for people over the age of 64. For the African group this indicates a very definite upward trend in terms of educational attainment. Although the samples that were used are not directly comparable, the differences in figures of the two surveys as well as the differences between the older and younger

groups in the 1999 survey show the increasing educational attainment among Africans, which will probably continue until parity with the other race groups is reached. While these figures are constantly changing and are likely to become more similar in time, for the foreseeable future, these differences will still have a serious impact on the South African society.

- **Unemployment**

  In the 1995 survey, a strict as well as an expanded definition of unemployment was used. The strict definition requires that a given individual should have taken specific steps to seek employment in the four weeks prior to a given point in time. The expanded definition takes into account the desire to work, irrespective of whether or not the person has actually taken active steps to find work (CSS, 1996a, p 15). Table 3.1 reflects the unemployment figures of the October 1995 Household Survey, using the expanded definition (CSS, 1996b). The 1996 census results are provided in brackets below (CSS, 1998, Table 2.30). The figures in brackets are based on the changed official definition of unemployment which was adopted for the 1996 census results according to which the unemployed are those people within the economically active population who

  (1) did not work during the seven days prior to the interview

  (2) want to work and are available to start work within a week of the interview

  (3) have taken active steps to look for work or to start some form of self-employment in the four weeks prior to the interview

TABLE 3.1 UNEMPLOYMENT RATES OF THE DIFFERENT CULTURAL GROUPS

Group	Total group	Males	Females
Total RSA	29,3% (33,9%)	22,5% (27,1%)	38,0% (42,0%)
Africans	36,9% (42,5%)	28,9% (34,1%)	46,9% (52,4%)
Coloureds	22,3% (20,9%)	17.8% (18,3%)	27.8% (24,1%)
Indians/Asians	13,4% (12,2%)	9,9% (11,1%)	19,9% (14,0%)
Whites	5,5% (4,6%)	3,7% (4,2%)	8,3% (5,1%)

* Top figure from the 1995 Household survey, figure in brackets from the 1996 census.

- **Job type** (CSS, 1996a, pp 19-21)

  Amongst employed Africans, 34 percent of males and 50 percent of females work in elementary occupations such as cleaning, garbage collection and agricultural labour (CSS, 1996a) - positions that are typically low in remuneration. A large proportion (35% males and 42% females) of Coloureds are still found in elementary occupations, but there is some movement into more skilled artisan and craft jobs (23% among Coloured males), and a move into sales and services (16%) and clerical jobs (16%) among females. A different picture emerges among employed Indians, which starts to resemble the picture found amongst Whites with only one percent of the males found in elementary occupations. Indian males are well represented in all occupational categories with a relatively large proportion (14%) in managerial occupations. A large percentage of Indian women (36%) can be found in clerical occupations. Whites, especially White males, tend to have access to

occupations requiring higher levels of competencies with 19 percent in white-collar management, 29 percent in artisan or craft positions and 17 percent in semiprofessional/ technical positions such as engineering technicians who require postschool technical qualifications. Forty-seven percent of White females can be found in clerical occupations. The picture that is presented by the above information indicates that in a typical pyramid structure, Africans tend to occupy elementary occupations, with progressively fewer people in those positions among Coloureds, Indians and Whites. The profiles of the Indian and White groups are similar, with the Coloured group falling somewhere between them and the African group. This trend is also evident in the consideration of socioeconomic circumstances and household incomes. The large percentage of Africans who are unemployed or employed in elementary occupations, compared with the other cultural groups, to some extent explains the differences in socioeconomic circumstances and household incomes found between the cultural groups.

- **Income**

  Of all employed South Africans, 26 percent earn R500 or less per month and 62 percent earn less than R1 501 per month. Only 11 percent earn more than R4500 per month. Table 3.2 summarises the income categories of males and females of the different population groups (CSS, 1998, Table 2.38).

- **Water in the home**

  One of the socioeconomic indicators used is whether there is tap water available inside the home. In terms of this indicator, the differences between the cultural groups are clear. Only 27,3 percent of African households have tap water inside the dwelling, while the figures for the other population groups are 72,4 percent for Coloureds, 96,4 percent for Indians and 97,6 percent for Whites (CSS, 1998). In the 1999 Reality Check Survey, the figures given for taps inside dwellings are 99 percent for Whites, 95 percent for Indians, 80 percent for Coloureds and 33 percent for Africans, showing some improvement in the three years between the two surveys.

- **Toilet facilities**

  Another indicator of socioeconomic status is the type of toilet facilities available in the household. Table 3.3 summarises the Census 1996 information in this regard, using percentages.

TABLE 3.2 INCOME CATEGORIES AMONG THE EMPLOYED, BY POPULATION GROUP AND GENDER (IN PERCENTAGES)

Income category	Percentage of the population group							
	African		Coloured		Indian		White	
	Male	Female	Male	Female	Male	Female	Male	Female
R3 501+	6,0	5,2	11,6	7,1	29,8	16,7	64,8	35,4
R1 501 - R3 500	20,1	13,3	27,5	21,5	38,4	32,4	22,5	40,4
R1 001 - R1 500	23,8	12,5	21,0	21,9	18,3	26,0	5,7	10,4
R 501 - R1 000	24,4	21,4	20,4	19,5	8,8	16,0	3,2	6,2
R0 - R500	25,8	47,5	19,4	30,0	4,8	8,9	3,9	7,6

TABLE 3.3 TOILET FACILITIES BY CULTURAL GROUP

Type of toilet facility	Cultural group			
	African	Coloured	Indian	White
Flush or chemical toilet	33,9%	79,7%	97,6%	99,2%
Pit latrine	43,5%	7,8%	1,8%	0,3%
Bucket latrine	5,6%	7,1%	0,1%	0,04%

None of the above	16,4%	5,1%	0,2%	0,09%
Other (unspecified)	0,6%	0,3%	0,3%	0,4%

- **Language**

  Language proficiency is an important contributing factor to cognitive and educational test performance, especially when the language of education and evaluation is not a person's first language. South Africa now has 11 official languages.

  Figure 3.1 indicates the percentage distribution of first-language speakers in South Africa, although the country is generally moving in the direction of using English as the main official language. It is clear from Figure 3.1 that, should English be the language used, for more than 91 percent of the population this will mean using a language which is not their first language.

The information in this section clearly indicates the large differences between the cultural groups in South Africa on various socioeconomic and educational indicators. Many of these are interrelated, and changes in one will certainly affect others. Gupta and Coxhead (1988) report that despite having equal opportunities and being exposed to the dominant culture, the abilities of children from different culture groups are still uniquely influenced by their particular home and cultural background. Although the government have committed themselves to the upliftment of the disadvantaged, this will still take many years, possibly decades to accomplish. In the meantime, the continued effect of cultural, educational and socioeconomic factors must be acknowledged, and steps taken to address them, also in the psychological measurement of cognitive development and performance. Whatever steps are taken to address the gaps that exist because of disadvantages in background, it needs to be acknowledged that people from underprivileged social conditions are at a disadvantage when exposed to tests that determine only current ability levels (Guthke, 1992), and that assessment of potential is required.

FIGURE 3.1 LANGUAGE DISTRIBUTION OF THE SOUTH AFRICAN POPULATION

* **Source**: The people of South Africa population census, 1996 (Census in Brief - Report no 1:03-01-11[1996]) (CSS, 1996c)

All of the educational and socioeconomic indicators reviewed in this section directly or indirectly impact on cognitive development and/or cognitive functioning. It is clear that extensive socioeconomic upliftment and educational and human capacity development are needed among the disadvantaged groups. Learning potential assessment can help direct resources where these are needed most and also help identify those individuals who are most likely to benefit from training and development - in particular among those currently functioning at low levels. Dynamic assessment and the measurement of learning potential can furthermore also accommodate the effects of continuing changes in socioeconomic status and educational attainment on test performance.

3.3.3 Dynamic testing of learning potential as a possible solution

Dynamic cognitive assessment, also known as learning potential measurement, is based on the view that cognitive processes are highly modifiable. It views interactive assessment as providing better insight into learning capacity. Owen (1998) describes dynamic assessment as involving a paradigm shift which cannot be viewed simply as another instance of psychometric assessment, while Lidz (1987b) regards dynamic assessment as a technique to be used in addition to currently available procedures and not a replacement for current approaches. Dynamic assessment includes a focus on the modifiability of the learner with a view to better understanding the level of current performance as well as the potential for improved future performance. Learning that is oriented towards developmental levels that have already been reached is regarded as ineffective from the viewpoint of a child's overall development. Teaching and learning should be aimed at improving current levels of attainment and reaching increased levels of performance. Measurement of learning potential helps to identify the present as well as the probable future levels of performance so that training can be aimed at realising the potentially improved performance levels.

Developmental levels and test performance continually change. This emphasises the need to move away from the static view of cognitive performance and the labelling that often accompanies it, towards a view that allows for the assessment of potential for future development. A basic view of cognitive ability as changeable is particularly appropriate if better educational and training opportunities can be provided. The focus therefore shifts to the identification of undeveloped potential. One way in which this can be achieved is by using dynamic testing, which offers a better chance of achieving fair results. Compared with conventional intelligence tests, learning tests reduce the differences between "disadvantaged" and "privileged" children as well as between members of different ethnic groups (Guthke & Stein, 1996). In this regard, Claassen (1997, p 305) states that "tests of learning potential show promise and are intended to serve a laudable purpose, but at this stage only limited information is available about the way they relate to more established measures of cognitive abilities".

Psychometric testing can make a contribution to the rebuilding of South Africa if it can deal with the realities of the present situation. In particular, there has to be a focus on assessment of people's capacities and undeveloped potential in preference to a focus on present ability and existing skills only. As indicated by international and national research, when the differences in living and educational conditions become smaller, there is a commensurate decrease in differences in test performance between groups. The people in the disadvantaged group, however, are not all disadvantaged to the same extent. Using existing or new instruments and making a constant score adjustment when a person belongs to this group will consequently not adequately address the differences that also exist within the group, and is therefore not a satisfactory option.

3.4 VYGOTSKY'S ZONE OF PROXIMAL DEVELOPMENT AS A THEORETICAL BASE

Vygotsky's (1978) concept and theory of the zone of proximal development (ZPD) is generally acknowledged as the theoretical base upon which dynamic assessment and the measurement of learning potential has been built. Various interpretations of the ZPD have been used. Because Vygotsky died before he could fully develop and operationalise his concept of the ZPD, what he has written is subject to different interpretations, some of which do not strictly adhere to his original ideas. Interestingly, although Vygotsky is generally viewed as the inventor of the ZPD concept, he is reported to have stated that the concept was not original and that the American investigators Meumann and McCarthy were the originators thereof (Van der Veer & Valsiner, 1991). The work of these investigators is, however, far removed from the concept of the ZPD as elaborated by Vygotsky.

In some of his early writings on the problems of deaf-mute, blind and retarded children, Vygotsky emphasised the importance of the social education of handicapped children and their potential for normal development. The way in which Vygotsky conceived the measurement of intelligence through IQ testing is in a way reminiscent of Binet's original concept which also made allowance for changes in IQ test performance.

Binet maintained that we can boost our IQ through instruction, and rejected the view that intelligence is an immutable inborn quantity. Vygotsky saw opportunities for understanding the mental processes of people and for establishing programmes for treatment and remediation.

Vygotsky (1978) linked learning and development in his definition of the ZPD. According to him (1978, p 85), it is "a well known and empirically established fact ... that learning should be matched in some manner with the child's developmental level". According to Campione, Brown, Ferrara and Bryant (1984), Vygotsky was concerned that the typical diagnostic, static tests used to assess children's competence within some domain might underestimate the potential of some individuals. He therefore suggested that each individual's response to instruction should *also* be investigated. In terms of Vygotsky's view, the task of assessment is to identify *not only* those cognitive processes that are fully developed, *but also* those that are in a state of being developed at the time of assessment and which can be identified by incorporating cooperative learning as part of the assessment procedure (Kozulin & Falik, 1995).

Based on this view, it is not sufficient to limit oneself to determining a single developmental level. There is a need to determine at least two developmental levels, namely the actual developmental level and the potential developmental level. The former refers to the results of already completed developmental cycles. The latter refers to the level that the learner can attain when some form of help has been provided. The zone of proximal development is "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978, p 86). Vygotsky criticised the practice of focusing only on the child's level of actual development "to the exclusion of the child's potential for growth" (Rogoff & Wertsch, 1984, p 2). Because the level of potential development may vary independently of the level of actual development, it has to be assessed separately in addition to the actual level of development. "The actual developmental level characterizes mental development retrospectively, while the zone of proximal development characterizes mental development prospectively" (Vygotsky, 1978, pp 86-87).

Vygotsky (1978, p 85) used “a simple example”, as he himself put it, to illustrate his view of and concern with the general practice of using only standard test results. The simple example that Vygotsky uses is the case of two children of the same age (10 years old chronologically) who initially measure at the same level of mental development (both eight years old in mental development level). These two children can therefore be considered exactly the same in terms of age and mental developmental levels. Based on this information alone, Vygotsky indicates that one would expect the future performance or “subsequent course of mental development and of school learning” of these two children to be the same (Vygotsky, 1978, p 86). However, he proposes that useful additional information can be obtained if one does not stop there. He argues that if these two children are shown additional ways of dealing with problems, differences between them may become apparent. Suppose that, following the additional training, it hypothetically, “turns out that the first child can deal with problems up to a twelve-year-old’s level, the second up to a nine-year-old’s. Now, are these children mentally the same?” Of course the answer is no. This clear and extremely simple example unequivocally explains Vygotsky’s theoretical concept of the ZPD as well as its practical implications.

Because most of the initial research on dynamic assessment and the measurement of learning potential involved low-ability disadvantaged or educable mentally retarded examinees, Vygotsky’s special case example could be applied directly. For these low-ability examinees who had generally performed quite poorly on initial unaided tests, the level of actual ability was comparably low. With the initial scores of the samples all being low and approximately equal, a number of researchers shifted their focus to the ZPD to investigate possible differences between these individuals. While Vygotsky (1978) used a simple example to illustrate his concept of the ZPD, he clearly indicated that both the actual developmental level and the ZPD should be used to interpret cognitive development. This would involve three measures, namely the actual developmental level, the potential developmental level and the ZPD which reflects the difference between the first two. In section 3.7, Vygotsky’s theory and its practical implications will be discussed further. For the present, the basic elements of his theory have been identified and can now be operationalised.

3.5 OPERATIONALISATION OF DYNAMIC ASSESSMENT AND LEARNING POTENTIAL

Researchers in learning potential assessment rely on Vygotsky's concept of the ZPD as their theoretical base. The ZPD is the difference between the *actual* (currently manifest) level of cognitive development and the level that can be reached with additional guidance, help or training. This concept forms the core of dynamic testing and the measurement of learning potential. Learning potential is what is measured, while dynamic assessment is the way in which it is measured.

The key characteristic of dynamic assessment is the test-intervention-retest format. Procedures differ considerably, although in principle they are based on specific interpretations of Vygotsky's theory. Quantitatively, three measures are implied, two of which are used to describe the cognitive development of the individual:

- The pretest reflects the actual (present) level of unaided performance.
- The ZPD is represented by the difference between the potential level of development after training (post-test) and the initial (pretest) level of development.

The ZPD reflects the individual's ability to further benefit from assistance and learning opportunities that are provided to improve upon present level of performance. It should also be noted that one cannot think of ability to learn independently of the mental operations required by the specific content taught (Dague, 1972). When learning potential is assessed, the domain that is used should therefore be clearly noted. Measurement of learning potential in a particular domain may be generalised to other contexts, but such generalisation needs to be substantiated by empirical validation.

In their interpretation of learning potential, many authors have focused only on the individual's potential to further benefit from instruction (ie the ZPD) as the principal variable. Current intellectual ability has therefore assumed secondary importance.

Vygotsky's special example, in which he uses children with the same current status on a standard IQ test but who are nevertheless different in terms of their cognitive potential, has thus been taken to represent the general case.

Using his example, Vygotsky indicated the limitations of standard static assessment of cognitive development. By also considering the ZPD, recognition is given to the fact that external factors affect cognitive development as well as cognitive performance. By incorporating training into the assessment context, provision is made for a more complete assessment of the person's current state of development. Hence a more comprehensive picture of his or her current developmental state can be obtained and used to predict the dynamics of development in the immediate future (Minick, 1987). Vygotsky's view acknowledges that even though people could presently be at the same level of performance, their future development may differ.

For two people who are initially on the same level of pretest performance, the person with the larger ZPD would probably benefit more from instruction to improve on his or her present level of performance. In general, people with larger ZPD scores are likely to improve their performance, while those with smaller ZPD scores are likely to maintain their present level of performance. A small ZPD indicates that performance is already close to optimal for that individual.

Vygotsky (1978) did not describe or give an example of a more general case. He did, however, clearly and specifically state that the zone of proximal development is to be used as a tool by means of which "we can take account of *not only* the cycles and maturation processes that have *already* been completed *but also* those processes that are currently in a state of formation, that are just beginning to mature and develop. Thus, the zone of proximal development permits us to delineate the child's immediate future and his dynamic developmental state, *allowing not only for what already has been achieved developmentally but also for what is in the course of maturing*" (own emphasis) (Vygotsky, 1978, p 87). He also states that "the state of a child's mental development can be determined only by clarifying its two levels: the actual developmental level *and* the zone of proximal development." (Vygotsky, 1978, p 87). What is clear is that Vygotsky incorporated both the present level of functioning and

the ZPD in his assessment of the individual's developmental level. These should therefore be the operational measures of interest in the assessment of learning potential.

3.6 DIFFERENT APPROACHES TO DYNAMIC ASSESSMENT AND THE MEASUREMENT OF LEARNING POTENTIAL

3.6.1 Introduction

Researchers have employed different approaches, procedures, techniques and measures in their use of dynamic assessment. The common link between all of these is that they involve some form of help or assistance to the person being assessed with a view to providing a more accurate assessment of individual differences than can be obtained with standard test scores.

The variety of terms used, such as "cognitive enrichment", "coaching", "learning tests", "graduated prompting", "testing-the-limits", "dynamic assessment" or "learning potential measurement" reflect the various approaches. Some try to bring about maximal levels of performance, others seek to measure the magnitude of response to instruction, while others still focus on the efficiency of operation of specific cognitive processes. Whatever the specific focus, they all have in common the evaluation of the extent to which the individual can improve his or her test performance. The most prominent approaches will be discussed here. It is clear that differences in goals of assessment will be reflected in the method of conducting the assessment as well as in the focus of measurement. In addition to differences in tasks, dynamic assessment researchers use different forms of teaching during their assessment. Decisions about the nature of the instruction are based on both theoretical and practical concerns.

Much of the early research on learning potential and dynamic assessment was concentrated on mentally retarded children (Budoff, 1967; Feuerstein, 1979), which limits the generalisability of results. However, recent research has included participants within the normal range of cognitive ability, gifted children and even

university students (Boeyens, 1989a; Passow & Frasier, 1996; Shochet, 1994; Zolezzi, 1995).

Various authors have categorised the different approaches to dynamic assessment (Campione, 1989; Grigorenko & Sternberg, 1998; Laughon, 1990; Lidz, 1991; Taylor, 1994b; Zolezzi, 1995). For the purposes of practicality, simplicity and clarity, two main approaches to dynamic assessment are distinguished here, based on the way in which Vygotsky's theory has been interpreted and operationalised as well as the desired outcome.

The first approach is the clinical or enrichment approach, where the focus is on the learning outcome of the individual. The aim of these approaches is to modify cognitive ability and achieve structural changes in cognitive functioning. The enrichment model makes use of remediation techniques. Mediation and enrichment of learning experiences are provided at individual level to overcome areas of deficiency and to improve thinking skills and cognitive functioning in the areas identified as underdeveloped.

The second approach is more psychometrically oriented, with the focus on the measurement of the magnitude of learning potential. The aim of this approach is not to effect enduring changes in cognitive performance, but rather to evaluate the capacity for acquiring new skills or knowledge when training is provided.

These approaches are used by different researchers who make use of a variety of mostly existing standard tests. Each type of application, however, includes its own dynamic assessment strategy commensurate with the particular approach.

3.6.2 Structural dynamic assessment: the enrichment approach

Feuerstein (1979) is generally regarded as the father of the cognitive enrichment approach to dynamic assessment, which represents the clinical approach to the measurement of learning potential. His research had a practical, empirical origin when

problems were experienced with the assessment of the cognitive functioning of culturally different and socially disadvantaged children. Feuerstein and his colleagues developed the Learning Potential Assessment Device (LPAD) when war orphans and young immigrants were sent to Israel, and the majority of these children appeared to be extremely low-functioning. Feuerstein (1972) reported that the scores of individuals from disadvantaged subgroups were on the average, almost always lower on tests, even on purportedly culture-free, culture-fair or developmental tests - leading to negative stereotypes and a pessimistic outlook. He warns that this negative and pessimistic outlook then becomes entrenched because it determines the amount, nature and quality of educational investment made in such children.

The focus of Feuerstein's approach is the modifiability of cognitive functioning. He views humans as open systems amenable to cognitive change. The approach is based on dissatisfaction with traditional measuring instruments to provide information about individual's learning ability. It is assumed that a lack or deprivation of mediated learning experiences is an important cause of low performance (Hamers & Resing, 1993). This approach is also based on Vygotsky's ZPD principle, although the emphasis is on the social interaction and qualitative aspects of the learning process. Vygotsky's concepts of current level of mental functioning and functions that are in the process of maturing are used, with the emphasis on developing those functions that are in the process of maturing. The dynamic approach proposed by Feuerstein and his co-workers represents an attempt to effect enduring change in the cognitive functioning of the individual. The LPAD is designed to measure an individual's cognitive modifiability, or the extent to which cognitive structures can be changed in response to a mediated learning experience (MLE) (Feuerstein, Rand, Jensen, Kaniel & Tzuriel, 1987). Attempts are made to obtain diagnostic information by analysing a child's activity in the ZPD to identify the strengths and weaknesses of his or her mental activity.

In the mediated learning experience, a human mediator is placed between the problem and the learner (Kozulin & Falik, 1995). Coaching or intervention is aimed at facilitating the individual's functioning in the proximal zone of his or her development (Feuerstein et al., 1987). An important element of this model is the analysis of tasks

which is guided by a cognitive map that is used to identify, clarify and modify a learner's deficiencies, attempting to locate the origin of success or failure (Feuerstein, Feuerstein & Gross, 1997; Lidz, 1991).

The result is a descriptive profile of modifiability with the primary focus on narrative and descriptive (rather than measurement) information. The Feuerstein mediated learning and cognitive enrichment approach makes use of the ZPD principles behind learning potential assessment but seems to emphasise post-test (after enrichment) performance more. Initial measurement is used to construct the cognitive map in order to direct the training provided. Standard tests, modifications of existing instruments or tasks specifically developed for the LPAD are used in dynamic mediational mode.

The LPAD research has been primarily focused on low-performing children using highly individualised clinical approaches. Individuals who already function at high levels are not viewed as legitimate targets for this kind of dynamic assessment (Feuerstein et al., 1987). Many decisions depend upon subjective judgment of the practitioner and the child's actions and responses determine the actions of the examiner. The role of the examiner is crucial, and because training is individual and not standardised, comparison of individual results is problematic.

The cognitive enrichment type of dynamic assessment, requires much skill, training, experience and investment in time and effort to administer (Tzuriel, 1997) and is consequently extremely expensive. Grigorenko and Sternberg (1998) also mention the lack of standardisation, low reliability and the substantial time and monetary investment required as limiting features of this approach.

Although Feuerstein's approach has many supporters, Frisby and Braden (1992) do not view it as a viable alternative to the proper use of rigorously researched individual IQ tests used by well trained professionals. Shayer and Beasly (1987) conducted a meta-analysis on data from three research programmes involving the comparison of experimental groups receiving instrumental enrichment (IE) based on Feuerstein's model and control groups. Most of the achievement subtest effect sizes they

reported were nonsignificant, and even those that did achieve statistical significance were of little practical significance. Considering the amount of time and human effort involved in IE, these results are disappointing.

Blagg (1991) evaluated the LPAD, and while commending some features of the programme, reported that the extensive training involved and the inconclusive results did not warrant unequivocal support. Over the duration of the programme, there were no significant improvements in reading skills, mathematical skills or work-study skills, nor was there any evidence of improved cognitive abilities as measured by the British Ability Scales (Blagg, 1991). Some positive effects were, however, found in that pupils became more active contributors to class discussions, more able to describe different strategies for solving problems and more likely to spontaneously read and follow instructions carefully. In reviewing the results of different mediated learning research projects, (Lidz, 1992) reported that mediation during assessment with the LPAD is associated with improved performance on a variety of tasks for a variety of learners. Two of the most powerful components in the mediational effect seem to be that of verbalisation and elaborated feedback. Mediated interventions seem most promising for lower-functioning students.

Feuerstein's LPAD has also been used in group dynamic assessment situations, although the information yielded by such administrations is less extensive than for individual administrations (Tzuriel & Haywood, 1992). According to Frisby and Braden (1992), the reliability of the LPAD administered in group format ranges between 0,7 and 0,95. However, the criterion validity of the LPAD with external criteria matching the nature of the testing has not been determined.

Skuy, Kaniel and Tzuriel (1988) investigated the use of the LPAD dynamic assessment techniques with academically superior children in a low socioeconomic status community in Israel. Their findings suggest that the LPAD can provide a basis for low SES children to be included in mainstream programmes for the gifted.

In South Africa, Van Niekerk (1991) investigated the effectiveness of Feuerstein's instrumental enrichment programme for a group of disadvantaged senior secondary

students. The aim was to evaluate the effects of the instrumental enrichment (IE) on verbal and nonverbal reasoning, perceptual speed, mathematical applications, vocabulary and study habits and attitudes using an experimental and a control group. The experimental group (N=13) participated in an IE course for 58,6 hours on average and was compared with a control group (N=15) for pretest and post-test performance. No significant differences were found between the two groups on verbal reasoning, nonverbal reasoning, perceptual speed, mathematical applications, vocabulary or study habits. Some of the experimental group members did benefit in terms of their study attitude.

The results of research investigating the effectiveness of these procedures have been mixed. There seem to be some gains on the softer variables such as motivation, attitude towards learning and student participation. However, in terms of the hard evidence for factors such as improved predictive validity, more research needs to be done.

In summary, the Feuerstein view of learning potential assessment is representative of the clinically oriented cognitive enrichment approach. The aim is to provide mediated learning opportunities, and to identify a cognitive map so that remedial mediated learning experiences can be designed to improve cognitive functioning. This approach differs from that of the present project, although both use Vygotsky's basic premise of the ZPD as their starting point. The aim of the Feuerstein approach is to bring about structural cognitive changes, while that of the present project is to obtain functional and standardised measures. What they have in common is that training is provided as part of the assessment procedure. The way in which the training is provided and the aim thereof are, however, distinctly different.

3.6.3 Functional dynamic assessment: the psychometric approach

This approach to the measurement of learning potential is also based on Vygotsky's ZPD theory, but the emphasis is on the measurement component. Here the focus is on standardisation so that measurement accuracy can be improved. Vygotsky's

(1978) theory of the ZPD with the use of both the actual (present) level of performance and the ZPD, clearly points to operationalisation using a pretest, training and post-test format. The pretest provides the actual developmental level, while the difference between the post-test and the pretest is taken as the ZPD measure.

Researchers have made persistent efforts to combine assessment of learning potential or measurement of the ZPD with sound psychometric principles (Budoff & Harrison, 1971; Guthke, 1992). The aim of more quantitative psychometric-oriented approaches is to obtain objective, valid, reliable and quantifiable measures of learning potential. These approaches have been criticised because in their quantitative focus, they fail to address Vygotsky's concern with child-oriented, qualitative evaluation with the emphasis on the social interaction elements of learning. However, with standardisation as the focus, this approach contributes to improved psychometric properties of the assessment of learning potential.

Although there are distinctly different approaches, a pretest-training-post-test procedure using standard psychometric tests describes many of these. The various procedures based on the psychometric approach differ in the degree to which the tasks used are domain-specific, the degree of standardisation in the interventions and the level of prescriptive or diagnostic information obtained. The similarity is that most are task oriented rather than child oriented (Kozulin & Falik, 1995).
The test material used by these approaches is similar to that found in traditional intelligence tests. Raven's matrices often provide material on which various versions of learning tests are built. The main psychometric approaches to measuring learning potential are discussed next.

3.6.3.1 Coaching on standard tests (the Budoff approach)

Budoff has been involved in learning potential assessment or training-based assessment measures since the early 1960s. His important contribution is the standardisation of instructions (Budoff, 1987a). His initial work was concerned with developing standardised procedures that would demonstrate whether optimised

procedures in testing would provide less biased estimates of intelligence or the ability to profit from experience. Reliable and extensively validated standard tests such as Raven's Progressive Matrices are used - but administered dynamically. In this approach there is a concerted effort to standardise the training, and the aim is to provide alternative measurements to conventional intelligence tests. In this regard, Budoff views learning potential as a measure of general ability (Grigorenko & Sternberg, 1998). Some of the research by Budoff and his co-workers (Babad & Budoff, 1974; Budoff, 1969; Budoff, 1987a, 1987b; Budoff & Corman, 1974) is characterised by a teach-within-test approach, while in other studies the pretest-train-post-test format was used. The bulk of Budoff's work concerns educable mentally retarded children of low initial ability level.

Budoff (1969) as well as Budoff and Corman (1974) studied the performance of educable mentally retarded children using Koh's block design test. They found that performance did not improve uniformly as a result of training, and consequently differentiated between "gainers", "nongainers" and "high scorers". High scorers demonstrate excellent understanding of the task prior to training. Gainers perform poorly on the pretest but improve markedly following instruction, while nongainers perform poorly initially and do not profit from the instruction provided. This broad classification was later replaced by a set of continuous scores. The post-test score, adjusted for the pretest level (ie a residualised score) was used. The potential to profit from intervention was found to be independent of current ability. They further found that learning potential (ie improvement or ZPD) scores did not correlate with either socioeconomic status or race.

Babad and Budoff (1974) developed a figural "series learning potential test" involving the choice of a geometrical shape that best completes each series and which involves colour, size, orientation and semantic content. It is aimed at the elementary and primary school level, and the results indicate that the post-training score predicts academic performance better than typical IQ measures. The fact that no mean difference in learning potential was found between subjects of higher and lower ability, supports the contention that ability and learning potential are independent. In a study using Koh's learning potential task and Raven's Progressive Matrices with subjects

who had been identified as educable mentally retarded, Budoff (1987a) found that grouping students according to their learning potential status provides a better prediction of their ability to profit from teaching compared with IQ or class placement.

The Learning Test for Ethnic Minorities (LEM) (Hessels & Hamers, 1993) is a specific learning potential test developed in the Netherlands for the assessment of general cognitive abilities of ethnic minority groups that generally fits in with the Budoff approach. Subtests of the LEM include classification, word-object association recognition, word-object association naming, number series, syllable recall and figure analogies. The testing procedure of the LEM consists of training within the test with test administration time similar to that of traditional intelligence tests. It was found that the LEM could strongly differentiate between children who had low scores on the conventional intelligence test, implying that some low-ability children may benefit from the learning potential test procedure.

A number of the South African studies that investigated dynamic assessment made use of the Budoff approach using standard tests administered dynamically with standard training provided (Shochet, 1994; Zolezzi, 1995). In each of these research studies, existing standard tests of cognitive ability were used in a test-train-retest approach with some form of standard training being provided between the pretest and the post-test. Shochet (1994) used two standard tests of the HSRC, namely the Deductive Reasoning Test (DRT) and the Pattern Relations Test (PRT) in an investigation of the role of dynamic assessment in the prediction of the success of Black undergraduate students. The aim of his study was to investigate whether the predictive validity of standard assessment methods is moderated by the cognitive modifiability of disadvantaged students. The independent variables were the results on the standard tests as well as the cumulative gain score on each of the two tests. The two gain scores were furthermore added to form a combined gain score. The criterion variable used was academic performance as measured by the number of credits obtained and average percentage grade obtained at the end of the first year. Multiple regression analysis indicated that only when cognitive modifiability was included, did the model of prediction become significant on both criteria, indicating that cognitive modifiability is an essential component in predicting academic performance.

Since Shochet's (1994) research found no valid predictor of success for highly modifiable students, he concludes that "modifiability does not necessarily constitute or predict success" - emphasising that the level of performance is also an important consideration.

Zolezzi (1995) conducted a similar study to that of Shochet (1994) and combined standard administration of standard tests with mediated administration. Mediation of 35 minutes was provided for the DRT and PRT and the post-test was an exact repetition of the pretest, except for some rearrangement of items of the DRT in the post-test. Zolezzi (1995) used Feuerstein's criteria for effective mediation as a guideline for the training provided. Although the results look promising, the small sample precludes generalisation of the findings.

The importance of Budoff's approach lies in his attempt to standardise procedures so that better psychometric measures can be obtained and the results of different examinees can be compared. According to Budoff and Pagell (1968), "the learning-potential assessment strategy seeks to obtain an estimate of general ability in a milieu that minimizes the possibly adverse effects of the child's prior experiences." Budoff generally followed Vygotsky's description of the ZPD reasonably closely by using a pretest-training-post-test strategy. The ZPD is taken as the difference between the post-test and the pretest scores. Budoff focused on the post-test score, also incorporating the level of pretest performance in a residualised gain score. This measure deviates slightly from Vygotsky's proposed combined use of the actual (pretest) level of performance and the ZPD which will be used in the present project. The present project will be similar to Budoff's typical research in that the pretest-training-post-test approach with standardised training will be used. The ZPD will also be calculated by subtracting pretest performance from post-test performance, albeit on an IRT-based scale and not with standard test scores as in Budoff's work.

The present project will make use of a test specifically designed for dynamic assessment instead of using existing standard tests. The type of figural nonverbal items used are, however, similar to those found in standard tests. Academic performance will also be used as criterion measures to investigate the predictive

validity of dynamic testing results. Another similarity between the Budoff group's work and the present project, is in the view of learning potential as a measure of general ability and of these measures as alternatives to conventional intelligence tests. Both also emphasise the use of nonverbal tasks to assess the reasoning abilities of people who may not have grown up in a rich verbal environment.

3.6.3.2 Graduated prompting (the Campione and Brown approach)

The graduated prompting model of dynamic testing involves the gradual transfer of control (Hamers & Resing, 1993). It is based on the information-processing theory of intelligence and makes use of standardised and hierarchically ordered hints for mediation. These predetermined prompts, based on task analysis, are provided in sequence from very general to very specific, until the final "hint" which is actually a blueprint for generating the correct answer (Brown & French, 1979; Campione & Brown, 1987; Campione, Brown & Ferrara, 1982; Ferrara, Brown & Campione, 1986). A pretest in the form of a standard test is first used to assess initial level of performance. A post-test similar to the pretest is administered dynamically and provides information on any improvement in performance (Brown & French, 1979). The testing procedure is standardised to produce psychometrically defensible quantitative data.

The focus of this approach is how much aid is needed to bring about a specified level of performance, rather than how much improvement is made. The number of hints as an index is likely to have psychometric properties only if the test administration is standardised as much as possible. Tasks of inductive reasoning (variants of progressive matrices problems and series completion problems) are mostly used because performance in such tasks is known to be related to scholastic success. These tasks feature in most ability tests, and consistently distinguish academically successful from less successful students.

Campione, Brown and Bryant (1985) used a matrices task and a series completion task with graded prompting procedures. They found that children of higher ability

tend to require fewer hints to solve the original sets of problems and to deal with new problems. An important finding by Brown and Ferrara (1985) was that learning disabled children needed far fewer prompts and hints to solve problems they were previously unable to solve compared with truly retarded children. Ferrara et al. (1986) found that children with lower educational performance required significantly more prompts than those with higher educational levels. The study by Ferrara et al. (1986) provided evidence that dynamic assessment can be used to supplement information found with static measures of ability.

In their studies of group differences comparing high-ability and average-ability children, Campione, Brown, Ferrara & Bryant (1984) found that group differences increased as transfer distance increased. Using the number of hints used as the measure of interest, they found that while correlations between IQ and the number of hints were nonsignificant for maintenance and near transfer, they increased and were reliable when far and very far transfer performance was considered. When the performance of a group of mildly retarded children was compared with a group of nonretarded children, the same pattern emerged. They generally found that "dynamic measures tended to be superior to the static measures in their ability to predict how much young children would profit from instruction" (Campione et al., 1984, p 89).

As far as is known, no South African studies have used this approach. Campione and Brown and their co-workers have operationalised Vygotsky's ZPD into practical measures in their attempts to obtain standardised and therefore comparable measures of the ZPD. They admit that the metric they use in their approach differs from the one that Vygotsky suggested (Campione et al., 1984). Whereas Vygotsky proposes that the extent to which an individual can improve upon initial performance should be the measure of interest, this group focused on the number of hints required for an individual to reach a predetermined level of performance. In an inverse measure, the larger the number of hints needed, the smaller the ZPD will be. This approach has certain drawbacks. Firstly, the measure used is the number of hints provided, and it is extremely difficult to compare hints in terms of the amount of help they actually provide. Also, the predetermined level of performance that is set as the

cutoff, builds in a ceiling effect on performance. Their approach has something in common with mastery testing, because a predetermined level of performance is required.

An area of similarity with the present project is the use of computerised test administration to standardise the procedures. Furthermore, the two approaches are also similar in that they can accommodate assessment of examinees from a wider spectrum of ability levels. However, the present project will differ from this approach in that it will use a test specifically designed for dynamic assessment and not standard tests as used by the Campione and Brown group. The approach for the present project differs further from that of the Campione and Brown group in that exactly the same standard training will be provided to all examinees. The measure of interest will include the individual's level of performance after standard training and not a fixed level of performance being set as the goal. Furthermore, for the present project, the initial level of performance as well as the difference score between post-test and pretest performance will be used and not the number of hints to a predetermined level of performance. In this regard, the procedure of the present research project will be closer to Vygotsky's proposed use of procedures and scores.

3.6.3.3 Testing-the-limits (the Carlson and Wiedl approach)

Testing-the-limits is based on the premise that intellectual and personality factors account for differences between individuals in processing of information. Assessment is focused on the effects of different methods of training on a transfer test to gain understanding of the examinee's specific ability to use the cues given by the assessor (Carlson & Wiedl, 1978, 1979). These authors used testing-the-limits procedures in the assessment of the intellectual capabilities of children with learning difficulties. Using Raven's Coloured Progressive Matrices, various different procedures were applied, namely standard instruction, verbalisation during and after solution, verbalisation after solution, simple feedback, elaborated feedback and elaborated feedback plus verbalisation during and after problem solution.

This approach represents an attempt to construct a theoretical framework that integrates empirical findings with information-processing theory (Grigorenko &

Sternberg, 1998). Testing-the-limits studies use standard tests in a novel way, and by focusing on the test situation, attempt to find a match between the training procedure and best performance for various disadvantaged subgroups. In comparing different methods, it has generally been found that verbalisation and elaborated feedback lead to higher levels of performance than the standard testing condition (Bethge, Carlson & Wiedl, 1982; Carlson, 1989).

One disadvantage of this approach is that because group performance is the focus of measurement, no individual comparison is possible. There is also no definite pretest, since the examinees are randomly assigned to the different mediation groups. Apart from the most basic elements involved in dynamic assessment - that is, training as part of assessment - the present project does not have much in common with this approach. The present project is aimed at accurate measurement and the provision of standard training in the test situation with the focus on individual results. Furthermore, the present project has definite pretest-training-post-test procedures with the focus on measuring the individual's pretest and ZPD scores.

3.6.3.4 Learning tests (Guthke's learning test approach)

According to Guthke (1998), the development of culture-fair tests, and subsequently, of learning potential tests, began because not everyone has the same learning history and opportunities to learn. In contrast to traditional tests, learning potential tests use standardised learning aids in the form of simple feedback or elaborated prompts. The aim of Guthke's particular approach is to meet the needs and demands of modern psychometrics while determining the individual's potential to learn. Guthke (1998) distinguishes between long-term and short-term learning tests. In the long-term learning tests, the training is much more elaborate and occurs over a longer period of time. Short-term learning tests require only one test session during which systematic feedback and assistance are offered (Guthke, 1998). The learning tests are strongly standardised and reintroduce psychometric standards together with the new learning or dynamic-oriented concept of testing.

Guthke's (1993b) aim has been to combine the advantages of assessment during a

training phase with the advantages of the psychometric methods in attempts to develop an objective and practical device. According to this approach, the extent to which individuals can improve their learning is needed in addition to their intellectual status. The focus of this approach is on the psychometric side, in attempts to achieve comparable results. With this focus, this approach is closer to the psychometric tradition of testing than most of the other dynamic testing approaches. A distinctive characteristic of this approach is that researchers use an adaptive procedure of feedback which is computerised and based on the examinee's specific responses.

There are five published German learning tests, some of which are based on conventional intelligence tests while others use new types of items (Guthke, 1992, 1993b, 1998). In the "classical learning test", a pretest battery of items involving a series of numbers, figures and analogies is administered. This is followed by individual or group training involving programmed manuals designed to teach problem-solving strategies. Afterwards, in the post-test, parallel items are used to examine the extent to which the subjects improved their performances as a result of the training (Guthke, 1992). The post-test score, which combines the initial level of performance with the rate of improvement as a result of training, is regarded as the decisive measure of long-term learning tests.

Attempts to have the same type of measures available but with shorter testing times, led to the development of short-term learning potential tests where the training phase is directly implemented into the procedure (Guthke, 1993b). In these approaches, different types of training have been evaluated, such as systematic feedback or extensive assistance and simple feedback. In this way, it is similar to the Carlson and Wiedl approach discussed earlier. Test scores are determined by the amount of help the examinee needs to reach a predetermined level of performance - similar to the Campione and Brown approach discussed earlier. In the short-term learning tests, the original coloured form of Raven's Progressive Matrices test is used. If the items are solved incorrectly, a puzzle format and a set of graded hints based on Galperin's learning theory are applied, and the child is guided to the correct solution. The primary variable of interest is the number of hints needed. Guthke (1992) reports that this kind of testing takes hardly any more time to administer than the standard test, but

leads to much higher predictive validity. Closer inspection reveals that learning tests seem to be particularly useful for children of below-average intelligence and those exposed to “irregular learning conditions” (Guthke, 1993a, 1998).

Although Guthke (1998) uses a standard test, additional items are used in the adaptive feedback procedure. In a longitudinal study over a period of seven years involving kindergarten children, Guthke (1998) found that when the group was used as a whole, the learning test version of Raven’s Coloured Progressive Matrices did not have improved predictive validity when compared with the standard version. However, when a low-functioning group was taken separately (identified by kindergarten teachers as slow learners), their results indicated that the predictive validity of the learning test version was superior to the conventional version at all times and in all measured criteria. It therefore seems that learning tests are particularly suited to the assessment of children with learning disabilities and/or irregular learning histories (Guthke, 1998; Lidz, 1987b). The learning tests have made a contribution to the field of dynamic assessment, in particular in terms of the accuracy of measurement and the standardisation of the training provided. General findings are that people differ in the extent to which they are able to improve their performance after receiving relevant training (Guthke, 1992) and post-test scores are usually more strongly related to external criteria than pretest scores.

The similarity between the present project and the learning test approach to dynamic assessment is their use of the test-train-retest approach with a shared focus on the psychometric properties of the procedures and tests used. The aim of both is to provide psychometrically sound learning potential instruments that can be generally used in cognitive ability assessment. Another similarity is the use of IRT in the development of the procedures. In terms of administration procedure, the learning test version directs the testing session in such a way that the next item that follows and the specific assistance provided are adapted to the examinee’s performance level and the errors that he or she makes. While the learning tests make use of this adaptive procedure for feedback, they do not incorporate dynamic assessment based on IRT, in the way that the present project does. In learning tests, the test score is determined by various methods such as the amount of help and the kind of prompts

the individual needs during the test or the level of post-test performance, while the present project focuses on level of performance only. The key characteristic of the Guthke (1998) approach is the combination of strongly standardised psychometric standards with a dynamic approach to learning potential testing. Acknowledging that many of the existing dynamic assessment procedures are extremely time consuming and awkward to apply, this approach (in particular the short-term versions), attempts to address some of the problematic areas of dynamic assessment, such as the lengthy administration time and lack of psychometric foundation. The latter is another point of similarity with the present project.

3.6.3.5 The IRT approach (recommended by Embretson and Sijtsma)

The aim of dynamic assessment is to modify an examinee's performance level by providing instructions as part of the assessment. While the use of classical test theory leads to measurement problems regarding difference scores, the use of IRT latent trait models provides a means of accurately equating scores. Embretson (1987) and Sijtsma (1993a, 1993b) propose that IRT-based procedures and, in particular CAT, provide a solution to many of the psychometric problems that have been associated with dynamic assessment and the measurement of learning potential. The psychometric features of dynamic assessment instruments can be vastly improved if IRT and CAT procedures are used. Learning potential assessment needs a sound psychometric foundation, and IRT and CAT can solve several measurement problems associated with this field (Embretson, 1987, 1991, 1992; Sijtsma, 1993a, 1993b).

Embretson (1991) developed a latent trait model for the measurement of learning and change where the change measurement is incorporated into the model, by taking into account that change measures at different levels have different meaning. Initial ability and change measurement (modifiability) is incorporated and used in this model for item parameter estimation. This model represents an important theoretical development for IRT procedures that proposes item parameter estimation by using a multidimensional model relating item responses to initial ability and modifiability.

The present project is based on standard IRT parameter estimation and CAT procedures. Two separate adaptive item banks are used for the pretest and the post-test respectively. Besides incorporating the latest trends in psychometric test development and DIF analysis, this approach provides a practical solution to the measurement problems traditionally associated with dynamic assessment and the measurement of learning potential. The details of the test construction are discussed in chapter 5, while the theoretical components of IRT and CAT are reviewed in chapter 4.

3.6.4 Conclusion

Dynamic testing and the measurement of learning potential have been actively researched for the last two to three decades. Although the results show much promise, further research, focusing mainly on obtaining empirical validity evidence, is required to fulfil this promise. According to Grigorenko and Sternberg (1998), there is a paucity of published empirical research on the reliability and validity of dynamic assessment. These psychometric elements are more complex in dynamic testing than in conventional testing, but the use of item response theory and computerised adaptive testing holds out the promise of addressing some of these measurement problems (Embretson, 1987; Sijtsma, 1993a, 1993b).

Whereas conventional tests may be better predictors of future school success, new and innovative dynamic assessment procedures may provide better insight into the intellectual development and capacities of disadvantaged examinees. According to Guthke (1993a) intelligence assessment will probably remain the focus of dynamic tests. The idea is that learning potential tests should have the psychometric properties of regular tests, but that their administration procedure should differ, because a training phase is incorporated and improvement in performance monitored. Consequently, not only previously acquired skills and knowledge are assessed, but also the ability to learn (Hamers & Resing, 1993).

Various different approaches to dynamic assessment have been reviewed.

Differences and similarities between these approaches and that of the present project have been noted. In broad, the present project belongs to the psychometric approach to dynamic assessment, with the focus on accurate measurement. Attempts at standardisation of procedures and training improve the comparability of the examinees' final test scores. Nonverbal test content is used in an attempt to provide more equitable multicultural measures of cognitive ability. Lastly, the test-train-retest approach is followed, with both the pretest and post-test being independent CATs.

3.7 PROBLEM AREAS AND POSSIBLE SOLUTIONS: VYGOTSKY REVISITED

Dynamic assessment has not yet lived up to its promise (Grigorenko & Sternberg, 1998). One of the reasons for this could be that there still seems to be some confusion about exactly what has to be measured, and how the measures obtained should be used. Thus far, no standard method of using the scores obtained has emerged. This has led to some confusion among practitioners and potential users of this approach. While dynamic measures are intended to provide additional information about individual's cognitive development, the expectation is also that these measures will be at least equal to standard measures in terms of their predictive validity. The acceptable psychometric properties of the procedures and measures also need to be proven. If this approach is to be used as alternative to, or in combination with the traditional measures of cognitive ability, further research is needed to place it on a surer psychometric footing.

Some of the main practical and technical problems with dynamic assessment are

- the time and difficulty involved in administering the tests
- the high cost because of the level of training required from the examiner
- subjective scoring of some procedures
- problems with the accuracy of the measurement of difference scores (ZPD)
- the lack of standardisation which limits generalisation and comparison
- the practice effect when the same instrument is used in both the pretest and the post-test

- problems in finding suitable criterion measures to provide predictive validity evidence for learning potential measures

Reviewing Vygotsky's theory again and investigating the possibility of using modern test development theory and techniques may prove useful for providing a theoretical and psychometric base for extending the application of dynamic assessment procedures.

While Vygotsky included both the initial level of functioning and the ZPD in his explanation of his theory, the focus has often been limited to either the ZPD or the post-test scores. For instance, while Lidz (1991) initially proposes the use of both fully matured processes as well as emergent developmental processes, she then isolates the ZPD as that which "can also be viewed as a definition of 'potential'" (Lidz, 1991, p 7). In such interpretations, no provision is made for differences in initial level of performance. Because much of the early research in dynamic assessment has focused on low-ability examinees, the seemingly popular or layperson's interpretation of learning potential has become one that focuses *only* on the ZPD or difference score obtained. This makes allowance *only* for the special example used by Vygotsky. In this special example, initial levels of performance are equal and can therefore be ignored during interpretation because the fact that they are equal, means that they do not contribute further to the interpretation. However, this special case does not allow for the interpretation and comparison of scores of individuals where there are differences in the initial level of performance and quite likely also in the ZPD. Vygotsky was only illustrating the principles of his theory by using a special example. He did not elaborate on the problematic interpretation of the majority of cases where both the initial level of performance and the ZPD are likely to be different. The pitfalls and logical consequences of incorrectly extending Vygotsky's special case example to the general and thereby incorrectly referring to the ZPD as defining "learning potential", can be illustrated by means of two practical examples.

Example 1

If someone were to say that a university mathematics professor has no learning potential, quite a few eyebrows would be raised. A person who functions at such a high level should by all accounts be able to cope better than most people with virtually any new learning situation. If the focus is on the ability to learn, then credit also needs to be given for learning that has already been accomplished and which forms part of the learner's repertoire. The professor will probably obtain a very high score on the initial (actual) level of performance and consequently can show only limited improvement. Within the restrictive framework of considering only the difference score as the score that indicates learning potential, it is therefore possible to say that she has very little learning potential. To take the example to the extreme, when selecting someone for further training, this professor could find herself being dropped in favour of a primary school pupil who showed more "learning potential", since the latter's difference score (ZPD) is larger - and this, in spite of the fact that the overall level of performance of the primary school pupil is substantially below that of the professor. It is clear, especially when one acknowledges that measurement of mental development is used in the framework of learning and training environments, that actual developmental level (pretest performance) cannot be overlooked in dynamic assessment. If it is assumed that by learning potential, we mean the potential to benefit from and cope with new learning situations, it is clear that Vygotsky's interpretation of using ***both*** the actual level of development ***and*** the ZPD should be adhered to.

A swimming analogy

The above explanation can be further illustrated in a swimming context. A champion swimmer, who has already equalled the world record, is not expected to improve much. If additional training is provided, a novice swimmer can be expected to show noticeable improvements in performance times. The same cannot be expected of the top-level swimmer, because just maintaining the present high level of performance is already an achievement. Improved times at lower levels are therefore much easier to attain. The higher the level of performance, the more difficult it is to improve upon performance. But a small or even no improvement at a high level of performance

does not mean that there is no potential for performance - only less likelihood of *improved* performance.

The use of ZPD (difference) scores without reference to the level at which they occur, provides incomplete information. Dague (1972, p 71) noted that "learning ability is not independent of education" and indicated that educability is partly a function of previous schooling. Jensen (1963) also indicated that using the gain score alone does not provide useful results. He (1963, p 1) drew the following conclusion:

> *when improvement with practice is thus measured from a different baseline for every subject, the results can be confusing and are often uninterpretable. A subject who is initially good at the task is already near the asymptote of his learning curve and can therefore show but little gain or improvement with practice. The slowest learners can often show the greatest gain. Consequently, correlations between gain scores on various learning tasks and psychometric measures of intelligence usually average close to zero.*

The next factor that requires close attention is the interpretation of (the same or different) difference scores (ZPDs) at different levels.

Example 2

What do equal ZPD scores mean? Using Vygotsky's special case again, where the initial (actual) developmental levels were the same, equal difference scores (ZPDs) could be interpreted as implying similar future mental development. However, in practice such situations rarely occur and few cases can comply with the strict (restricted) conditions set by special examples. When the performances of two or more people have to be compared, their pretest (actual development) performance will not always be equal. To illustrate this point, take another special case where the ZPDs are equal, but where the levels of actual development are different. Let us take an academic example and compare student A who improved from 30 to 40 percent

and student B who improved from 80 to 90 percent after training. Their ZPD scores are both 10 percent, but does that mean that they have the same learning potential? Surely not? Student B should pro rata be given more credit for improving 10 percent at an already high level of performance compared with student A who improved at a level where much more improvement is (theoretically) possible. Surely, when new learning situations are encountered, student B, with such a high level of overall performance, can be expected to show more learning potential in the new context than student A? This example once again illustrates the problem when *only* the difference score (ZPD) is considered in interpreting learning potential. Once again, if Vygotsky's suggestion of using both "what already has been achieved developmentally but *also* for what is in the course of maturing" is adhered to, the problem can be addressed.

A high jump analogy

A high jump analogy will be used to illustrate the point. Suppose a high school is required to put together an athletics team which includes athletes to compete in the high jump event at the next athletics meeting. During the initial selection process, all pupils are asked to participate. At first, their initial levels of performance are assessed and recorded. A special training coach is then brought in to teach them a special technique (commonly known as the "Fosberry Flop") where instead of facing the jump head on, one jumps backwards and head-first over the horizontal bar. This is the first time this technique has been taught and even those athletes who had previously competed in high jump events are expected to improve their performance. The results of three athletes will be used to illustrate the point. Athlete A started at an initial level of 0,75 m and improved to 1,00 m after training. Athlete B started at 1,10 m and improved to 1,35 m after training. Athlete C, who had taken part in the high jump in previous years, started at an initial level of 1,5 m and improved to 1,6 m after training. If these athletes had to be rated (ordered) in terms of their overall performance, what would the rating look like? When only the improvement in performance is considered, athletes A and B are equal and both showed more improvement than athlete C. Athlete C is, however, the best overall performer, and even at an initial higher level still showed an improvement in performance.

Furthermore, even after their improvement in performance has been taken into account, neither athlete A nor athlete B reaches the initial level of performance attained by athlete C. Athlete C should therefore be first on the list. The other two have the same improvement score, but athlete B performs at a higher level initially, which puts him in second place and athlete A in third place. Using equal improvement scores for athletes A and B represents another "special case" to illustrate a point.

The above examples clearly illustrate that both initial level of performance and improvement should be taken into account to provide a fair and more equitable description of likely future performance. If we take as our starting point that learning potential results are supposed to assess the capacity of a person to make progress in a learning, scholastic or academic environment, such tests should predict "the ability to learn". This would imply that both the present level of performance and the ZPD are needed to improve prediction of performance in new learning situations.

For real-life decisions to be made, the information presented seldomly reflects the convenient characteristics of special cases. It is clear that Vygotsky's proposed use of both the actual developmental level (level of initial performance) *and* the ZPD is essential to achieve logical and useful interpretations. In the above examples, using both the actual developmental level *and* the ZPD (difference score) as suggested by Vygotsky, also allows for the interpretation of more general cases and generalising his theory to all ability levels. While Vygotsky also emphasised the social component of interactive learning experiences, his ideas can be applied in psychological test development to enable more equitable assessment of disadvantaged examinees. The distinction between the enrichment (structural) and the psychometric (functional) approaches to dynamic assessment should again be emphasised. Although they are both based on Vygotsky's theory, their aims differ and they therefore emphasise different aspects of his theory. The one is not necessarily better than the other, unless the particular aim is brought into consideration. The enrichment approach is the obvious choice when one has structural change in mind, while the psychometric approach becomes the better choice when the focus is on measurement and comparison.

According to Van der Veer and Valsiner (1991, p 329) “Vygotsky claimed that the concept of the zone of proximal development was particularly helpful to distinguish between normal and retarded children”. This view is reminiscent of Binet’s original intention with his measuring instrument. Binet and Vygotsky not only share the original intention or aim to enable one to distinguish between normal and retarded children, but they also both supported the concept of cognitive ability as something that is dynamic and that can change from the present measured level if additional training inputs are made. Binet referred to the “mental gymnastics” that could be used to improve the level of mental functioning, while Vygotsky’s ZPD is indicative of measurement of improved performance after help has been provided. Whereas Binet did not incorporate these different levels of performance into his measure, Vygotsky specifically refers to different measures, although he did not suggest an instrument to achieve this. However, the measurement implied in Vygotsky’s ZPD concept is clear. The aim of accurate and psychometrically sound measurement of all the scores concerned points toward standardisation of test procedures and of the training or help provided.

Using the scores identified by Vygotsky’s ZPD theory, the following four options are possible:

(1) Only the pretest scores can be used, but that would provide the same type of information as that obtained with standard static tests.

(2) Only the post-test scores can be used, thereby taking the altered performance after training into account. However, this would preclude distinction between two individuals who, although their post-test performance is the same, had different ZPDs (pretest performance different). In a reverse of Vygotsky’s special case, it should be clear that these two individuals are developmentally different, even though their post-test performance is the same. Some researchers have made use of the post-test score adjusted for pretest level of performance. However, the basic premise of measurement of learning potential is that of looking *forward,* starting with present level of performance and making allowance for improved performance following relevant training in the interpretation of developmental level.

(3) Only the difference score can be used, but this assumes comparable pretest performance, which is not always the case. In specific situations where only very low-ability individuals are used or where the ability range is extremely restricted, such an assumption may be valid and justifiable. In cases where examinees with a wider range of initial ability are tested, the assumption of comparable pretest performance is not valid and consequently precludes the use of only the difference score or ZPD.

(4) Lastly, as proposed by Vygotsky, both the initial (actual) level of performance as well as the ZPD can be used together. This option would allow the comparison of people at all initial levels of performance and with different ZPD scores. When it is recognised that Vygotsky's example was an extremely restricted special case by means of which he explained his theory, the extension of his theory to more general cases is a next logical step. Keeping to Vygotsky's proposed way of interpreting the scores allows for this generalisation.

According to Shirley (1992), affirmative action inevitably involves training, implying cognitive modification, which is why tests of potential, which may be equated with modifiability, represent the future of psychometrics in South Africa. Taking into account some of the problems discussed so far, a dynamic assessment procedure that uses modern test theory and test technology in the form of IRT-based computerised adaptive testing, can help address some of the practical and technical problems that have hampered progress in dynamic assessment. Various difficulties seem irresolvable when change measurements are conceptualised in classical test theory, but some of these seemingly irresolvable problems can be resolved by conceptualising change measurement in item response theory. The problems with measurement of the difference score can be addressed with the use of IRT and CAT because the pretest and post-test scores are on the same scale. Furthermore, CAT is time-effective since items are selected from an item bank to match the examinee's estimated ability level. The scoring of performance with IRT and CAT procedures is accurate, objective, standardised and psychometrically sound. Since the training is computerised, it is automatically standardised, which will improve the generalisation of the results. Lastly, using two separate item banks for the pretest and the post-test

will eliminate the effect of memory in post-test performance and thereby also provide a more accurate measure of the level of performance compared with that attained in the pretest.

A new approach to dynamic assessment that uses Vygotsky's interpretation of scores and that makes use of modern test theory and technology in the form of IRT procedures and CAT can address most of the problems that have generally been experienced with dynamic assessment.

3.8 A PROPOSED NEW APPROACH TO DYNAMIC ASSESSMENT AND THE MEASUREMENT OF LEARNING POTENTIAL

For the present project, a new approach to dynamic assessment is proposed that will allow the extension of Vygotsky's ZPD to the broader ability spectrum. At the same time, it will focus on providing a test with the necessary sound psychometric characteristics for such an endeavour. The extension of the ZPD concept to the broader ability spectrum is possible only if Vygotsky's proposed use of both the initial ability level as well as the ZPD score is included in the interpretation of the developmental (ability) level. Use of IRT-based CAT procedures will address the technical and psychometric problems of dynamic assessment.

Because much of the previous research covered in this chapter involved the use of subjects who were at the lower ability levels or educable mentally retarded subjects, the use of Vygotsky's special example was relevant. However, as soon as examinees at higher levels of ability are included in dynamic assessment and the measurement of learning potential, the restrictive conditions of this special example no longer apply. For the more general application of dynamic assessment of general ability including learning ability, both the initial level of performance as well as the ZPD needs to be included in any proposed interpretative framework. The way in which this will be accomplished for the present project is discussed in detail in the section on the construction of the instrument (chapter 5).

3.8.1 Definition of learning potential

For the present project, learning potential is defined as the ability to benefit from (new) learning experiences by using appropriate existing and realisable skills. This definition recognises that a person will also apply existing skills and competencies to any new learning situation. While provision is explicitly made for improved abilities through the development of new skills, the contribution of existing skills to learning is not disregarded, as it often has been. In the terminology of assessment, learning potential is therefore defined as a combination of actual (initial) level of performance and the ZPD score.

Das (1987) emphasises that the extent to which skills can be transferred to new learning situations must depend, to some degree, on initial ability, and warns that differences between disadvantaged and advantaged groups in their level of performance in cognitive ability tests should be recognised and not minimised or explained away by dynamic assessment. This also underscores the need for using both initial level of performance and the ZPD score for the interpretation of cognitive development.

3.8.2 Operationalisation

For the present study it is proposed that techniques based on item response theory methods combined with computerised adaptive testing be used. This approach has been suggested by Embretson (1987) and Sijtsma (1993a, 1993b) as providing a solution to many of the problems that have been encountered in dynamic assessment. This approach has various advantages (Sijtsma, 1993b):

(1) Because separate item banks are used for the pretest and the post-test, items are not repeated and memory does not confound test performance.
(2) Because both the pretest and the post-test items are adapted to the

performance level of the examinee at the time of testing, ceiling or floor effects are unlikely to occur.

(3) Because measures are obtained on the IRT latent ability scale, the difference between post-test and pretest scores reflects change in performance (latent ability) due to training and not change due to different difficulty levels of the two tests. If training is ineffective for a person, then the same theta value (latent ability level) will be estimated on both the pretest and the post-test.

(4) Reliability for each measurement is estimated separately. Hence a more refined assessment of change is possible than with the number-correct scale of classical test theory.

The measure of interest is a combination of the initial level of performance and the ZPD, as proposed by Vygotsky. This will give an indication not only of the level of present performance, but will also allow for the size of the ZPD to indicate possible improvement over and above the initial level of performance. Sternberg's (1991) reservation that a numeric value attached to the ZPD will lead to it being interpreted as a fixed entity is noted. Both the actual level of performance as well as the ZPD should be seen as malleable and subject to (further) change following intervention or instruction. This is a crucial basic tenet of dynamic assessment.

IRT-based CAT is particularly suitable for learning potential testing (Embretson, 1987, 1992; Sijtsma, 1993a, 1993b). Tailoring the pretest and the post-test to each individual's performance level not only leads to shorter tests having the same precision for each individual but can also be expected to motivate the examinee during the training programme.

3.9 CONCLUSION

If the ZPD is to be incorporated into the overall interpretation of an individual's cognitive development, it is imperative that it should be accurately measured. While the measurement of difference scores has been considered problematic for many years, item response theory provides a useful and practicable solution to this problem

in the context of dynamic assessment and the measurement of learning potential (Embretson, 1987; Sijitsma, 1993a, 1993b). The use of CAT addresses the issue of length of testing time, since it provides for both testing efficiency and measurement accuracy.

Biesheuvel, who is regarded as the "doyen of psychologists and particularly of cross-cultural psychologists in South Africa" (Mauer & Retief, 1987, p iii) is quoted by Cronbach and Drenth (1972, p 477) as having said that "I think we psychologists ought to have the guts to stand up for the instruments which we have produced which we know will do a better job [than other methods] of sorting out those people who can take advantage of the very limited educational opportunities that are available". Although this statement was made almost 30 years ago, it is still relevant and applies especially to the emerging field of dynamic assessment and the measurement of learning potential. According to Foxcroft (1997a, 1997b), users and developers of psychological tests in South Africa face numerous challenges during this time of transformation and nation-building. She emphasises that the development of culturally relevant tests is paramount to enhancing the practice of psychological testing and assessment. "Assessment used to identify individuals with potential so that they can be linked into developmental programs can, however, be a very effective and ethically defensible approach to redressing past imbalances" (Foxcroft, 1997b, p 234).

The use of IRT and CAT procedures can provide the sound psychometric base needed to put forward a technically sound dynamic assessment instrument for the measurement of learning potential.

CHAPTER 4

ITEM RESPONSE THEORY AND COMPUTERISED ADAPTIVE TESTING

4.1 INTRODUCTION

The development of item response theory (IRT) over the last 30 to 40 years has brought about significant changes in psychometric theory and test development. These changes have been incorporated into a general new set of rules of measurement that are fundamentally different from the old rules of classical test theory (Embretson, 1996; Reckase, 1996). This chapter will show how IRT and computerised adaptive testing (CAT) can be used to address some of the most urgent problems currently experienced in dynamic testing. These include problems relating to the measurement of difference scores, extended testing times and possible floor or ceiling effects, as indicated in the previous chapter.

The main applications of IRT are in test construction, test equating, detection of DIF and adaptive testing. Adaptive tests of today are based on sophisticated IRT-based test theory developments and make use of powerful computer technology. Interestingly, even in these modern developments, Binet's legacy still features prominently, since his test can be regarded as the first adaptive test to be developed (Weiss, 1983b). The fact that CAT has been described as "A good idea waiting for the right technology" (Reckase, 1988), therefore seems most appropriate. Binet's test had several key features of current adaptive tests, namely:

- A variable entry point was used, depending on the examinee's ability level as estimated by the examiner.
- Items were scored during administration and the results used for further branching and selection of additional items.
- The test featured a variable termination criterion which resulted in different individuals receiving varying numbers of items. The test was terminated when a

ceiling level was reached (Weiss, 1983b).

Statistical procedures form an integral part of psychometric test development. At item level, item analysis is performed to identify the psychometric properties of items, helping to identify good and poor items and investigating DIF in items for various groupings. The characteristics of tests can also be investigated by means of statistical analysis. For many years, classical test theory (CTT) was the only approach in the statistical analysis of items and tests in test development procedures. Recent statistical and test theory developments in IRT and CAT provide new features to improve the psychometric characteristics of tests in general and of dynamic assessment instruments in particular. Before discussing the principles, theoretical concepts and main features of IRT, a brief overview of the main features of CTT is given.

4.2 GENERAL FEATURES AND LIMITATIONS OF CLASSICAL TEST THEORY

Classical test theory item analysis provides information on the difficulty of the item as well as the extent to which it contributes to overall test performance. The item characteristics used in classical test theory are:

- **The p-value**
  This is the proportion of examinees who select the correct alternative. It is also referred to as the “item difficulty”.

- **The discrimination value**
  This value is the point-biserial correlation between the item score and the test total, indicating to what extent someone’s performance on the item is commensurate with overall test performance.

These two indices are used to evaluate the available items developed for a test in order to construct a psychometrically sound instrument for a given purpose. Other characteristics that are considered in the CTT approach to test development are the

examinees' mean performance, the standard deviation of their scores and the skewness of the distribution of their scores.

However, one of the drawbacks or limitations of CTT is that the item parameters defined in CTT are dependent upon the sample of subjects to whom the items were administered and are relative to the characteristics of the test and examinees. For instance, the p-value is relative to the ability level of the group to which the items are administered. The same item given to a high-ability group will have a relatively high p-value (ie a large proportion of examinees selecting the correct alternative), whereas in a low-ability group it will have a relatively low p-value (ie a small proportion of examinees selecting the correct alternative). The discrimination value is relative to the homogeneity of or distribution of ability levels of the examinees in the sample, as well as to the subject matter homogeneity of the items in the test and the distribution of p-values of items in the test (Warm, 1978). It is evident that the mean, standard deviation and skewness indices will also vary according to the characteristics of both the examinees and the test. In CTT, the total number-correct scores are used to obtain the scores of individuals, and test scores are therefore dependent on the difficulty levels of the items included in the test. Furthermore, true scores cannot be directly measured and must be estimated from observed scores, where the observed score is considered to consist of a true score and an error score:

$$X_o = X_t + X_e$$

with X_o the observed score
X_t the true score
X_e the error score

The error score represents the difference between true performance on the construct of interest and the observable data (Hambleton & Slater, 1997). The reliability of observed scores is influenced by factors such as the standard deviation of the test and the difficulty and discrimination values of the items. Consequently, it also depends upon the examinees' particular abilities as well as the characteristics of the test.
General limitations and problems with CTT that have been noted (Hambleton & Swaminathan, 1985; Sijtsma, 1993a; Weiss & Yoes, 1991) are as follows:

- CTT does not provide the means to empirically establish the measurement properties of test scores.
- The average ability level and the range of ability scores in a sample influence the values of the item statistics.
- The abilities of different examinees cannot be compared independently of a specific test and test scores are item-dependent.
- Precision of measurement of a test is assumed to be identical for the whole range of the scale - that is, the standard error of measurement is assumed to be the same at all levels of the ability under consideration.
- There is no base for predicting how an examinee might perform on a particular test item.
- In CTT, procedures for technical problems such as identification of biased items or equating of test scores are limited and difficult to handle.

Two of the most noted problems in CTT are that measurements are not invariant with respect to the instrument (or items) used and that the properties of instruments (items) are not invariant with respect to the persons being tested (Muniz, 1998). This makes comparison of test scores particularly problematic.

Dynamic assessment is characterised by the test-train-test model of testing with a specific focus on the comparison of the two test scores obtained. Instead of using the same test twice, the pretest and post-test could be designed to be different but parallel tests (Lord & Novick, 1968), or equivalent tests according to some other, more liberal definition of equivalence (Sijtsma, 1993b). If parallel tests are used, memory does not affect test results, and the difference score will reflect change rather than a variable confounded with memory. However, in the CTT model, regression effects, the unreliability of change scores, ceiling or floor effects and unequal intervals on the scale of measurement remain problematic issues. The problems in CTT apply both to the comparison of scores across different administrations of the same test and to the comparison of results on two different tests (ie two possibilities for the pretest and post-test in the case of dynamic testing). Measurement of difference scores has been identified as one of the key problem areas in dynamic testing when viewed from the

classical test theory perspective.

The development of IRT has provided solutions to many of the problems experienced in CTT. One of the advantages of IRT is that it resolves the problem of sample dependency by providing ability parameters which are invariant over samples of subjects, and not dependent on the particular items that are administered. This allows for more accurate comparison of different test scores of the same individual as well as comparison of test scores between individuals - an important feature for dynamic assessment. Some background on the history and development of IRT is provided next.

4.3 A BRIEF HISTORY OF IRT

IRT has been described as the most significant development in psychometrics in many years - "perhaps to psychometrics what Einstein's relativity theory is to physics" (Warm, 1978, p 11). IRT is characterised by a mathematical function that represents the relation between an underlying (latent) ability and the probability of a correct response, enabling estimation of an individual's trait level from observable responses. The adoption of IRT has been surprisingly slow. One of the reasons why this theory has not been used to the extent that was first anticipated, is that it is mathematically and statistically complex. However, the continued development of computer technology with the availability of ever more powerful microcomputers and of test development packages that incorporate IRT procedures, has improved the accessibility of these procedures for test developers. User-friendly test development systems have been made available commercially (Assessment Systems Corporation, 1995), which gives increasingly more users access to these procedures. As in many other sciences, less than a full understanding of the mathematical theoretical underpinnings need not prevent researchers from applying the theory and incorporating its benefits in research and practical applications.

Although the history of IRT dates back to the early 1940s, it only became useful to test constructors with the development of computer technology in the 1960s. Early

progress was slow because of both the mathematical complexity of the theory and lack of availability of suitable computer programs. A brief summary of the historical development of IRT (Assessment Systems Corporation, 1989, 1995; Hambleton & Slater, 1997; Hambleton & Swaminathan, 1985; Hambleton & Zaal, 1991; Lord; 1980; Lord & Novick, 1968; Sands, Waters & McBride, 1997; Van der Linden & Hambleton, 1997; Van Tonder & Claassen, 1992; Warm, 1978; Weiss, 1983a, 1983b) is provided below.

- In the early 1940s, several researchers became involved in groundwork related to IRT. Relationships were derived between IRT parameters and CTT parameters, providing an initial way for obtaining IRT parameter estimates.
- The work of Ferguson and Lawley on latent trait theory is also regarded as important. Mosier described relationships between latent trait concepts and psychophysics while Guttman developed the basics of latent trait theory to solve scaling problems in attitude measurement.
- In 1952, Lord published his PhD thesis in which he presented IRT as a model or theory in its own right, calling it item characteristic curve theory. Lord is generally regarded as the father and founder of IRT. However, he stopped work on IRT for 10 years shortly after publishing his thesis, apparently because of a seemingly unsolvable problem concerning the assumption that the item response function takes the form of the normal ogive (cumulative graph).
- In 1960, Rasch published his one-parameter sample-free model which led to considerable research during the next decade.
- In 1965, Lord conducted a massive study with a sample exceeding 100 000, which led to the solution of the earlier problem and to his taking up his work on IRT once again.
- In 1968, Lord and Novick published a psychometrics textbook, including chapters by Allan Birnbaum on the mathematical underpinning of the two- and three-parameter normal ogive and logistic models.
- The availability of computer technology in the late 1960s stimulated the development of IRT test theory and computerised adaptive testing
- In 1970, Urry completed a PhD thesis, concluding that the three-parameter model best described the real world for multiple-choice items. Since then, the

three-parameter model has received most research attention, including work by Frederic Lord, Vern Urry and David Weiss. The LOGIST program for item parameter estimation was made available in 1976, using Lord's description of parameter estimation methods.

- In 1977, Lord changed the name of his model from item characteristic curve theory to item response theory. The 1970s saw IRT becoming the dominant topic for study by measurement specialists.
- In the 1980s, the use and further development of IRT continued with books related to the topic being published making the information more accessible to measurement specialists and other users, allowing them to solve practical testing problems with IRT procedures. The first commercially available computer package for incorporating IRT and CAT procedures in test development became available. In South Africa, development of the first computerised adaptive psychometric test was started in the late 1980s.
- The 1990s saw continued large-scale use of IRT procedures in test development and other psychometric applications such as test equating, DIF analysis and large-scale CAT program development. The first South African computerised adaptive psychometric test was published by the HSRC in 1992. The present project, which commenced in 1993/4, represents the second such an endeavour in South Africa as far as is known.

The development of computer technology and the availability of powerful personal computers heralded the era of combining the developments on the statistical front with the available computer technology for computerised adaptive testing (see 4.6). The implementation of IRT and CAT in test development and use have resulted in major changes in the way that psychological testing is done (Hambleton, 1994; Hambleton & Slater, 1997; Weiss, 1983b, 1985).

4.4 PRINCIPLES AND THEORETICAL CONCEPTS OF IRT

4.4.1 Introduction

One of the most useful features of IRT is that examinee estimated ability level and item difficulty level are put on the same scale. Another important feature is that in IRT,

item and test statistics are dependent neither on the examinees' characteristics nor on the other items in a test. It therefore becomes possible to describe the characteristics of a test before it is administered, allowing the construction of more efficient tests. It also facilitates improved DIF investigations and makes possible computerised adaptive testing.

IRT in its most basic form postulates that a single ability underlies examinee performance on a test and that the relation between this ability and the probability of a correct response on an item is a monotonically increasing curve (Hambleton & Slater, 1997). IRT models specify a function depicting the relation between the probability of correct responses of an individual to a test item and the individual's level on the latent trait. There are three broad categories of IRT models, namely the one-, two- and three-parameter models. The item parameters of each of the models are used to describe the item characteristic curve (ICC) or item response function (IRF), which represents the relationship between ability and probability of a correct response. Each of the models uses a different number of descriptors or item characteristics to describe the ICC. The three main models are distinguished by the number of item characteristics incorporated. The different models are discussed in 4.4.2.

In general when using IRT, the following is assumed (Hambleton & Swaminathan, 1985; Ree & Jensen, 1983; Warm, 1978):

- If the examinee knows the correct answer to the item, he or she will answer it correctly.
- The item response function (IRF) takes the form of the logistic ogive, which is an approximation to the normal ogive.
- Items are unidimensional, which means that the items measure one and only one area of knowledge or ability and that performance is thus attributable to a single latent ability/trait.
- There is local independence between items, which means that the probability of an examinee getting an item correct is unaffected by responses to other items. Local independence does not mean that there is no correlation between the items.

- Tests are not administered under speeded conditions.

These assumptions were also made for the present project. Where practical, these assumptions were built into the test programme as in the case of unspeeded administration and the construction of items to function independently. Unidimensionality was empirically verified.

4.4.2 IRT models

The three general IRT models vary in terms of the item characteristics they include. The one-parameter model is based on only the item difficulty value. This model is the simplest in that only the difficulty level (b-value) of a test item and the examinee's ability level is taken into consideration. This model, while allowing differences between items in terms of their difficulty level, does not allow differences in other characteristics of items (Weiss, 1983a). In the two-parameter model, the item discrimination (a-value: the rate of change of the probability of a correct response as a function of the underlying trait level), is also considered, together with the difficulty level of the item. When multiple-choice items are used and items can be answered correctly by guessing, the third parameter, namely the pseudo-chance parameter (c-value), can be added to form the three-parameter model. In the one-parameter (Rasch) model, only the b-value varies while the a-value and c-value are taken as constants (usually the a-value is set at 1,00 while the c-value is set at 0,0). In the two-parameter model, the b-value as well as the a-value varies, while the c-value is still set at 0,0. The three-parameter model allows all three parameters (a, b and c) to vary. The parameters and the range of possible values that they can take are explained in 4.4.3.

The choice of the appropriate IRT model is determined by a combination of different factors such as the size of the available sample, the quality of the data, the choice of estimation procedure and the availability of computer programs. For instance, sample sizes smaller than 200 would dictate the use of the one-parameter model. In the three-parameter model, large samples are needed (N > 1 000), with sufficient

numbers of low ability examinees to allow accurate estimation of the pseudo-chance (guessing) index. Other factors that affect the appropriateness of a particular model could be the type of items and cost factors. For example, calculations for the three-parameter model are much more extensive and therefore also more costly. Schoonman (1989) proposes that the choice of psychometric model should be based on practical considerations, even if the model does not fit the data entirely.

The three-parameter model was chosen for the present project. The reasons for this choice are that multiple-choice items were used (and c-values therefore > 0), and that large enough samples were available to allow this model to be used. Furthermore, sufficient numbers of low-ability examinees were included in the sample to allow estimation of the c-parameter. The MicroCAT system (Assessment Systems Corporation, 1995) for item analysis, which runs on a personal computer, was available with no additional cost for running the three-parameter item analyses programs. Specific assumptions of the three-parameter model that were also made for the present project are unidimensionality, the invariance of item parameter estimates and the invariance of ability parameter estimates (Hambleton & Swaminathan, 1985), all of which were empirically investigated.

The advantages of IRT models will only be obtained when there is a close match between the model selected for use and the test data (Hambleton & Swaminathan, 1985). It is therefore important to determine if the test data satisfy the assumptions of the test model of interest and whether the expected advantages are in fact obtained. The three-parameter model is the most general of the unidimensional models in common use and is generally recommended as the best model for multiple-choice items.

With respect to evaluating the fit between an item response model and a set of test data, it is necessary to design and implement a wide variety of analyses, to interpret the results and to judgmentally determine the appropriateness of the intended application. It is advisable to accumulate a considerable amount of evidence to assess the appropriateness of a particular item response model. Analyses should include investigations of model assumptions, the extent to which desired model

features are obtained and comparisons between model predictions and actual data. Hambleton and Swaminathan (1985) consider unidimensionality to be the most important assumption to verify. To investigate the invariance of item parameter estimation, plots of item parameter estimates obtained in two groups can be compared but no general required format for “baseline plots” is provided. The shortage of computer programs to carry out the necessary analyses is also a limiting factor. For the present project, unidimensionality and the invariance of item parameters were empirically investigated. The results are discussed in chapter 5 where test construction information is provided. Time limits were not used in the administration of test items for item analysis purposes, thereby complying with that assumption.

To implement an IRT model, the parameters of the items and the trait levels of the individuals need to be estimated. The first phase involves the estimation of the item parameters. Although this process of estimating item parameters also involves the estimation of the trait level parameters of those individuals to whom the items were administered, the latter parameters are usually of secondary importance and not used for individual assessment. Once the item parameters have been calculated, the ability levels of (other) individuals can be estimated.

4.4.3 The item characteristic curve (ICC) and item parameters

Much of educational and psychological measurement concerns underlying (latent) variables of interest and involves determining how much of such a latent trait a person possesses. A correct response depends on the characteristics of the item and the ability of the person. The probability of a correct response is expressed as a mathematical function of examinee ability and item characteristics - also known as the item characteristic curve (ICC). The ICC graphically represents the basic tenet of IRT with the ability level of examinees plotted on the X-axis, against the probability of answering an item correctly. Each examinee is considered to have an ability score which places him or her somewhere on the ability scale. An examinee’s ability is denoted by the Greek letter theta (t). At each ability level there is a certain probability that an examinee with that ability will answer the item correctly. This probability is

indicated by P(theta) - or in symbol form P(t). In typical items, this probability is smaller for individuals with low ability than for those with higher ability levels. Therefore, if the probability function P(t) is plotted against ability level, the result is the typical S-shaped form of the ICC. Each item will have its own ICC. The typical form of an ICC is illustrated in Figure 4.1.

FIGURE 4.1 AN EXAMPLE OF AN ITEM CHARACTERISTIC CURVE

The ICC curves are strictly monotonic functions - in other words, as the values along the X-axis increase, so too do the corresponding values on the Y-axis. This shape is known as an ogive (Warm, 1978). The ogive with which we are concerned is the normal ogive, representing the cumulative frequency distribution of the normal distribution. However, no algebraic function can be found to describe this ogive, which makes it extremely cumbersome to work with mathematically. The logistic ogive, which is so close to the normal ogive as to be hardly distinguishable graphically, on the other hand, is easier to work with. The logistic ogive is therefore substituted as a convenient and very close approximation to the normal ogive (Warm, 1978). Characteristically, the ogive always rises from left to right, is never completely horizontal and never goes down. The point where the ogive changes from being convex upward to concave upward is called the “inflection point”, which is the point where the *slope* of the ogive is at its maximum. The distinctive characteristics of the

three parameters that determine the shape of the ICC are provided next (Baker, 1985; Hambleton & Swaminathan, 1985; Lord, 1980; McBride, 1997; Warm, 1978; Weiss, 1983a, 1985).

- **The b-parameter (difficulty value)**

One way in which ICCs differ from each other, is in the horizontal location of the inflection point on the ability or theta axis, which indicates the difficulty level of the item. The horizontal position of the inflection point is called the "b-parameter" or "*b-value*", reflecting the *difficulty level* of the item. The b-value represents the point on the ability scale where the probability of a correct response is 0,50 (ie a 50% chance of getting the item correct). The larger the b-value, the more difficult the item will be. Although b-values theoretically range from $-\infty$ to $+\infty$, typical b-values range from -2,5 to +2,5. A b-value of -2,5 indicates a very easy item and a b-value of +2,5 indicates an extremely difficult item.

- **The a-parameter (discrimination value)**

The second parameter of the three-parameter IRT model is the "a-parameter" or "*a-value*" which is related to the slope of the ogive at the inflection point (b-value) and indicates the precision of measurement at the particular difficulty level of the item. The a-parameter is called the "*discrimination index*" of the item response function. The steeper the slope of the curve, the greater the discrimination will be, but the smaller the range of discrimination. Theoretically the a-value can range from 0 to $+\infty$, but a-values typically range between 0,0 and 2,0 with values exceeding 2,0 seldom being found. Items with a-values below 0,5 are insufficiently discriminating for most testing purposes. With a high a-value, the item has a steep ICC and discriminates well, but over a small range of theta. The larger the discrimination value, the better the item can separate examinees into different ability levels in the region of the item difficulty level.

- **The c-parameter (pseudo-chance level)**

The ICC has two asymptotes that the ogive approaches at its extremes. The upper asymptote is located on the vertical axis at 1,00, while the lower asymptote never quite reaches 0,00. The lower asymptote is called the "c-parameter" or the "*c-value*" and reflects the probability that a person with very little of the particular ability will answer the item correctly. This value is also known as the "*pseudo-chance parameter*", so called because most items used for the three-parameter model are of the multiple-choice format, which makes guessing possible. This parameter is included in the model to account for item response data from low-ability examinees, where guessing is a factor in test performance. Theoretically, c-values range from 0,0 to 1,0. The general recommendation is, however, that items with c-values of 0,30 or greater should not be used. According to Baker (1985), a side effect of using the guessing parameter is that the definition of the difficulty parameter is changed. Instead of the b-parameter being described as the position on the ability scale at which the probability of a correct response is 0,5, this probability becomes the value halfway between the value of c and 1,0. Thus, the difficulty parameter then defines the point on the ability scale where the probability of a correct response is halfway between the floor value (c) and 1,0.

For the three-parameter model in particular, stable and accurate estimation of the item parameters requires large numbers of subjects over a broad range of ability. It is generally recommended that samples of at least 1 000 be used for the three-parameter model (Baker, 1985; Hambleton, 1994; Hambleton & Swaminathan, 1985; Weiss, 1983a) The accurate estimation of the c-parameter also requires large numbers of subjects at (very) low ability levels. Nevertheless, the advantages offered by these models far outweigh the efforts involved in using them, despite drawbacks concerning sample sizes required and the mathematical/statistical complexity of the theory.

4.4.4 The test information function

Using IRT, the main purpose is to estimate the value of an examinee's ability parameter. The variance of estimates around the value of this parameter indicates the

precision with which a given ability level can be estimated (Baker, 1985). The accuracy of measurement at a particular ability level is related to the amount of information at that particular ability level. Depending on the values of the item difficulty (b) and the item discrimination (a) values of an item, a specific amount of information can be obtained with a particular item. Thus, when the IRT parameters for a set of items are known, the item information values for each item can be calculated at each possible ability level, indicating how precisely the item measures at particular ability levels. Because ability is on a continuum, information will also be provided on a continuum. The amount of information based upon a single item can be computed at any ability level. In general, an item measures ability with the greatest precision at the ability level corresponding to the item's difficulty value. The item information curve peaks at the difficulty of the item, and the degree of precision at any point on the continuum is related to the item discrimination or a-value (Weiss, 1985). When the amount of information supplied by a *set* of items is plotted against ability, the graph of the test information function is obtained.

The test information function is the sum of the information functions of all items included in that particular test. It indicates how well the test measures over the whole range of ability scores. If the amount of information at a particular level is large, this means that the ability of an examinee whose true ability is at that level can be estimated with precision and that the estimates will be reasonably close to the true ability level. If the amount of information is small, this means that the ability cannot be estimated with precision and the estimates will be widely scattered about the true ability.

The desired shape of the test information function depends upon the purpose of the test. More information is required at levels where increased accuracy of measurement of ability is needed. The information function indicates how well each ability level is estimated, irrespective of the distribution of examinees over the ability scale. A flat or horizontal information line indicates that all ability levels can be estimated with similar precision (Assessment Systems Corporation, 1989, 1995; Hambleton & Swaminathan, 1985; Warm, 1978; Weiss, 1983a). Typically, however, the information function does not indicate equally precise measurement at all levels.

To attain more accuracy at levels where the information function is lower, more items may need to be administered in regions where less information is available. This is particularly true for computerised adaptive tests (see 4.6), where accuracy of ability estimation is often used as the termination criterion. For interpretation of the test information function, the reciprocal relationship between the amount of information and the variability of the ability estimates should be noted. To translate the amount of information into a measure of accuracy of estimation (standard error of estimation), the reciprocal of the square root of the amount of test information is used. The formula used to depict this relation is

$$SE_{(t)} = \frac{1}{\sqrt{I_{(t)}}} \qquad \text{(Baker, 1985)}$$

This equation directly depicts the relation between accuracy of ability estimation and the amount of information available at that ability level. If it is known what the level of information available at a particular ability level is, then the accuracy of measurement of ability at that level can be calculated in the form of the standard error of estimation (SE).

4.4.5 Conclusion

The specific features and characteristics of IRT discussed in this section indicate that this approach can contribute to the improvement of psychological test construction. Factors such as the use of item parameters that are sample invariant and the graphic representation of item characteristics in the form of the ICC contribute to effective test construction procedures. Furthermore, having the amount of information and thereby also the accuracy of measurement available at each ability level, will further improve test construction.

4.5 ADVANTAGES OF IRT

The central feature of IRT is the specification of a mathematical function relating the probability of an examinee's response on a test item to an underlying ability. In IRT the item parameters are not dependent upon the ability level of the examinees responding to the item. This group invariance of the item parameters is a powerful feature of IRT, reflecting that the item parameters are a property of the item and not of the group that responded to the item. This does not mean that item parameter estimation using different groups will yield identical numerical values for the item parameters, but the values obtained with different groups should be of a similar magnitude (Baker, 1985). The group invariance of the item parameters also reflects the feature that the item parameters are independent of the distribution of examinees over the ability scale.

Another basic feature of IRT is that the examinee's ability is invariant with respect to the items used to determine it. The ICC spans the whole ability scale so that any set of items can be used to estimate an examinee's ability. When the difficulty of items used is not located near the examinee's ability, the standard error of the estimates may be quite large. The optimum set of items for estimating an examinee's ability would have all its item difficulties equal to or close to the examinee's ability parameter and have items with large a-values (high item discrimination indices). Different sets of items will yield values of estimated ability near the examinee's actual ability level. It has to be understood that it is impossible to obtain an exact value of an examinee's ability - at most, we can obtain an estimate of it (Baker, 1985).

Having used IRT to analyse items, it becomes possible to construct a test and include in it items of a particular difficulty and discrimination level. This also simplifies building of parallel tests for simultaneous administration with psychometric properties that are virtually equivalent.

4.5.1 General advantages of IRT over classical test theory

The development of IRT has addressed a number of psychometric problems or

limitations that were generally experienced in the use of CTT. The following are some of the main advantages of IRT over classical test theory as summarised from Baker (1985), Hambleton (1994), Hambleton and Slater (1997), Hambleton and Swaminathan (1985), Sijtsma (1993a), Van der Linden and Hambleton (1997), Warm (1978) and Weiss (1983a):

- The difficulty level of the items and the ability level of examinees are on the same scale.
- Item parameters are independent of the population of examinees.
- Ability parameters are independent of the particular choice of items.
- Item characteristics can be calculated beforehand, with the result that tests consisting of items of a particular nature can be constructed to suit the purpose of the particular test or to yield tests with desired characteristics (ie tests that function at a particular level, or better matched parallel test versions).
- The precision of ability estimates can be determined at each ability level, from the test information function.
- Item banking means having available a large pool of items for which the item indices are known and for which the difficulty level of the items is depicted on the same scale so that items are directly comparable. Tests can be specifically constructed on the basis of the characteristics of the available items.
- Equating is the technique of depicting ability measures obtained from different tests for the same ability on one common scale. Furthermore, item indices obtained from different subgroups can also be equated.
- Adaptive testing becomes possible, whereby items that match the ability level of the examinee are administered interactively, which leads to greater reliability using fewer items than in a conventional test.
- IRT provides more powerful procedures for detection of DIF.
- The functional relationship that IRT models specify between observable responses and unobservable trait levels allows computer simulation to investigate the behaviour of models and their applicability to a wide range of measurement problems.

The advantages that have been mentioned thus far are for psychological tests in general. However, IRT also holds specific advantages in the field of dynamic testing for the measurement of learning potential.

4.5.2 Advantages of IRT for learning potential measurement

Sijtsma (1993a) considers IRT to be an excellent choice for applications such as change measurement. He suggests learning potential research as a most likely application of IRT because learning potential testing involves training with the purpose of improving test performance, thereby involving measurement of change. Measurement of change is a well-known and problematic issue in psychometrics (Lord, 1967; Embretson, 1991). Consequently, one of the main problem areas in dynamic testing has been the measurement of difference scores, or the measurement of change between pretest and post-test scores (Grigorenko & Sternberg, 1998). Problems with the measurement of change have long featured in psychological testing. In 1962, a special three-day conference was held to discuss various aspects of problems in measuring change (Harris, 1967). Lord (1967) who also participated in the conference, proposed elementary models for measuring change. At the time of the conference, IRT was still in the beginning stages of development. At present it is generally acknowledged that measurement of change can be better addressed with IRT models compared with any other means previously employed (Embretson, 1987, 1992; Sijtsma, 1993a, 1993b). An important advantage of IRT is that because of the improved test equating, the problem with measurement of change and difference scores of dynamic testing can be better addressed than with CTT. Item difficulties and ability measures are on the same scale and are comparable. Comparison of the ability levels of persons is further simplified because measurement on the theta scale is independent of the items used for measurement.

According to Sijtsma (1993b), in order to improve the quality of the measurement of change, different pretests and post-tests that comply with an IRT model should be used. This will eliminate several problems with measuring change. Firstly, test sessions can be regarded as independent, and memory will not confound test

performance. Secondly, because both the pretest and the post-test are adapted to individual performance levels at the time of testing, ceiling (or floor) effects will not occur or will be modest. Thirdly, because measurement takes place on an interval scale, equal change scores can be considered to reflect equal change, with the typical distortions of the number-correct scale being eliminated. Fourthly, there is no need to construct parallel tests - the only requirement being that both tests measure the same ability. Effective training is indicated by systematic improvement in theta values on the post-test. Ineffective training, or the absence of change, will be indicated by the same theta-value estimate on both the pretest and the post-test.

The main advantage of IRT for learning potential measurement lies in the improved accuracy of measurement of difference scores, as well as improved means to compare scores of the same or of different examinees, since in IRT, all measures are on the same scale.

Embretson (1991, 1995), developed a latent trait model for the measurement of learning and change. Her proposed multidimensional Rasch model for learning and change (MRMLC) is based on contemporary IRT as the foundation for measuring learning ability. It provides some answers to seemingly unsolvable problems with change measurement involving a new IRT-based model with modifiability conceptualised in an IRT framework, relating it to the latent ability. Expected changes in performance for a given level of learning ability will depend on both the level of initial ability and the item difficulties. The purpose of this new psychometric model is to function as a latent trait IRT model for item parameter estimation, by using a multidimensional model relating item responses to initial ability as well as modifiability. The focus is therefore on predicting item responses by using both the initial level of performance and modifiability. In IRT, the learning ability that is being measured, relates to a change in the latent ability. In the MRMLC model (Embretson, 1991, 1995), the possibly unequal impact of underlying abilities on latent response potential is shown, acknowledging that modifiabilities at different levels are not directly comparable. The difference score remains the essential concept.

The present project uses a different approach in that a standard three-parameter IRT

model is used for item parameter estimation procedures. The focus then turns to the measurement of initial level of ability as well as modifiability. The aim is not to predict performance on specific items or to develop a new psychometric model, but to obtain more accurate measurement of the measures concerned. The present project differs from Embretson's model in that learning potential is defined as a combination of pretest and difference scores and the difference score is therefore not the essential concept. However, it shares the view of Embretson's (1991, 1995) model that modifiabilities at different levels are not directly comparable.

Guthke (1992) and his co-workers have used IRT item analysis procedures in their dynamic assessment research. They have incorporated a form of dynamic assessment where the feedback provided and the next item administered are based on the particular distractor selected. Thus each examinee will not necessarily receive the same set of items. The items administered and the feedback provided are dependent upon the particular answer selected. This strategy, although administered by computer and adaptive to some extent, is not fully-fledged IRT-based CAT. IRT-based CAT provides additional features that can further contribute to more effective test procedures for dynamic assessment of learning potential.

4.6 COMPUTERISED ADAPTIVE TESTING (CAT)

Psychometrists have long been concerned with the fact that in standard tests, many items in a test are not appropriate for a given examinee. In standard tests, items are usually sequentially ordered from easy to difficult. The easy items are appropriate for low-ability examinees, but offer no challenge to high-ability examinees, who probably become bored by them. The more difficult items are appropriate for high-ability examinees, but are likely to frustrate low-ability examinees, who will probably guess at answers for items that are too difficult for them. With IRT and personal computer technology came the possibility of "tailored testing" or "adaptive testing" by computer, so called because it allows the "tailoring" of the test to the ability of the examinee (Weiss, 1983a). The adaptive testing strategy requires an item bank to store large numbers of items, a computer program to select and present items and process the responses and an IRT model to estimate theta and to compare the thetas obtained (Baker, 1985; Hambleton & Swaminathan, 1985; Warm, 1978; Weiss, 1983a).

Computerised adaptive testing (CAT) is one of the most exciting developments that flowed from IRT. It is based on the premise that "an examinee is measured most effectively when the test items are neither too difficult nor too easy for him" (Lord, 1980, p150). CAT involves the interactive selection of items during test administration so that item difficulty is matched to the examinee's (estimated) ability level throughout the test session. The item selected each time, is the one that provides the most information at the examinee's current estimated level of ability. A test thus "adapted" to each individual examinee's ability level, results in various advantages such as more precise measurement and higher examinee motivation (Hambleton & Swaminathan, 1985; Lord, 1980; Weiss, 1983a, 1983b). Although, as mentioned earlier, Binet's first test can be regarded as an individually administered adaptive test, this mode of testing was not fully explored until the 1960s when theoretical developments and the availability of computer technology allowed the development of IRT-based CAT.

CAT uses a bank of precalibrated items, which were analysed with IRT and whose statistical characteristics are known so that suitable items may be selected. The interactive selection of appropriate items from the item bank throughout the test is possible because the difficulty level of items and the examinees' ability level are on the same scale. Items at or close to the examinee's estimated ability level improve measurement accuracy. A statistic indicating the precision with which each examinee's ability is estimated is provided and can be used as a termination criterion in CAT. Adaptive tests appear to be more intrinsically motivating for low-ability examinees (Weiss, 1983c), an advantage that is particularly relevant for learning potential assessment, which is usually aimed at low-performing examinees.

Practical requirements for adaptive testing are as follows (Green, Bock, Humphreys, Linn & Reckase, 1984; Reckase, 1989; Weiss & Vale, 1987):

- an adequate pool of items with well-estimated item parameters
- an item selection procedure with rules for selecting the next most optimal item to be administered
- a scoring procedure to produce ability estimates on the same scale after each

item has been administered

- stopping rules - that is, a specified level of information, a specified posterior variance for Bayesian procedures or a certain number of items administered

In general administration of CATs, because the theta value is unknown at the outset, a first item of average difficulty is usually administered to the examinee. If the correct answer is given, the theta value is re-estimated (adjusted upward) and the next item will be more difficult to match the examinee's estimated ability level. If an incorrect answer is given, the theta value is re-estimated (adjusted downward) and the next item will be less difficult, once again to match the examinee's estimated ability level. Based on each response and the preceding responses, the computer estimates the theta value and its standard error and uses the information continually to select the next item to be administered. Testing is ended if some termination criterion is satisfied, for example, if the standard error of the ability estimate drops below a predetermined value.

The adaptive testing procedure quickly converges to the true theta value, using significantly fewer items than required in a traditional test to obtain the same measurement accuracy (Hambleton & Swaminathan, 1985; Reckase, 1988; Weiss, 1983a). CAT makes possible equiprecise measurement at different ability levels, since the termination criterion can be linked to the level of accuracy of measurement that has been achieved. The latter is of particular importance to measure the ability of examinees at the two extremes of the ability range. Adaptive test procedures can provide more information at the extremes of the ability distribution compared with standard tests (Hambleton & Swaminathan, 1985). The termination criteria for an adaptive test are commonly a minimum and maximum number of items to be administered, together with a required level of accuracy of measurement in terms of the variance of the ability estimation. To attain equal accuracy of measurement, more items may need to be administered at certain levels of ability where less information is available, or in cases where the answer pattern is somewhat erratic, which will influence the accuracy of ability estimation. With adaptive testing procedures it is possible to administer varying numbers of items to different individuals while scores remain comparable.

Although IRT and CAT procedures seem particularly suited to learning potential assessment, no previous application of CAT procedures based on IRT for learning potential assessment could be found in the literature.

4.7 ADVANTAGES OF USING CAT FOR DYNAMIC ASSESSMENT AND THE MEASUREMENT OF LEARNING POTENTIAL

According to Sijtsma (1993b, p185):

> *computerized adaptive testing could be used for learning potential assessment purposes. The use of different pretests and post-tests can be attained at the individual level, by means of computerized adaptive testing. Tailoring the pretest and the post-test to the performance level of each individual not only leads to shorter tests having the same precision for each individual, but also can be expected to motivate the testee during the rest of the training programme.*

In IRT and CAT, ability level is estimated on the basis of the difficulty level of items that are answered (correctly or incorrectly), and not on a number-correct score as is the case with classical tests. This allows the administration of different sets of items to different examinees, while the ability measures obtained are still directly comparable, thereby solving the problems regarding measurement of difference scores. This same principle allows direct comparison of pretest and post-test scores of the same examinee as well as comparison of scores of different examinees. In addition, because two separate CATs using different items for the pretest and post-test are used, memory does not confound test performance. The termination criteria for the pretest and post-test can be set independently, and can involve both a minimum and a maximum number of items to be administered, in combination with a required level of accuracy of the ability estimation. Another distinct advantage is that the pretest level of performance can be used as the entry-level ability of the post-test, thereby further improving testing efficiency. A commercial programme, MicroCAT (Assessment Systems Corporation, 1989,1995) is available for the construction of CATs with the

above features.

4.8 CONCLUSION

Relatively few tests of learning ability are available commercially, and according to Hegarty (1988), such testing will remain peripheral to the psychometric enterprise until practitioners have the means to include it in their assessments. Only when dynamic learning potential tests can be administered without significant additional testing time or effort and have been proven to be psychometrically sound, will they become a regular feature of cognitive assessment practice.

Because of improved measurement accuracy of difference scores and the time-saving measurement efficiency of CAT, the development of a dynamic CAT offers a useful solution to the measurement of learning potential. It can address most of the problems that have been identified, in particular those concerning the measurement of difference scores and test administration time. This approach was used in the construction of a new psychometric test, namely the Learning Potential Computerised Adaptive Test (LPCAT). The aim with the development and construction of the LPCAT is to provide a psychometrically sound instrument that meets the general requirements set for psychological tests, and that will be useful for the measurement of learning potential of all population groups in South Africa. The construction of this new instrument is discussed in the next chapter.

CHAPTER 5

CONSTRUCTION OF THE LEARNING POTENTIAL COMPUTERISED ADAPTIVE TEST (LPCAT)

5.1 INTRODUCTION

5.1.1 Need for a new instrument

A need was identified to construct an instrument for the measurement of learning potential in the domain of general reasoning ability. Such a test can lessen the impact of socioeconomic, cultural or educational background on test performance so that cognitive ability can be assessed in a more equitable and culture-fair manner. To this end it was deemed necessary to develop a dynamic test of learning potential that includes training in the test administration to cater for improvement of test performance and to accommodate people in various stages of development. The aim of the test is not to rely on prior school learning, language proficiency or the socioeconomically related aspects of knowledge. Such a test can provide important screening information for South Africa's multicultural society.

Previous chapters have indicated the need for a learning potential instrument that can address some of the key dynamic testing problems such as administration time, accuracy of measurement of difference scores as well as interpretation of difference scores at different performance levels. Such a test would need to accommodate not only differences between cultural groups, but should also make provision for differences in each group. Because of the complex language situation in South Africa with 11 official languages, language content should be avoided for learning potential measures in particular. Although language proficiency has been shown to be a good predictor of academic performance, in particular of previously disadvantaged students (De Beer & Van Eeden, 1997), the aim of a learning potential measure is to avoid material that is related to socioeconomic or educational background and to move away from the measurement of that which has been learnt previously.

To make a useful contribution, any new instrument should be standardised for South African use, should provide information not already provided by available instruments, or should allow increased measurement efficiency or accuracy of measurement compared to existing instruments. According to Claassen (1997, p 305): “tests of learning potential show promise and are intended to serve a laudable purpose, but at this stage only limited information is available about the way they relate to more established measures of cognitive abilities”. Any new instrument therefore also has to be investigated for construct validity by comparing results with existing standard cognitive test results.

Although learning potential tests differ in terms of their administration procedure, they are expected to have the psychometric properties of a standard test (Hamers, Hessels & Pennings, 1996). The problems experienced with existing dynamic tests for the measurement of learning potential concern the extended administration time, non-standard training, low measurement accuracy, difficulty in interpreting scores and the prevailing use of retarded subjects in research studies. Grigorenko and Sternberg (1998) emphasise the need for conducting studies that involve larger participant populations, the need to validate dynamic testing results against educational or professional criteria as well as the need to replicate results. They list the following important aspects of the evaluation of dynamic tests:

- whether new information is provided over and above that obtained with conventional measures
- how successfully the new methodology predicts performance in a designated population for a given set of criteria
- the time and effort required compared to the unique contribution of the information obtained
- whether the results have been shown to be replicable across studies and research groups

These aspects will be investigated and reported on for the Learning Potential Computerised Adaptive Test (LPCAT), the construction and evaluation of which is the

focus of the present study.

5.1.2 Overview of general steps in test construction

Test construction is a long and involved process which often takes years to complete. In their recent policy document for the classification of psychological tests (South African Professional Board for Psychology, 1998/18/9/B), the Psychometrics Committee of the South African Professional Board for Psychology acknowledges the long process involved in test construction, but makes an urgent appeal for the development of culturally appropriate instruments in South Africa. The generic steps in the construction of psychological tests provided by various authors (Aiken, 1996; Anastasi & Urbina, 1997; Gregory, 1996; Hambleton & Swaminathan, 1985;, Hambleton & Zaal, 1991; Kline, 1991; Reckase, 1996; and Weiss 1983a) can be summarised as follows:

- **STEP 1: Defining the test and preparation of test specifications**
  The test developer should, in the light of existing instruments, show that the proposed instrument has a contribution to make that existing instruments cannot make - in other words, that it is different from and/or better than existing instruments.

- **STEP 2: Choice of psychometric model and selection of scaling method**
  The overall purpose of psychological testing is to assign numerical values to test performance. Decisions have to be made about the level of measurement and scaling method to be used.

- **STEP 3: Writing of items and preparation of the item pool**
  The process of item construction is time-consuming. Having clear guidelines in terms of the item format, table of specifications for content, difficulty level and other important factors helps to ease the process. To make provision for items that will be discarded during later item selection, approximately one and a half times the number of items eventually needed are written initially.

- **STEP 4: Field testing of the items**

  To obtain useful information for item analysis purposes, the items have to be administered to a large and representative sample of the population for whom the final test is intended.

- **STEP 5: Item analysis and selection of test items**

  In order to construct a final test that is psychometrically sound, item analysis has to be performed to identify the best items. Indices that are typically used to evaluate items are the item-difficulty index, item-reliability index, index of discrimination, IRT parameters and item-characteristic curve. This information can be used to construct a final test that adheres to the construction criteria initially set.

- **STEP 6: Construction of the final test**

  The steps provided below focus specifically on the construction of computerised adaptive tests.

  - choice of appropriate IRT model
  - choice of starting point
  - choice of item selection model
  - item scoring/ability estimation
  - choice of stopping rule

- **STEP 7: Reliability and validity studies**

  Evidence of the reliability and validity of the instrument needs to be provided. For the LPCAT, the planning of this phase and the results in support of its use are discussed in chapters 6 and 7 respectively.

- **STEP 8: Final test production**

  Production of test materials, user manuals and technical manuals form the last phase of test construction. User-friendly material that facilitates smooth and correct administration should be the aim. The psychometric properties of the

test, particularly reliability and validity have to be investigated and reported in the test manual. The LPCAT will have both a user's manual and a technical manual to provide users with the relevant information they might need to administer the test as well as to psychometrically and technically evaluate the instrument.

5.1.3 Main features of the LPCAT and an overview of its construction

The main features of the LPCAT reflect the initial objectives for its development, which are to construct a test for the measurement of learning potential that will

- use nonverbal, figural items that can be administered to all cultural groups
- make use of computerised adaptive testing (CAT) to save administration time without forfeiting quality or accuracy of measurement
- use IRT and computerised adaptive features for more accurate measurement of change scores
- use the dynamic test-train-retest approach
- incorporate a standard training section similar to typical group training situations
- focus on learning potential and monitor not only present performance, but also to what extent examinees are able to improve their performance after the relevant training
- use multicultural groups both for item analysis and standardisation and validation of the test to provide the required evidence for the psychometric properties and validity for the use of the LPCAT for cross-cultural or multicultural assessment

Development of the LPCAT started in 1993 with the initial conceptualisation and planning of the project. In the intervening years, the following phases followed:

- 1994: items developed
- 1995: field testing of the items for item analysis
- 1996: item analysis, computerisation and initial computerised adaptive administration to investigate validity
- 1997: analysis of initial validity information, development of a second (nontext) version of the LPCAT and further administration to study validity

- 1998: translations of test instructions and analysis of the second (nontext) version to obtain validity information
- 1999: final validity administration, data analysis, completion of test manuals (user's and technical) and finalising of software programs to run the test and provide the test results in a user friendly manner

Initial development of the LPCAT focused on a single form of the test, but it soon became evident that a second version was needed to accommodate examinees who do not have the required level of reading proficiency in English or Afrikaans required to read and follow the test instructions, training and feedback independently on the screen. This led to the development of a second form of the LPCAT, namely LPCAT-2, with the initial form being renamed LPCAT-1.

5.2 DEFINING THE TEST

Questions about the cross-cultural fairness of standard cognitive tests led to the conceptualisation and development of a learning potential test for cross-cultural cognitive assessment in the South African context. The LPCAT was developed as a nonverbal, culture-fair measure of learning potential in the domain of cognitive ability or general reasoning ability. It is intended to serve as a screening instrument that can be used mainly to counter inadvertent discrimination against disadvantaged groups. It provides a measure of learning potential using nonverbal general reasoning ability that is not dependent upon either language proficiency or prior school learning. The LPCAT consists only of figural, nonverbal item types and uses a dynamic test-train-retest format combined with computerised adaptive testing. Relevant training is provided as part of the standardised test administration. The LPCAT thus provides a culture-fair measure of learning potential, indicating present level of performance as well as potential future level of performance after relevant training. The difference between these two measures can be interpreted as the magnitude of *undeveloped* learning potential. Legislative requirements for psychological testing in South Africa (Employment Equity Act 1998) were complied with in the construction of the LPCAT.

By also focusing on the potential to improve performance and not only on developed abilities, it is acknowledged that people have different backgrounds, and because of inadequate educational or training opportunities, may not be functioning at their optimal level. A larger ZPD indicates probable improvement in future levels of performance if relevant training can be provided. A smaller ZPD indicates that the person will probably continue to function at or close to his or her present level. A smaller ZPD is likely to be found for people who either already function close to their optimal level or for those who function at a high level and therefore do not have much room for improvement. In South Africa in particular, where there are large differences in the educational opportunities for and socioeconomic backgrounds of people, it is necessary to identify undeveloped potential, for this will help to provide learning and training opportunities for those who will benefit most from them. Learning potential for the LPCAT is defined as a combination of the pretest performance and the magnitude of the difference between the post-test and the pretest scores. It is important that the difference score should not be used alone, but that present level of performance should also be taken into account. The present level of functioning, the potential future level of functioning and the potential for improving on the present level of functioning are all considered in the interpretation of the results.

5.3 CHOICE OF SCALING METHOD

The LPCAT is constructed as a computerised adaptive test using the three-parameter item response theory model. As such, the scaling of the (latent) ability level is on the theta scale with a mean of 0 and standard deviation of 1. The three-parameter model has been used most widely in CAT and can be regarded as a general model for dichotomously scored items (Assessment Systems Corporation, 1989). The Bayesian modal method is used for ability estimation and item selection in the LPCAT (see 5.7.4). The Bayesian item-selection strategy selects items on the basis of minimising the Bayesian posterior variance of the ability estimate (Assessment Systems Corporation, 1989).

Since the theta scale used for ability estimates in the standard three-parameter IRT

model includes negative values, the final scores for the LPCAT will be provided in the form of T-scores with a mean of 50 and a standard deviation of 10 as well as percentile scores and stanines.

5.4 LPCAT ITEMS AND PRACTICE EXAMPLES

An important ingredient of a good CAT is a large, well-distributed pool of items for a domain with one dominant dimension (Green et al, 1984). Based on the results of previous South African research (Hugo & Claassen, 1991) and general international consensus on culture-fair test content (Jensen, 1981), it was decided to use nonverbal items of the figural type only. Verbal items and number series were purposefully excluded in an attempt to negate the effects of prior learning, so that language proficiency and/or prior scholastic background would not affect test performance.

The item types chosen for the LPCAT were figure analogies, pattern completion and figure series items. These item types are typical of the figural items found in most cognitive ability tests and are generally considered to provide a fairly pure measure of Spearman's g-factor (Jensen, 1981, p 133):

> *Culture-reduced tests try to minimise culture loading by not using words, letters, numbers, or even pictures of familiar common objects. They consist of only simple elements - lines, curves, circles and squares - and they involve such universal concepts as up/down, right/left, open/closed, whole/half, larger/smaller, many/few, full/empty, and the like. Quite complex problems involving relational reasoning can be made up of such elements - for example, figural analogies, figure series completion, and matrices. Such tests are near the opposite extreme on the culture-loading continuum as compared with tests involving specific factual knowledge or scholastic content.*

Considering the type of figures used and the similarity of patterns formed by the figures in the three formats used, it is assumed that there is probably considerable overlapping

in the processes underlying the solution of these item types. On the face of it, the test can be considered to be largely one-dimensional. Evidence in support of the one-dimensionality of the LPCAT items is discussed in 5.6.2.1. The format of the three item types used in the LPCAT is described next.

- **Figure series**

  A series consisting of four figures is presented, each in a square. A fifth square is empty, and the examinee is expected to deduce a rule from the given part of the series and to complete the series accordingly.

- **Figure analogies**

  Two figures that correspond in a certain way are given in a block. In a second block, a third figure is given, and a fourth one should be selected so that it corresponds with the third figure in the same way as the second figure corresponds with the first. The first and third figures also correspond in a certain way (eg size) - hence the second and fourth figures should correspond in the same way as the first and the third.

- **Pattern completion**

  A block consisting of nine squares, in three rows and three columns, is presented. In each row and column, three figures form a pattern. The figure in the last row and column should be found. The examinee is expected to deduce a rule from the given part of the pattern and to complete the pattern accordingly.

A pool of 270 new items - 90 each of the three item types - of varying difficulty was constructed. The items were aimed at the lower-ability levels, although an attempt was made to have items of each of the three types available at all ability levels. Although the LPCAT is intended to measure at all ability levels, most interest in learning potential measurement is aimed at lower-ability level and low educational level examinees. It is at the lower-ability levels that most information is needed for disadvantaged persons, who tend to obtain lower scores on standard tests of cognitive ability. LPCAT items across the difficulty-level/ability-level spectrum allow estimation of learning potential at all levels. Items were evaluated by a committee of specialists on cognitive assessment and

changes were made on the basis of the feedback received. These items were administered in paper-and-pencil format for item analysis purposes (see 5.5). For this administration, the items were grouped into three subtests, one for each item type. Each subtest started with six practice examples to ensure that the examinees understood the questions and how to answer them.

5.5 ITEM ANALYSIS ADMINISTRATION

5.5.1 Introduction

Administration of all items in the item bank to a multicultural sample of examinees took place during 1995. The samples were large enough to analyse items by means of the three-parameter IRT model, which is best for scoring multiple-choice items (McBride, 1997). Although item parameters obtained from paper-and-pencil administration may differ from those obtained in computer-administration, practical considerations made it impossible to administer LPCAT items by computer for item analysis purposes. However, according to Hetter, Segall and Bloxom (1997), item parameters calibrated from paper-and-pencil administration of items can be used in power CATs of cognitive constructs without changing the construct being assessed and without reducing reliability. The number of items that needed to be administered was too large to administer to examinees in a single test. This necessitated the construction of two paper-and-pencil forms with sufficient anchor items - items answered by both groups - to calculate item parameters on the same scale.

In IRT, the population used for determining the item parameters requires that a group roughly comparable to the target population be used in order to obtain accurate estimation of item parameters (Green et al, 1984). This standardisation sample is the group against whose general performance eventual test scores are interpreted. Psychological test norms are based on the test performance of individuals of the standardisation sample, and it is therefore important to obtain a sample that is representative of the population for which the test is designed.

5.5.2 LPCAT standardisation sample

Owing to the transformation process in education at the time, access to schools was severely restricted. Nevertheless, three regions were identified where the school psychologists indicated their willingness to administer the items in paper-and-pencil format. Although only three of the 10 provinces were included in the item analysis sample, there is no reason to believe that the pupils in these provinces are any different from those in the other provinces.

Forty-one schools were selected. These included 15 schools (37%) from the Northern Cape, 12 schools (29%) from the Northern Province and 14 schools (34%) from Mpumalanga. The schools had been identified, on a random basis, by the HSRC Centre for Statistical Support, taking into account the urban and rural distribution and the sizes of the school populations. School psychologists from the three provinces were provided with a list of school names included in the sample and the necessary test material was sent to them by post. At each school, 60 pupils, 30 from grade 9 and 30 from grade 11 were randomly selected for testing. Furthermore, in each grade group of 30 pupils, half the examinees were boys and half girls. In each grade sample group of 30, Form A and Form B of the test were alternated, thereby ensuring an equal distribution of the two forms between both the gender and the grade groups. The cultural group and language group allocation of each school was based on the education body that had previously been responsible for that school. In 1995, when the LPCAT items were administered in paper-and-pencil form, schools were still relatively homogeneous in terms of language and culture.

Of the four main cultural groups in South Africa (African, Indian, Coloured and White), all but the Indian group were included in the paper-and-pencil sample. The reason for the exclusion of the Indian group was threefold. Firstly, they form only 2,5 percent of the South African population (CSS, 1996c). Secondly, in cognitive test performance as well as in socioeconomic status and educational attainment, they are very similar to the White group (CSS, 1996c). Thirdly, the province with the highest representation of the Indian population was not one of the three provinces included for the item analysis test

administration. Indian examinees were later included in the validation of the LPCAT in its computerised format. The cultural composition of the sample for the item analysis is given in Table 5.1.

TABLE 5.1 CULTURE AND GENDER COMPOSITION OF THE STANDARDISATION SAMPLE

Group	African pupils	Coloured pupils	White pupils	Total
Male	600	300	328	1 228
Female	597	299	330	1 226
Total	1 197	599	658	2 454

According to the 1996 census information (CSS, 1996c), the percentage of the different cultural groups in South Africa is 76,3 percent African, 12,7 percent White, 8,5 percent Coloured and 2,5 percent Indian. The representation of these groups in the LPCAT standardisation sample is 49 percent African, 27 percent White and 24 percent Coloured. The African group is therefore underrepresented, while the Coloured and White groups are proportionally overrepresented. This distribution does, however, provide adequate numbers of examinees of the different subgroups for item analysis purposes. There was an almost equal gender distribution with 1 228 male and 1 226 female pupils included. Despite the fact that the sample cannot be considered to be statistically representative of the South African population (because of the lack of both regional and full cultural representation), in practical terms, it can be regarded as being representative of groups in South Africa. The sample sizes for the different groups were large enough to meet the requirements of the procedures used for analysis - in particular for three-parameter IRT item analysis. The composition of the sample in terms of regional distribution is provided in Table 5.2.

TABLE 5.2 CULTURE AND REGIONAL COMPOSITION OF THE

STANDARDISATION SAMPLE

Culture group/ rural-urban	African pupils		Coloured pupils		White pupils		Row Total		TOTAL
Province	R*	U*	R*	U*	R*	U*	R*	U*	All
Mpumalanga	360	300	-	-	-	180	360	480	840
Northern Province	240	240	-	-	-	240	240	480	720
Northern Cape	-	60	60	540	-	240	60	840	900
Total	600	600	60	540	-	660	660	1 800	2 460

*R = Rural U = Urban

Table 5.3 provides the numbers of pupils from the different culture groups in the item analysis sample that completed the items in the two test forms. Owing to missing values for certain variables, the numbers of the different subgroups in the description of the composition of the sample henceforth will not always add up to the same overall total number of examinees in the planned sample.

In Table 5.4, a summary of examinees by test form and culture group is provided. One thousand two hundred and seventy-seven pupils (52%) completed Form A, while 1 173 pupils (48%) completed Form B.

TABLE 5.3 CULTURE, GENDER AND TEST FORM DISTRIBUTION IN THE STANDARDISATION SAMPLE

Culture	African				Coloured				White			
Gender	M		F		M		F		M		F	
Form	A	B	A	B	A	B	A	B	A	B	A	B
N	301	296	337	258	153	147	150	149	182	146	153	177
Gender total	597		595		300		299		328		330	
Culture total	1 192				599				658			

TABLE 5.4 CULTURE AND TEST FORM COMPOSITION OF THE STANDARDISATION SAMPLE

Culture group	Form A only	Form B only	Total
African	639	554	1 193
Coloured	303	296	599
White	335	323	658
Total	1 277	1 173	2 450

For each of the three item types, 56 of the 90 items were included in each of Form A and Form B, with 22 of the 56 items of each type being anchor items, repeated in both Form A and Form B. IRT procedures can use anchor items to combine samples for item analysis purposes, and to this end, the 66 anchor items were used. The 90 items of the three item types were distributed between the two paper-and-pencil forms as shown in Table 5.5.

TABLE 5.5 ITEM TYPE DISTRIBUTION FOR THE ITEM ANALYSIS ADMINISTRATION

Item types	Items 1-34 Form A only	Items 1-34 Form B only	Items 35-56 Forms A&B anchor	Form A total	Form B total	Item type total
Figure series	34	34	22	56	56	90
Figure analogies	34	34	22	56	56	90
Pattern completion	34	34	22	56	56	90
Total	102	102	66	168	168	270

The results of the paper-and-pencil administration were used for item analysis and DIF analysis.

5.6 ITEM ANALYSIS

5.6.1 Classical test theory item analysis

The entire pool of 270 items was analysed by means of both classic item analysis and IRT item analysis. The ITEMAN program of MicroCAT (Assessment Systems Corporation, 1995) which was used for the classical item analysis, scores items that are not reached as incorrect, and this affects the values obtained. Therefore, for the

classical test theory item analysis, the items of the two forms had to be kept separate as two tests with 168 items each. Consequently, for the anchor items included in both forms, two sets of values were calculated. The classical item analysis information included the item difficulty value (p-value) as well as the item discrimination value (r_{it}). The p-value indicates the proportion of examinees who answered that particular item correctly. Item discrimination (r_{it}) reflects the correlation of the item with the total score. The means for these values for Form A and Form B are reported in Table 5.6.

TABLE 5.6 MEAN VALUES OF CLASSICAL TEST THEORY ITEM PARAMETERS

	Form A	Form B
p-value	0,656	0,657
r_{it}	0,498	0,476

Although these indices were used to help evaluate the properties of the items of the LPCAT, most of the standard classical indices do *not* apply for computerised adaptive testing, because a standard number and sequence of items are not applied, as in a standard test.

TABLE 5.7 COEFFICIENT ALPHA VALUES FOR THE TWO TEST FORMS FOR DIFFERENT GROUPS

GROUP	Form A (168 items)		Form B (168 items)	
	N	Alpha	N	Alpha
Total group	1 277	0,981	1 173	0,978
African	639	0,975	554	0,971
Coloured	303	0,969	296	0,970
White	335	0,925	323	0,926

African language group	639	0,975	554	0,971
English/Afrikaans language group	638	0,973	619	0,971
Male	636	0,981	589	0,979
Female	640	0,980	584	0,978
Grade 7	622	0,980	600	0,977
Grade 9	653	0,981	572	0,979

Another classical test theory index that was calculated is coefficient alpha, a measure of internal consistency or test homogeneity. Using all the items that were administered for item analysis, the alpha values ranged between 0,925 and 0,979 for the various subgroups. The alpha value for the total group was 0,981 for Form A and 0,978 for Form B, indicating high internal consistency. Table 5.7 provides the coefficient alpha values for the two test forms (A and B) for both the total group and various subgroups.

Coefficient alpha is regarded as an index of reliability in standard tests, and according to Gregory (1996), can be regarded as an index of the degree to which a test measures a single factor. In this regard the high values obtained for coefficient alpha provide support for the one-dimensionality of the LPCAT items.

5.6.2 Item response theory item analysis

Applications of IRT depend upon the item parameters, which are obtained by using computer programs designed to estimate them. Use of the three-parameter model allows for variation among the items in their level of difficulty, their discrimination power and also for guessing on the multiple-choice test items by low-ability examinees. According to Hambleton, Zaal and Pieters (1991), the three-parameter model fits test data better than either the one- or two-parameter models, and the three-parameter logistic model is the model of choice by most CAT advocates. Its requirements for item analysis, namely large sample sizes and sufficient numbers of low-ability examinees, were met by the LPCAT standardisation sample.

For the LPCAT IRT item analysis, Form A and Form B were combined into a single test of 270 items by using the 66 anchor items which all examinees had completed. The anchor items had the same positions (nos 35-56) in each of the three subtests of the two paper-and-pencil forms. For each of the three item types, items were combined using Form A items 1 to 34 as the first 34 items, using the anchor items (nos 35-56) of both Form A and Form B for item numbers 35 to 56, and lastly, using items 1 to 34 of Form B for item numbers 57 to 90. It was possible to combine the two forms for IRT item analysis because the ASCAL program of Microcat (Assessment Systems Corporation, 1995, p 12-13) works with dichotomously scored items and makes an important distinction between items that are coded as omitted and items that are coded as not reached. Items that are not reached are excluded from the analysis for the examinee concerned. The 34 items from the alternative form which the examinee did not complete, were therefore coded as “not reached”, which made it possible to combine the two groups, thereby increasing the available sample size for the anchor items.

Five of the original 270 items were discarded during initial analysis, because of problems with some of the distractors. The ASCAL program of MicroCAT (Assessment Systems Corporation, 1995) was used to calculate the IRT item parameters: the ASCAL parameter estimation program estimates IRT item parameters according to the two- and three-parameter IRT models. ASCAL estimates item parameters using a combined maximum likelihood and modal Bayesian procedure. The initial theta distribution is broken up into 20 fractiles and then the mean theta in each fractile is calculated and used as the theta level for all examinees in that fractile. Item parameter estimation then proceeds using a Bayesian adaptation of Lord’s maximum likelihood equations with the a-parameter bounded at 0,4 and 2,4; the b-parameter bounded at -3,0 and +3,0; and the c-parameters bounded at 0,0 and 2/K (where K is the number of alternatives provided). In the estimation procedure, each item’s lack of fit to the IRT model is indexed by a Pearson chi-square statistic. The procedure estimates item parameters through an iterative process. This means that it estimates the parameters several times, each time using the previous estimates as starting points from which to make better estimates.

A descriptive summary of the values of the item bank before item selection is provided in Table 5.8. The mean a-value indicates that, on average, items discriminate well, while the mean b-value, being less than 0,0 indicates that most items are reasonably easy.

TABLE 5.8 DESCRIPTIVE STATISTICS OF ITEM PARAMETERS OF THE ITEMS SUBJECTED TO IRT ITEM ANALYSIS

IRT parameter	N	Mean	SD	Minimum	Maximum
a-value	265*	1,435	0,486	0,442	2,500
b-value	265	-0,231	0,829	-1,558	3,000
c-value	265	0,179	0,0853	0,000	0,470

* Five of the 270 items were discarded during IRT item analysis.

Selection of the items to be included in the final version of the LPCAT was based on both classical and IRT item analysis, although greater weight was attached to IRT item parameters.

Three of the most important general assumptions of IRT are one-dimensionality, item parameter invariance and ability parameter invariance. These three assumptions were empirically investigated for the LPCAT item bank. It was decided to use the entire bank of items and not only those items that were included in the final version of the test. Exclusion of items that were eventually discarded because they failed to comply with the standards set, is likely to positively affect the results reported in the following subsections.

5.6.2.1 One-dimensionality

Because one-dimensionality is a general assumption in the use of IRT, the factor structure of the LPCAT items was investigated for both the total group and specific subgroups. This was done to determine whether the same constructs were being measured for the different groups. The factor analysis was performed at *item* level and not subtest level. LPCAT items were constructed to measure a single domain (general nonverbal, figural reasoning). Principal component factor analysis is regarded as appropriate when the primary concern is to predict the minimum number of factors needed to account for most of the variance. In the case of the LPCAT, factor analysis had to be executed separately for Form A and Form B. The results indicate support for a one-dimensional structure for both the total group and the various subgroups. The eigenvalues for the different groups are reported in Table 5.9.

For both Form A and Form B, the eigenvalues for the first factor were between 6,54 and 8,92 times larger than the eigenvalue for the second factor for the total, African and Coloured groups. The eigenvalues of subsequent factors were significantly closer to each other. The exception to the above ratios of eigenvalues was for the White group where for Form A the first eigenvalue was only 2,65 times the size of the second. This ratio was 3,33 for Form B. Considering the item types and item content used and the similarity between strategies required to solve the items, the above results provide support for the expected one-dimensionality of the LPCAT items.

According to Hair, Anderson, Tatham and Black (1995), to ensure practical significance for the derived factors, extraction can be stopped at the point where the last factor accounts for only a small proportion (less than 5%) of the variance. The scree test can also be used to determine the cutoff point for the number of factors. The point at which the curve begins to straighten out indicates the maximum number of factors to extract. In the case of the LPCAT, the scree plots also provide support for the one-dimensional nature of the LPCAT item domain, as shown in Figures 5.1 and 5.2. Scree plots for various subgroups are provided in the *LPCAT technical manual* (De Beer, 2000b).

TABLE 5.9 EIGENVALUES AND PERCENTAGE OF VARIANCE FOR DIFFERENT GROUPS FOR FORM A AND FORM B

Group	Factor 1 Eigenvalue	Factor 1 % variance	Factor 2 Eigenvalue	Factor 2 % variance	Factor 3 Eigenvalue	Factor 3 % variance
Form A: total group	44,552	26,519	5,721	3,406	3,488	2,076
Form B: total group	41,736	24,843	4,678	2,784	3,329	1,982
Form A: African group	37,264	22,181	4,537	2,701	3,961	2,358
Form B: African group	33,645	20,027	3,784	2,252	3,473	2,067
Form A: Coloured group	34,012	20,245	5,200	3,095	3,991	2,376
Form B: Coloured group	33,990	20,232	5,054	3,008	4,238	2,522
Form A: White group	17,618	10,487	6,659	3,964	6,343	3,776
Form B: White group	18,032	10,734	5,418	3,225	4,532	2,698

FIGURE 5.1 SCREE PLOT FOR ALL GROUPS (FORM A)

FIGURE 5.2 SCREE PLOT FOR ALL GROUPS (FORM B)

Another reason why the one-dimensional nature of the LPCAT is important is to accommodate and simplify the training that is provided as part of the dynamic assessment approach. The items used in the LPCAT are similar in that they are all nonverbal, figural items that require similar strategies to find the correct answer. This makes it easier to provide relevant training for these types of items. Since dynamic testing attempts to assess the extent to which a person can use relevant training to improve test performance, the one-dimensionality of the test content improves the

efficiency of the training.

5.6.2.2 Item parameter invariance

The second general assumption of IRT that was investigated was that of item parameter invariance. According to Lord (1980, p 35), “the invariance of item parameters across groups is one of the most important characteristics of item response theory”. He warns that we are so accustomed to thinking of item difficulty in terms of the proportion of correct answers, that it is sometimes hard to imagine how item difficulty can ever be invariant across groups that differ in ability level. The invariance of item parameters across groups means that if we determine the item parameters for a set of items with two separate groups of examinees independently, we can expect a linear relation to exist between the item parameters. Lord (1980) warns that we should not expect the parameters to be identical. This relation can be empirically investigated by means of scatter diagrams of the parameters calculated for two separate groups and also by obtaining the correlation between the two sets of values.

According to Hambleton and Swaminathan (1985), it is desirable to identify subgroups of special interest in the examinee population and use them to study item parameter invariance. The item parameters of the LPCAT were investigated by using two sets of independent groups, namely the two gender groups (male vs female) and the two home language groups (English/Afrikaans speaking vs African languages). The item parameters for these groups were calculated separately with the MicroCAT ASCAL program (Assessment Systems Corportation, 1995). The scatter diagrams and correlation results were obtained for both comparison sets for all three of the parameters. Scatter plots show the relationships between the sets of item parameter values obtained for the independent subgroups. Since the main interest for the LPCAT is in terms of the cross-cultural application, the plots obtained for the two gender groups can be used as baseline plots for comparison. The best results were obtained for the b-parameter, which may be regarded as the main parameter for comparison, since it reflects the difficulty of the item. The b-parameter plots of the gender and language groups (Figure 5.3 and Figure 5.6) are similar. There are some differences in the

a-parameter and c-parameter distributions of the gender and language groups, but in particular in the light of the similarities found in terms of the b-parameters and the multiple-choice item format, this is not considered to be a limiting factor in the use of the three-parameter IRT model. The scatter diagram results are reported in Figures 5.3 to 5.8. All correlations found were highly significant, which provides support for the invariance of the three parameters obtained with these independent groups.

FIGURE 5.3 SCATTERGRAM OF THE B-PARAMETER (ITEM DIFFICULTY) OF THE GENDER GROUPS (r=0,948; p<0,001; N=265)

FIGURE 5.4 SCATTERGRAM OF THE A-PARAMETER (DISCRIMINATION) OF THE GENDER GROUPS (r=0,813; p<0,001; N=265)

FIGURE 5.5 **SCATTERGRAM OF THE C-PARAMETER (PSEUDO-CHANCE) OF THE GENDER GROUPS (r=0,715; p<0,001; N=265)**

FIGURE 5.6 **SCATTERGRAM OF THE B-PARAMETER (ITEM DIFFICULTY) OF THE LANGUAGE GROUPS (r=0,945; p<0,001; N=265)**

FIGURE 5.7 **SCATTERGRAM OF THE A-PARAMETER (DISCRIMINATION) OF THE LANGUAGE GROUPS (r=0,558; p<0,001; N=265)**

FIGURE 5.8 **SCATTERGRAM OF THE C-PARAMETER (PSEUDO-CHANCE) OF THE LANGUAGE GROUPS (r=0,454; p<0,001; N=265)**

5.6.2.3 Ability parameter invariance

The other general assumption of IRT that was investigated for the LPCAT items was that of ability parameter invariance. This refers to the fact that in IRT, the ability parameter of a person is not affected by the items that are used to estimate it. According to Lord (1980), ability parameters are invariant from one test to another, except for the choice of origin and scale, assuming that the two tests both measure the same ability or (latent) trait. This characteristic can be empirically investigated by calculating the ability parameters of a group of examinees with two different sets of items. In the case of the LPCAT, this was done for three different sets of item combinations by using the separate item types to independently calculate the ability estimates for the total group of examinees. The MicroCAT ASCAL program was used (Assessment Systems Corporation, 1995). These results are reported in Figures 5.9 to 5.11.

FIGURE 5.9 SCATTERGRAM OF ABILITY ESTIMATION OF EXAMINEES USING FIGURE SERIES AND FIGURE ANALOGY ITEMS (r=0,859; p<0,001; N=2450)

FIGURE 5.10 **SCATTERGRAM OF ABILITY ESTIMATION OF EXAMINEES USING FIGURE SERIES AND PATTERN COMPLETION ITEMS (r=0,836; p<0,001; N=2450)**

FIGURE 5.11 **SCATTERGRAM OF ABILITY ESTIMATION OF EXAMINEES USING FIGURE ANALOGY AND PATTERN COMPLETION ITEMS (r=0,873; p<0,001; N=2450)**

Figures 5.9 to 5.11 provide support for the invariance of ability parameter estimation using different subsets of LPCAT items, indicating that similar ability estimates are obtained with different sets of items. The distributions and correlations found here are similar to those found in other such studies (Gierl & Hanson, 1995).

The evidence supporting the one-dimensionality of LPCAT items as well as the invariance of item parameters and ability estimates, justifies the use of the three-parameter IRT model for the LPCAT.

5.6.3 IRT differential item functioning (DIF) analysis

According to Zieky (1993), the investigation of DIF helps to identify test items that may be unfair for members of certain groups. According to Linn (1993), DIF analysis cannot be expected to convince people who do not want tests to be used. DIF statistics are often difficult to interpret and procedures for the use of DIF in test development are still evolving.

According to Osterlind (1983), bias is a technical term which indicates some systematic error in the measurement process, while Holland and Wainer (1994, p xiv) state that "the ambiguous term *item bias* is used to refer to an informed judgment about an item that takes into account the purpose of the test, the relevant experiences of certain subgroups of examinees taking it, and statistical information about the item". In

general, bias is considered to be a technical matter which requires careful scrutiny and statistical investigation of test items. Fairness of a test, on the other hand, indicates whether it is an equally valid measure of ability for individuals from different groups and deals with the social consequences of test use - often involving socially-based and more subjective evaluation of information.

More recently, there has been a change in terminology, with most writers now preferring to use the term “differential item functioning” (or DIF) to refer to biased items - items that function differently for different groups of examinees. IRT has provided a major breakthrough in the study of DIF and its more sophisticated techniques contribute to improved procedures for measuring and analysing DIF. For example, DIF can be investigated at particular ability levels or over the entire ability spectrum, which provides a distinct advantage over classical methods. Despite its complexity, the IRT-based approach is a most valuable tool for investigating DIF.

The way in which the item characteristic curves (ICCs) are used to evaluate DIF is to compare the ICCs of two groups. In DIF analysis, the examinee group of interest is referred to as the focal group, while the group to which its performance on the item is being compared is called the reference group (Holland & Wainer, 1993). After calculating the item parameters separately for the two groups, the theta scales are equated (Lord, 1980; Osterlind, 1983; Van den Berg, 1989). The ICCs can then be drawn on the same graph and compared for DIF. If a test item has exactly the same item response function for each group, persons at any given level of ability will have exactly the same probability of getting the item right. This would be true even though one group may have a lower mean theta, and thus lower test scores than the other group (Lord, 1980). The basic approach to measurement of DIF therefore lies in the difference between the probability of getting the item correct if one is a member of one (focal) group, versus what would have been the probability of a correct response if one were a member of the other (reference) group. “A test item is said to be unbiased when the probability for success on the item is the same for equally able examinees of the same population regardless of their subgroup membership” (Osterlind, 1983, p 3). If there is a distinct difference between the ICCs of the two groups, the item shows DIF. Such items should be flagged so that they can be further evaluated and possibly

scrapped if they do not meet the requirements set for inclusion into the test bank.

The most common procedure for detecting bias is by means of calculating the area between the two ICCs (Wainer, 1993). Angoff (1993) warns that restricting the ability (theta) values between -3 and +3 is required to limit the influence of differences in the c-parameter. This precaution was taken for the present project when the areas between various ICCs were calculated for DIF analysis.

Another approach to investigating bias is to compare test results for particular groups with some outside criterion, as is done when assessing the criterion-related validity of the test. Test performance can be compared with present performance on some criterion measure (concurrent validity) or with future performance on a criterion (predictive validity). In both cases, differences between the test scores of subgroups should primarily be caused by differences in whatever the test purports to measure. Criterion-related validity results for the LPCAT are discussed in chapter 7.

A distinction is made between uniform DIF and nonuniform DIF. In uniform DIF, the probability of answering an item correctly for one group is consistently lower than that of the other group. This results in the ICC for one group being below that of the other group over the entire ability range (see Figure 5.12). In nonuniform DIF, the curves cross at a certain point. Whereas for one range of ability the one group has a lower probability of answering the item correctly, the reverse is true for another range of ability. Figure 5.13 illustrates an item that shows nonuniform DIF.

FIGURE 5.12 ITEM SHOWING UNIFORM DIF BETWEEN TWO GROUPS

FIGURE 5.13 ITEM SHOWING NONUNIFORM DIF BETWEEN TWO GROUPS

FIGURE 5.14 ITEM SHOWING NO DIF BETWEEN GROUPS

Of course, the ideal is that there should be little difference between the ICCs of the two groups being compared. An item with no DIF is shown in Figure 5.14.

To investigate differential item functioning in the present study, ICCs for the following four sets of groups were compared:

Language: African home language vs English/Afrikaans
Culture: African vs White
Gender: Male vs female
Grade: Grade 9 vs grade 11

The only sample that could be considered somewhat small for the three-parameter item analysis to obtain the ICCs was the White group (N=658). All the other subgroups were sufficiently large (samples larger than 1 000) to justify the IRT analysis.

The following procedure was used to calculate the area between the ICCs:

The theta range was limited from -3,0 to +3,0 and divided into sections of 0,1 in width (ie -3,0; -2,9; -2,8; ... + 2,8; +2,9; +3,0). At each of these 61 points, the probability of a correct response (P(theta)) was calculated separately for the two comparison groups using the three-parameter formula (Lord, 1980; Osterlind, 1983). The absolute value of the difference between the two P(theta) values was then determined and multiplied by 0,1 (the width of the interval), to obtain the area for that particular rectangular region. The areas over the entire ability range (-3,0 to +3,0) were then added together to obtain the overall area between the two ICCs. Using the absolute value of the difference between the two P(theta) values meant that for uniform or nonuniform DIF, all areas were added together to give the total area.

The mean values of the areas calculated for the four sets of comparison groups are provided in Table 5.10. Deciding how large an area would justify scrapping an item is somewhat subjective since no clearcut indices are provided in the literature. The general consensus is that a combination of visual inspection and empirical estimation of cutoffs should be used for flagging DIF items to be scrapped from the item pool.

Considering the nature of the LPCAT items, no bias was expected for any of the subgroups identified.

TABLE 5.10 DESCRIPTIVE STATISTICS FOR DIF AREAS BETWEEN ICCs FOR DIFFERENT COMPARISON GROUPS

DIF comparison groups	N	Mean	SD	Minimum	Maximum
Grade groups (grade 9 vs grade 11)	265	0,1789	0,1471	0,0025	1,2338
Gender groups (male vs female)	265	0,1672	0,1616	0,0089	1,4375
Culture groups (African vs White)	265	0,3307	0,2081	0,0254	1,4050
Language groups (African vs English/Afrikaans)	265	0,2336	0,1570	0,0083	0,9762

For the LPCAT, an item was considered to show DIF (ie to be biased) if the area between the two curves (uniform DIF or nonuniform DIF) exceeded 0,5. This value was determined by visual inspection of ICCs and by considering the mean areas for the various comparison sets. DIF items were discarded purely on the magnitude of the DIF indices, irrespective of the particular group against which it was considered biased or whether the DIF was uniform or nonuniform. This resulted in eight figure series items, 17 figure analogy items and 10 pattern completion items being discarded.

5.6.4 Criteria for item selection

CTT, IRT and DIF analyses were used to identify items suitable for inclusion in the final LPCAT item pool. For the three-parameter IRT model used, the general consensus among researchers (Baker, 1985; Hambleton & Swaminathan, 1985; Sands, Waters & McBride, 1997; Weiss, 1983a) is that a-values should be within the range 0,8 to +2,0

and c-values within the range 0,0 to 0,3 for items to be included in a test.

Classic item parameter values were also considered and no item with r_{it} below 0,3 was included, *unless* the a-value (IRT) for the same item was above 1,00. The condition in terms of the IRT a-value was included because items that discriminate well (a > 1,00) at a high ability level may not have high item reliability values (r_{it}), since very few examinees would get the correct answer for these items.

Items were excluded on the basis of any one of the following:

IRT: c-values: c > 0,3
IRT: a-values: a < 0,80
CTT: r_{it} < 0,3 unless IRT a > 1,0
DIF: area between the ICCs of any of the four DIF comparison groups > 0,5

Altogether 47 items were discarded on the basis of these IRT and CTT criteria, and an additional 35 items were discarded on the basis of DIF, bringing the total of discarded items to 82, or 30 percent. This percentage is comparable to the findings of similar research projects. Adaptive testing demands higher quality (more discriminating) test items than conventional testing as well as more variability in item difficulty level, and in practice only about one in three items is useful for adaptive testing (McBride, 1997).

Table 5.11 provides a summary of the number and types of items that were discarded from the LPCAT item pool.

TABLE 5.11 NUMBER AND TYPES OF ITEMS DISCARDED AS A RESULT OF ITEM ANALYSIS AND DIF ANALYSIS

Procedure	Figure series	Figure analogies	Pattern completion	Total
Item analysis (IRT & CTT)	17	15	15	47
Bias analysis	8	17	10	35

TOTAL	25	32	25	82

5.6.5 Selection and allocation of the final test items

Once the items that did not meet the selection criteria had been identified and discarded, the remaining items were allocated to the final pretest and post-test. Altogether 188 items remained (65 figure series, 58 figure analogy and 65 pattern completion items). As a first step, the remaining items of each item type were arranged in ascending order of item difficulty (b-values). Thereafter the items were allocated to the pretest and the post-test sequentially in a 1:2 ratio (one item to the pretest, and the next two to the post-test). This was done separately for each of the three item types to ensure an even spread of item types and item difficulties in the pretest and post-test. Approximately one-third of the selected items were thus allocated to the pretest (N=63) and the remainder to the post-test (N=125). McBride (1997), suggests that the number of items in the bank should exceed by a ratio of 5 or 10 to 1, the number of questions an individual examinee will encounter. For the LPCAT, the number of items in the respective item banks exceeded (by a ratio of between 5 and 8 for the pretest and by a ratio of between 7 and 10 for the post-test), the number of questions an individual examinee will encounter. Fewer items are administered in the pretest (between 8 and 12) than in the post-test (between 12 and 18). The pretest provides an initial general level of performance. In the post-test, the pretest level of performance is used as entry level, and therefore a more accurate measure of performance is possible. This requires more items at each difficulty level in the post-test. The resulting item distribution following the procedure described above is summarised in Table 5.12.

TABLE 5.12 NUMBER OF ITEMS OF DIFFERENT TYPES ALLOCATED TO THE PRETEST AND POST-TEST

Item type	Pretest	Post-test	Total
Figure series	21	44	65
Figure analogies	20	38	58

Pattern completion	22	43	65
Total	63	125	188

Descriptive statistics of the IRT item parameters for the pretest and post-test items are provided in Table 5.13.

TABLE 5.13 DESCRIPTIVE STATISTICS FOR IRT ITEM PARAMETERS OF THE LPCAT PRETEST AND POST-TEST

	Items in pretest (N=63)			
	Mean	SD	Min	Max
a-value	1,554	0,395	0,828	2,422
b-value	-0,316	0,749	-1,507	1,753
c-value	0,163	0,065	0,030	0,280
	Items in post-test (N=125)			
a-value	1,504	0,433	0,815	2,464
b-value	-0,276	0,781	-1,280	3,000
c-value	0,169	0,071	0,000	0,300

Because of the way in which items were allocated to the pretest and the post-test respectively, the mean b-parameters of items in the pretest and post-test are very similar. The mean item discrimination values (a-values) and mean pseudo-guessing values (c-values) of the pretest and post-test are also very similar.

To assess the availability of items at all ability levels, the *distribution* of b-values in the pretest and post-test is also considered. The distribution of b-values for the pretest and the post-test of the final LPCAT are presented in Figures 5.15 and 5.16. The distributions indicate that a large proportion of the LPCAT items can be regarded as being reasonably easy for someone at approximately grade 10 level (theta equal to 0).

This is in accordance with the original aim of the test as an instrument to measure learning potential, specifically aimed at people from disadvantaged backgrounds. However, there are sufficient items at higher difficulty levels to allow the adaptive procedure to select appropriate items for higher ability examinees.

FIGURE 5.15 DISTRIBUTION OF B-VALUES IN THE LPCAT PRETEST ITEM BANK

FIGURE 5.16 DISTRIBUTION OF B-VALUES IN THE LPCAT POST-TEST ITEM BANK

It should be noted that none of the examinees tested in the validation studies reached the maximum T-score of 80 on the LPCAT (see Chapter 7). This supports the claim for a sufficient number of difficult items in the item bank. Having items available over a wide range of difficulty levels, and administering items in an adaptive manner, means that the LPCAT can provide information over a wide spectrum of ability levels. A b-value of 0 can be regarded as average in difficulty level. A person with a grade 10 level of education should have a 50 percent probability of answering an item at a difficulty level (b-value) of 0 correctly. Furthermore, the average person with a grade 10 education should obtain a theta-value (ability) of about 0, which, transformed to a T-score, will be equal to 50.

5.6.6 Test information functions for the LPCAT pretest and post-test

Using IRT, one can predict certain characteristics of a test before it is administered, since the item parameters have been previously determined. Test information is an index of the precision of measurement that a test can provide. Use of IRT test information has two main advantages. The first is that it is provided in the form of a function, which allows calculation of the standard error of measurement at various levels of ability. The second advantage is that the test information is directly related to

the measurement effectiveness of a test. The test information function graphically indicates the amount of information at various ability levels, when specific items are included in a test. It is furthermore possible to compare the effect of administering various numbers of items on the information levels achieved.

Items with high discrimination values (a-parameters) provide more information and it is preferable to include such items. Highly discriminating items provide information over a narrow range of theta and little or no information outside that range, while less discriminating items provide information over a much wider range of theta. Based on the assumption of local independence, item information functions can be summed for the set of items that comprise a test to provide the test information function (TIF). The TIF shows the relative amounts of information provided by the test at each level of ability (theta) (Assessment Systems Corporation, 1989, 1995).

Reliability concerns consistency of measurement. The classical indices of reliability, namely test-retest reliability, parallel forms reliability and split-half reliability do *not* apply to computerised adaptive testing. This is because of the interactive selection of items from an item bank which results in different sets of items being administered to each examinee. The IRT equivalent to test score reliability and standard error of measurement is the test information function. Since the information function may vary from one ability level to the next, the standard error may also vary. Hence the standard error needs to be calculated for a specific ability level. In CAT, where the variance of the estimation of ability is incorporated as one element used for test termination for each individual, equal accuracy of measurement is more attainable than with standard tests.

To translate the amount of information at a specific level into a standard error of estimation, one need only take the reciprocal of the square root of the amount of test information. For example, if the maximum amount of test information is 2,383 at an ability level of 0,0, this translates into a standard error of 0,65 which means that roughly 68 percent of the estimates of this ability level fall between -0,65 and +0,65 - the calculated range around the given ability level of 0,0. Thus this ability level is estimated with a modest amount of precision (Baker, 1985).

Alternatively, the standard error of estimation (SEE) equals one divided by the square root of the information. Therefore, if the SEE is expected to be 0,40, then the information should be 6,25 at that ability level. Thus to obtain estimates of ability to the desired degree of precision across the ability scale - from -2,0 to +2,0 - items must be selected from the item pool to produce a test information function with a height of over 6,25 from -2,0 to +2,0 on the ability scale. The tails of the target information function, those sections below -2 and above +2, are not of interest to test developers and can take on any form (Hambleton & Swaminathan, 1985). Accuracy of measurement is somewhat different in adaptive testing, since no fixed number of items is selected. In adaptive testing, accuracy of ability estimation is furthermore used as a termination criterion, which helps to provide equiprecise measurement at all ability levels.

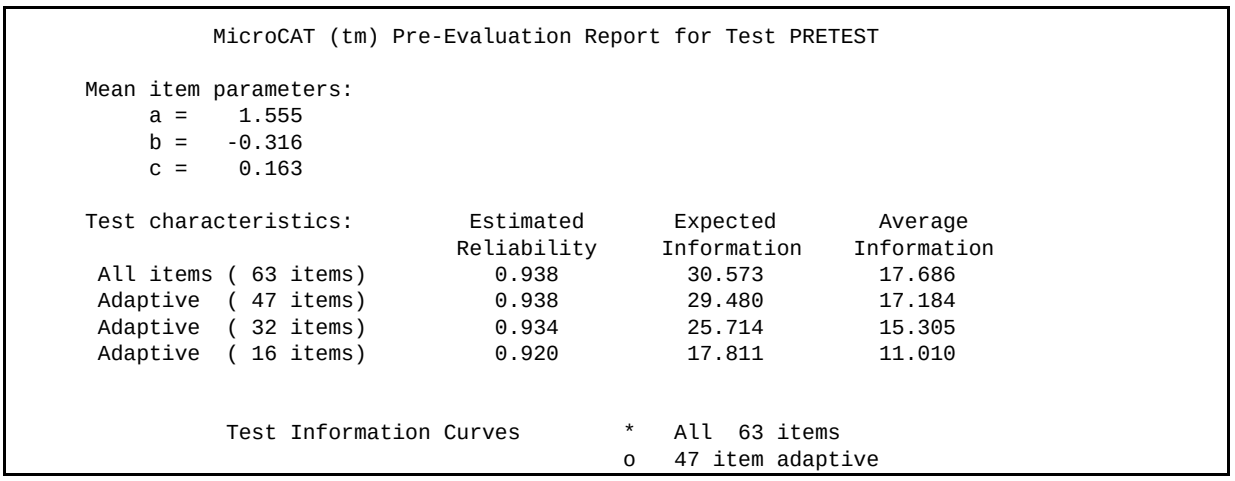
The MicroCAT test evaluation program EVALUATE (Assessment Systems Corporation, 1995) was used to obtain the test information functions as estimates of the reliability of the pretest and the post-test of the LPCAT respectively (see Figures 5.17 and 5.18). The MicroCAT test pre-evaluation program computes the average item parameters using all the items included in each test or subtest specified (Assessment Systems Corporation, 1995). The reliability is estimated from the conditional standard errors of measurement provided by the IRT model. The reliability estimates are computed for four different test lengths under the assumption that the items administered at each level of ability are those with the highest information at that level. The first is a conventional test containing all the items. The other three are theoretical adaptive tests. These three adaptive tests represent the best tests that an adaptive testing strategy could provide with one-fourth, one-half and three-fourths as many items as the full-length conventional test. For these tests, items are ranked in order of their psychometric information at each of the ability levels considered. Then, at each level, the most informative items are chosen for each short test form. *Note that the substantial decreases in test length result in relatively minor changes in test quality for the adaptive tests.* Test information functions for the LPCAT pretest and the post-test are provided separately, since the two tests function completely independently.

The expected information for the LPCAT pretest is 17,811, and for the post-test, 32,699 for 25 percent of the items administered. The standard error of estimation is

equal to the reciprocal of the square root of the information available at that theta level. The expected information for the pretest and post-test will therefore translate into standard error measures of 0,24 and 0,17 theta units respectively, which means that generally speaking, roughly 68 percent of the estimates will fall between -0,24 and +0,24 (for the pretest) or between -0,17 and +0,17 (for the post-test) from the estimated ability level in standard theta units. Translated to T-scores, this means that roughly 68 percent of the T-score estimates will fall between -2,4 and +2,4 T-scores from the estimated ability level (for the pretest), and for the post-test roughly 68 percent of the T-score estimates will fall between -1,7 and +1,7 T-scores from the estimated post-test ability level. The estimated SE values at various key theta values are provided for the pretest and the post-test respectively, in Table 5.14.

Note that the number of items that are administered in the LPCAT are slightly lower than the ones for which the information in Table 5.14 is supplied. The table values are based on 16 items adaptively administered from the pretest item bank and 31 items adaptively administered from the post-test item bank, determined by the 25 percent level used by the MicroCAT test information program (Assessment Systems Corporation, 1995). In the LPCAT pretest, between eight and 12 items are adaptively administered, while in the LPCAT post-test, between 12 and 18 items are adaptively administered.

FIGURE 5.17 TEST INFORMATION FUNCTION OF THE LPCAT PRETEST



MicroCAT (tm) Pre-Evaluation Report for Test PRETEST

Mean item parameters:
a = 1.555
b = -0.316
c = 0.163

Test characteristics:	Estimated Reliability	Expected Information	Average Information
All items (63 items)	0.938	30.573	17.686
Adaptive (47 items)	0.938	29.480	17.184
Adaptive (32 items)	0.934	25.714	15.305
Adaptive (16 items)	0.920	17.811	11.010

Test Information Curves * All 63 items
o 47 item adaptive

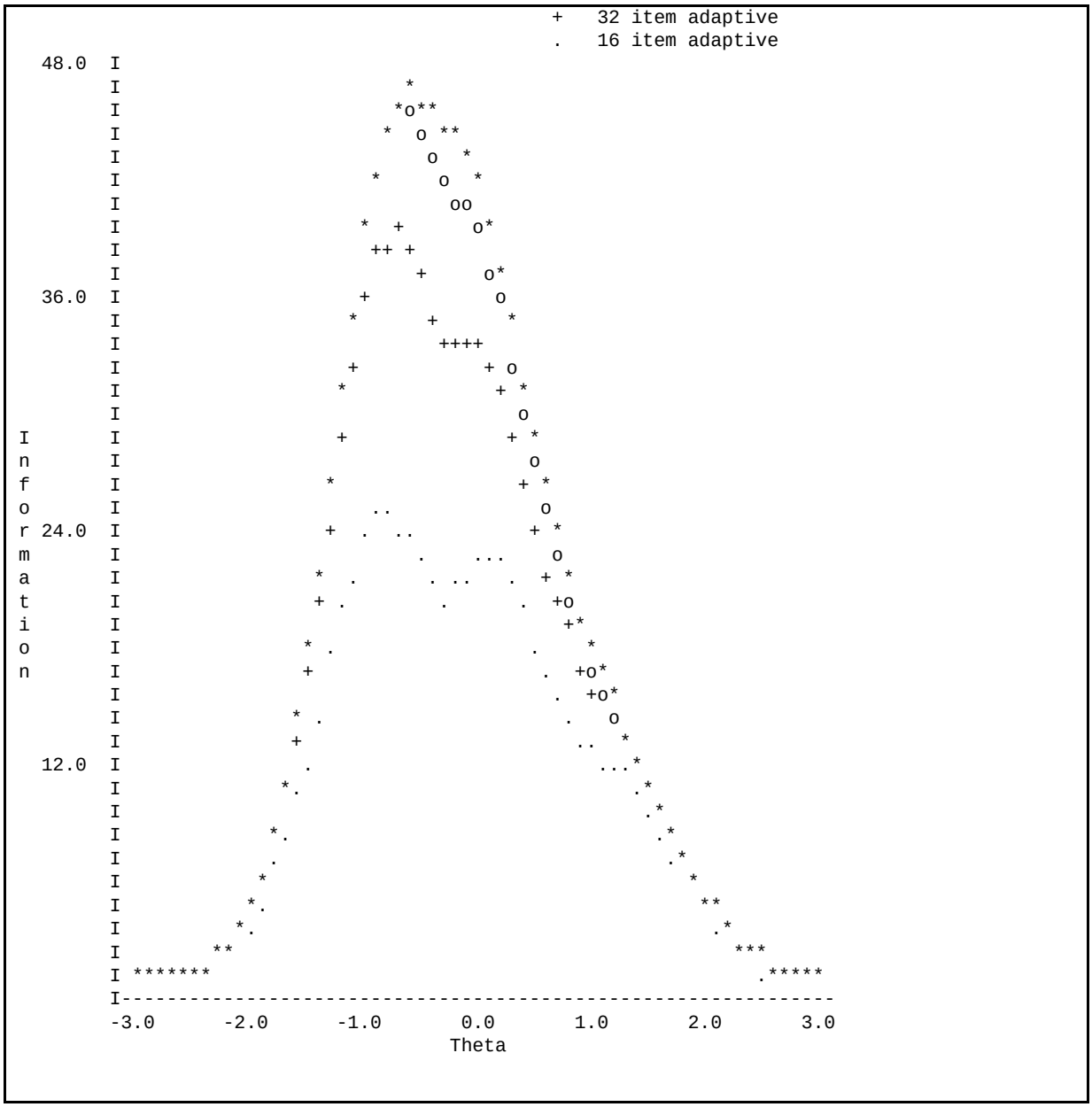

+ 32 item adaptive
. 16 item adaptive
48.0
36.0
24.0
12.0
Information
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

FIGURE 5.18 TEST INFORMATION FUNCTION OF THE LPCAT POST-TEST

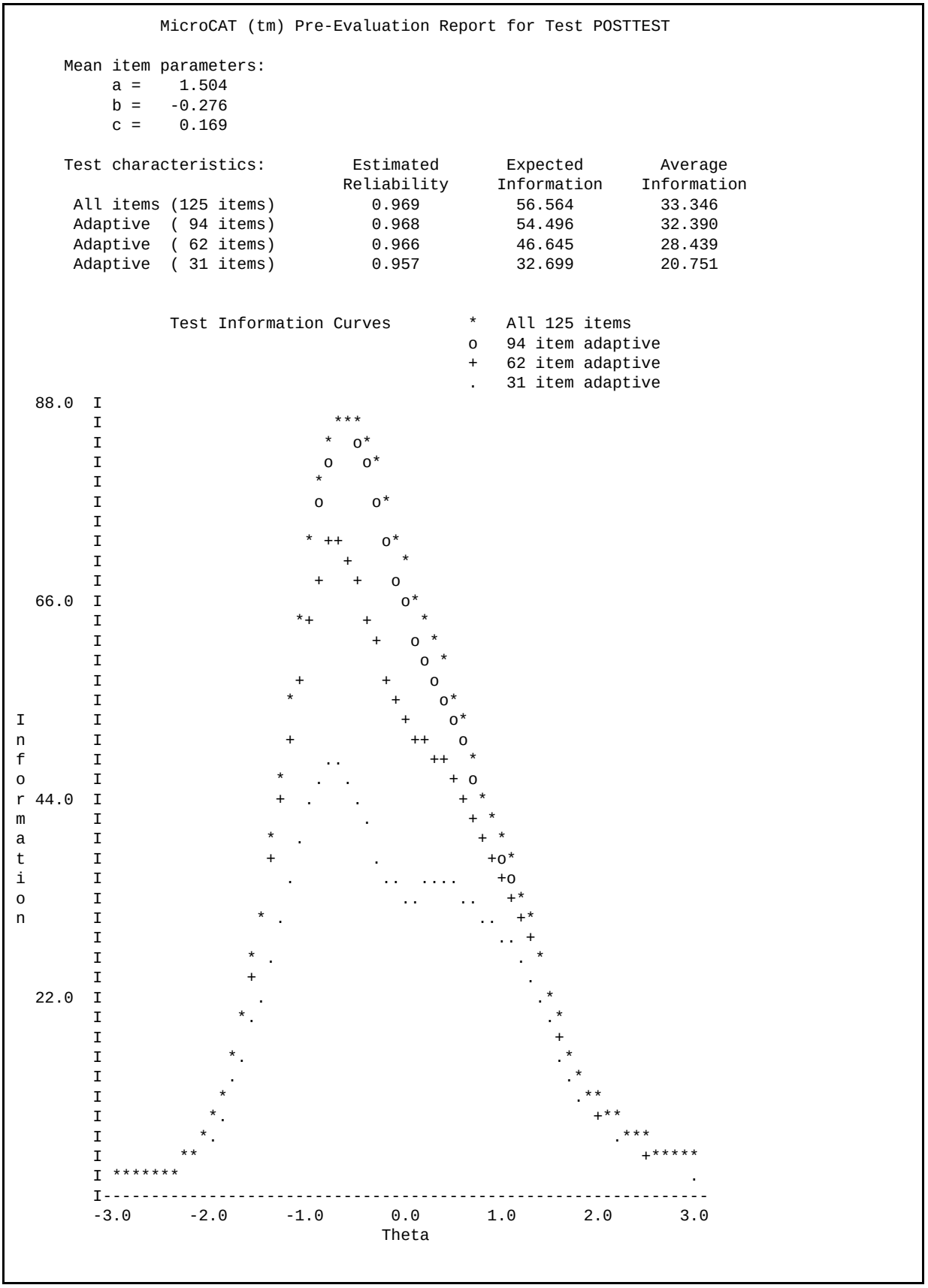


TABLE 5.14 ESTIMATED SE VALUES AT VARIOUS ABILITY LEVELS, BASED ON THE TEST INFORMATION FUNCTIONS OF THE PRETEST AND POST-TEST RESPECTIVELY

	Pretest (16 items adaptive)			Post-test (31 items adaptive)		
Theta levels	Information	SE (theta) [T-score units]		Information	SE (theta) [T-score units]	
-2,0	5	0,45	[4,5]	10	0,32	[3,2]
-1,5	10	0,31	[3,1]	22	0,21	[2,1]
-1,0	22	0,21	[2,1]	40	0,16	[1,6]
-0,5	24	0,20	[2,0]	50	0,14	[1,4]
0,0	22	0,21	[2,1]	35	0,17	[1,7]
+0,5	22	0,21	[2,1]	32	0,18	[1,8]
+1,0	14	0,27	[2,7]	30	0,18	[1,8]
+1,5	12	0,29	[2,9]	15	0,26	[2,6]
+2,0	6	0,41	[4,1]	10	0,32	[3,2]

Software to calculate the amount of information for the specific number of items administered in the LPCAT was not available. Furthermore, in adaptive testing, a varying number of items are administered. This is, however, tempered by the fact that measurement accuracy is used as the termination criterion, thereby ensuring reasonably similar accuracy of measurement at all ability levels. The fact that the information at the extremes of ability is less than in the centre region, means that more items will have to be administered to examinees who fall close to either of the extremes in their ability level. The test information functions for the pretest and post-test provided in Figures 5.17 and 5.18 indicate that more information is available at the

central regions of ability than at the extremes. This is usually what happens. It should be remembered that the LPCAT items are administered interactively, which results in the best items at each ability level being selected and administered for optimally precise measurement.

5.7 CONSTRUCTION OF THE FINAL COMPUTERISED ADAPTIVE LPCAT

In 5.1.2 and 5.1.3, general steps in test development as well as specific steps followed in the development of the LPCAT were reviewed. Having documented the definition of the LPCAT, the scoring method, item construction, item analysis administration, item analysis and item selection, this section deals with the final stages of the construction of the LPCAT.

5.7.1 Computerising the items and practice examples

The items were computerised with the MicroCAT Testing System (Assessment Systems Corporation, 1989). The Graphics Item Banker of this system allows specification and editing of characteristics associated with each item. Test items containing text, graphics or a combination of both, can be entered and edited. Among the characteristics entered are the item's unique identifier, the number of response alternatives, the correct answer and the item parameters.

Screens to introduce the examinee to the test, to familiarise him or her with the keyboard and the keys that will be used and to explain the answering procedure were computerised. No computer literacy is required of examinees. Initially, examinees are given the opportunity to locate and practise using the SPACE BAR and the ENTER KEY. These are the only two keys used throughout the test. The answering procedure, in which the SPACE BAR is used to move between the different answers and the ENTER KEY is used to choose a particular answer, is also explained and practised. Copies of these screens for the English version of LPCAT-1 can be seen in Appendix A.

After the initial introduction, practice examples are administered to familiarise the examinee with the types of items included in the test. Three screens were prepared to show the format of each of the three item types together with two practice examples for each item type to be administered before the pretest. These examples give the examinees an opportunity to practise the answering procedure, and also to familiarise themselves with the strategies used to find the correct answer. In the practice examples, feedback is provided after each answer to inform the examinee whether the answer he or she chose was the correct one. Feedback on the practice examples is individualised in that, if the examinee selects the wrong answer, it is marked with an "×" at the chosen distractor. The correct answer is indicated by a ("_") together with an explanation of why it is the correct answer. The screens to accomplish this were also prepared and computerised.

The dynamic test-train-retest format of the LPCAT involves a training section between the pretest and the post-test. In the training section, the screens for the three item types are repeated again, followed by information highlighting specific aspects that should be noted in finding the correct answers to these types of questions. More practice examples and additional training screens were prepared for this section of the test. At first, in the training section, two items of each item type are presented, each illustrating a specific aspect to be noted when looking for patterns and features to solve the items. The correct answers are already indicated for these examples.

In the LPCAT-1 version, where the examinee reads and follows the instructions and explanations independently, four items to check the understanding of both the concepts and the terms used in the explanation are administered after the initial training screens. These "language" items are scored and a percentage mark allocated. If an individual answers any of these extremely easy questions incorrectly, it probably indicates that he or she did not understand the terms and/or concepts used in the feedback and training. Limited understanding of the instructions and feedback may consequently have affected the results negatively. It is important to emphasise that the LPCAT-1 (text on screen version) should be used only for people who are adequately proficient in either English or Afrikaans to enable them to read and understand independently the instructions and feedback provided. A reading proficiency level of

grade 6 for English or Afrikaans should be sufficient.

To conclude the training section, another seven practice examples including items of each of the three item types, are administered. Feedback on the answers chosen is provided again. This time the feedback on the examples is very specific, in that it takes the specific distractor that was chosen into account in the feedback provided. All examinees are given exactly the same examples in the training. The reason for not administering the training and practice examples adaptively, is mainly to standardise the training, thereby simplifying comparison of test scores of different individuals. Another reason is that standard training improves the perceived fairness and face validity of the test. In the training and with the examples that are provided, an effort was made to highlight the principal strategies to be used to answer the questions. The feedback, hints and guidelines that are provided are intended to assist the examinee to learn how to solve these types of problems. A copy of the English instructional and practice example screens and feedback is provided in Appendix A. Complete instructions are provided in the *LPCAT user's manual* (De Beer, 2000a).

5.7.2 Overall structure of the LPCAT-1 and LPCAT-2

The LPCAT has a test-train-test format, with two independent CATs and a standard training session in between. There are no overall test time limits, although, for practical reasons, a maximum screen time of three minutes per test item was built into the test. Owing to the adaptive nature of administering items that are suited to the estimated ability level of the examinee, this time limit is rarely exceeded. For practical purposes the LPCAT can be regarded as an untimed power test. Because of the properties of IRT-based CAT, the scores on the pretest and post-test are directly comparable and on the same scale, although items are administered independently for each examinee.

After initial administration of the LPCAT to mostly post-grade 12 examinees, a need was identified to allow testing at lower levels of literacy. Because lack of language proficiency and/or reading skills could prevent individuals from adequately understanding the test instructions independently, it was decided to construct a second

version of the LPCAT. This second version uses exactly the same examples and item banks and the same basic procedure, but all text is removed from the screen. A new set of test instructions was prepared for this version so that the instructions and feedback on practice examples could be read to the examinee(s). The screens and the instructions are numbered so that it is clear which instructions have to be read at which screen. The instructions were also translated into the other nine official languages (besides English and Afrikaans). A copy of the English version of these screens and instructions are provided in Appendix B.

To distinguish between the two versions, the initial LPCAT version with the text on the screen was named LPCAT-1 and the version in which the instructions are read to the examinee (no instructions or feedback on the screen), LPCAT-2. Items for the LPCAT-1 and LPCAT-2 are selected from the same two item banks, and the same interactive testing procedure is used.

5.7.3 Choice of starting point

In adaptive testing, the entry level of difficulty of the first item to be administered can be specified when the test is constructed. This means that if a group of examinees' approximate level of ability is known beforehand, the test can be constructed in such a way that from the very first item, the difficulty level of the items is appropriate for the examinees' ability level. In general, an item of average difficulty is usually presented first, after which the adaptive item selection process commences. In the case of the LPCAT-1, the difficulty level (entry level) of the first item was set at 0, which is the mean value on the theta scale. Hence an item of average difficulty level for an average grade 10 level examinee will be administered first. In the LPCAT-2, which will probably be used for examinees with either lower ability levels or lower educational levels, the entry level was set at -1,0 on the theta scale. The result is that an easier first item is administered, whereafter the adaptive testing process continually matches the difficulty level of items to be administered to the examinee's estimated ability level. Although the two entry level items differ in difficulty level for the LPCAT-1 and the LPCAT-2 respectively, through the adaptive testing process, examinees with any level of ability

can be tested with any of the two test forms. The level of reading proficiency of the examinee will determine which version is most appropriate.

5.7.4 Selection of test items

The two common methods for estimating examinees' theta levels are maximum likelihood estimation and Bayesian modal estimation. The Bayesian procedure, which was used for the LPCAT, is widely applied in adaptive testing programs (Hankins, 1990). In the Bayesian item selection strategy, the item pool is searched to find the one item which, when administered, will maximally reduce the posterior variance of an individual's ability estimate (Weiss, 1983b). Items are selected so that the estimated posterior variance is minimised after each item administration using a complex set of formulae to re-estimate the individual's ability each time (Hankins, 1990). With the MicroCAT procedure (Assessment Systems Corporation, 1989, 1995), when items are administered using the adaptive testing process, the following information is automatically available for each item that is administered:

a counter number of the item just administered - keeping track of the number of items that are administered
the unique identifying item bank number of the item just administered
the distractor (answer) chosen by the examinee
the correct distractor
the Bayesian mean (estimated ability level on the theta scale)
the Bayesian variance (accuracy of ability estimation)
the time that has elapsed since the start of the test

In order to utilise the MicroCAT Bayesian computerised adaptive testing procedures for constructing a CAT, the following information had to be provided for the LPCAT pretest and post-test respectively:

a list of items for the item pool from which items can be selected
the variance of ability estimation to be used as the termination criterion
the minimum number of items to be administered
the maximum number of items to be administered

(Assessment Systems Corporation, 1989, 1995)

During the CAT procedure, items are sampled without replacement from the specified pool and administered to the examinee until one of the termination criteria is reached.

5.7.5 Stopping rule

In order to standardise the administration procedure of the computerised adaptive LPCAT for its validation, a fixed number of 10 items were administered in the pretest and 16 in the post-test. On the basis of the results of the standardisation, the true adaptive procedure was later built into the test, also in terms of the number of items administered. Adaptive test termination depends on accuracy of measurement while a minimum and maximum number of items to be administered is also specified. According to Weiss (1983a), short adaptive tests of about 15 items are sufficiently reliable for general assessment purposes and additional items do not yield psychometric returns proportional to the added administration time required. In the LPCAT standardisation, the empirically obtained mean variance associated with ability estimation after the 10 pretest items was 0,1143, while the mean variance after the 16 post-test items was 0,04355. These values were used in the construction of the final LPCAT version where the cutoff in terms of variance was put at 0,10 for the pretest and 0,05 for the post-test respectively. These values translate to a standard error of 0,31 for the pretest and 0,22 for the post-test, which is in line with that of other adaptive tests (Hankins, 1990). These cutoffs entail that once the set minimum number of items has been answered in the CAT (pretest or post-test respectively), the termination criterion is either that the set maximum number of items has been administered, or alternatively, that the predetermined level of accuracy of ability estimation has been reached and surpassed. For the final version of the LPCAT, between eight and 12 items are administered in the pretest and between 12 and 18 in the post-test.

5.7.6 Scoring

The fact that the difficulty of items and examinees' ability level are measured on the

same scale (the theta scale) allows for the interactive selection of items during test administration which is characteristic of computerised adaptive testing. The theta scale has a mean of 0 and standard deviation of 1, but can be transformed to any other standard scale. In the case of the LPCAT, scores are transformed to T-scores, stanines and percentile scores. The pretest and post-test are scored separately.

The LPCAT results consist of four different scores, namely:

the pretest score (T-score, stanine, percentile score)
the post-test score (T-score, stanine, percentile score)
the difference score (T-score)
the composite score (single score on T-score scale)

The pretest score represents the level of performance at the end of the pretest, which is indicative of the actual developmental level in Vygotsky's ZPD terminology. The post-test score represents the potential level of performance, while the difference score represents the ZPD. The composite score is a combined score which provides a single score incorporating both the pretest level of performance and a proportional credit for the ZPD, depending on the level of performance. For the pretest and the post-test scores, the accuracy of estimation is also provided in the form of a posterior variance score. This score, when transformed to a standard deviation score, can be used to provide a band (range) within which the actual score falls with a particular probability.

These four scores allow for a richness of interpretation not possible with conventional tests. The following four cases illustrate the different ways in which the emphasis on specific scores can shift, depending on the context.

Pretest as focus

In cases where a certain existing level of performance may be required, the pretest score may be considered to be most important. One scenario is selection for extremely expensive training, necessitating a certain level of present ability level as a prerequisite for the training to be offered.

ZPD as focus

In other cases, where the focus may be on affirmative action or individual development, the most emphasis may be placed on the difference score, since this will indicate those individuals who are likely to benefit most from training, irrespective of their present level of performance. In such a case, the desire to provide opportunities for those who manifest the most potential to change their existing levels of performance may lead to the ZPD (difference score) being regarded as the most important score.

Post-test as focus

In some cases, the focus may be on the performance level after training, which may include the aim of affirmative action or development of disadvantaged individuals. In such cases, the pretest score may be ignored - on account of unequal prior opportunities for learning - and only the post-test score may be taken into consideration. This would mean that both previously advantaged or previously disadvantaged individuals can be selected with selection being based solely on performance after relevant training has been provided.

Overall performance as focus

Although it is suggested that all of the above scores (pretest, ZPD and post-test) be taken into consideration when assessing performance, it can be cumbersome to work with three separate scores and comparison between individuals can be difficult for the user. A single composite score that represents a justifiable and reasoned combination of the first three scores, allows for easier comparison of the cognitive developmental level of different persons.

In an attempt to allow for the fact that the same ZPD at different ability levels cannot be interpreted in exactly the same way, a composite score which combines both the actual level of performance and some portion of the ZPD is provided. The same ZPD at different initial levels of performance does not have the same meaning. A small ZPD at a *low* initial level of performance indicates a lack of learning potential owing to the combination of both small ZPD *and* low level of performance. However, a small ZPD at a *high* initial level of performance may result mainly from the fact that the examinee is already close

to the maximum attainable performance level and in this case does not indicate lack of learning potential. The maximum attainable performance level will be about three on the theta scale and about 80 on the T-scale. The idea underlying the composite score is that the amount of credit given for the size of an examinee's ZPD (or LPCAT difference score), should be adjusted on the basis of the maximum ZPD that he or she could theoretically have attained from his or her pretest level of performance. Denoting the composite score by C, the pretest score by I, and the difference score by D, the composite score (on the theta scale) is defined as

$$C = I + \frac{D^2}{(3 - I)}$$

On the T-scale (with scores C, I and D in terms of T-scores), the composite score will be defined as

$$C = I + \frac{D^2}{(80 - I)}$$

Irrespective of what an examinee scores initially, it is assumed that he or she can improve up to a maximum of three on the theta scale or up to a maximum of 80 on the T-scale. Hence the maximum possible improvement on these two scales will be (3 - I) or (80 - I) respectively. The difference score is then expressed as a proportion of this maximum: D / (3 - I) {for the theta-scale} or D / (80 - I) {for the T-scale}. Credit is given for this proportion of the difference score, that is for an improvement of D^2 / (3 - I) {theta-scale} or D^2 / (80 - I) {T-scale}. Adding this proportional credit to the initial score provides the composite score as indicated in the formulas above. This composite score allows for the fact that

(1) the level of initial performance contributes to the ability to learn and should therefore be taken into consideration in any learning potential score (ie top level performance with a very small ZPD does not imply no learning

potential) and

(2) the size of the ZPD at different ability levels does not have exactly the same meaning (ie improvement at top level is much more difficult to attain than at lower levels)

Some researchers (Babad & Budoff, 1974; Budoff, 1969; Budoff, 1987a, 1987b; Budoff & Corman, 1974; Lidz, 1991) use a post-test score adjusted for the pretest level (ie a residualised score) which is in a way similar to the composite score used for the LPCAT. These researchers' interpretation, starting from post-test performance and making an adjustment downward for pretest level of performance, does not strictly adhere to Vygotsky's proposed use of the pretest score and the ZPD for interpretation. The LPCAT composite score is similar to this residualised score, but starts from the pretest score and represents a more conservative estimation of learning potential. The latter interpretation takes into account that the optimal conditions usually provided in the dynamic testing situation do not always materialise in real-life training situations. Also, the entire premise of learning potential is based on looking *forward* from the present level of performance while making allowance for improved performance following relevant training.

The advantage of the LPCAT composite score is that people at different levels of initial performance and with different ZPDs can be compared in a systematic manner. In cases where there is no improvement following training and the pretest score is higher than or equal to the post-test score, the pretest score is taken as the composite score.

5.8 CONCLUSION

Despite growing interest and increased research activity in dynamic testing (Grigorenko & Sternberg, 1998), measurement and administration problems of these instruments have hampered progress. Computerised adaptive testing, based on IRT, can address most of the problems that have been identified. Firstly, it allows shorter testing times compared with standard tests, without forfeiting measurement accuracy. It also allows administration of independent pretests and post-tests, with items interactively selected

to match the examinee's estimated ability level, thereby further improving measurement accuracy and testing efficiency.

The LPCAT has combined the dynamic (test-train-test) approach to the measurement of learning potential with IRT-based CAT. In the construction of the LPCAT, two independent and separate CATs are used for the pretest and the post-test. However, because measurement is on the same scale, these scores are directly comparable, thus improving the measurement accuracy and psychometric soundness of the LPCAT.

The research design and results of the study that dealt with the validity investigation of the LPCAT will be discussed in chapters 6 and 7 respectively.

CHAPTER 6

PROCEDURE FOR EVALUATING THE VALIDITY OF THE LPCAT

6.1 INTRODUCTION

This chapter deals with the planning and execution of the empirical research for gathering validity information for the LPCAT - for both LPCAT-1 and LPCAT-2 versions. The validity of a test concerns what the test measures and how well it does so, mostly by assessing the relationships between performance on the test and on other measures of the behaviour under consideration. Although the validity of a test cannot be reported in general terms (ie a test cannot be described as having high or low validity), the validity information tells us what can be inferred from test scores with reference to the particular use for which the test is being considered (Anastasi & Urbina, 1997). The aim of the LPCAT validity investigation is to evaluate its usefulness as a measure of learning potential within the general nonverbal reasoning domain, including an evaluation of its fairness as a cross-cultural measure of general reasoning ability.

The construction and evaluation of the LPCAT comprised two distinct phases, both of which concern specific aspects of validity.

- The first phase, namely test development, was reported on in chapter 5. This phase dealt with the construction of test items, IRT and CTT item analysis as well as DIF analysis for item selection to help ensure the psychometric soundness of the LPCAT. Large and representative samples were used for the item analysis in compliance with the requirements for the use of three-parameter IRT analysis. Selected items were allocated to the pretest and post-test in such a way that an even distribution was achieved in terms of item type and item difficulty. The reliability indices of the pretest and the post-test were reported in the form of two separate test information functions. In terms of test validity, the first phase involved the evaluation of content validity and face validity, which will be discussed in 6.3 of this chapter.
- The second phase, involved the administration of the two versions of the LPCAT

in computerised adaptive form in order to gather empirical information in support of its validity in various contexts and with different sample groups. The samples, measuring instruments used and measures obtained as well as the procedures followed, are described in this chapter. The results of this process of obtaining empirical results with the LPCAT together with other relevant results, are discussed in the next chapter. In terms of test validity, this second phase concerns mainly criterion-related validity.

Validity is built into a test from the outset and is not limited to the last stages of test development. The ways in which validity is addressed throughout the test development process as described by Anastasi and Urbina (1997) can be summarised in the following steps:

(1) formulation of a detailed trait definition derived from psychological theory
(2) preparation of test items to fit the construct definition
(3) empirical analysis for selecting the most effective or valid items from the initial item pools
(4) further internal analysis of test items, subtests, et cetera
(5) validation of scores and interpretive combinations of scores through statistical analyses against external real-life criteria
(6) investigation of factors such as test bias

The first four steps were dealt with in the previous chapter dealing with the construction of the LPCAT. The constructs to be measured, the item types used as well as item analysis for item selection were discussed in chapter 4. Content and face validity, which form a part of the test development phase, will be discussed in this chapter.

This chapter also deals with the planning and operationalisation of steps 5 and 6. The results of these investigations will be reported and discussed in chapter 7. Before providing information on the samples, the measures obtained and the procedures involved in the validity investigation of the LPCAT, an overview of validity evaluation in general will be provided.

6.2 AN OVERVIEW OF VALIDITY EVALUATION IN GENERAL

Validity information assists in the prediction process, by providing useful information on which to base decisions. Validity needs to be systematically addressed with a variety of validity evidence being integrated to provide the required support for the use of an instrument. Messick (1994) describes test validation as the empirical evaluation of the meaning and consequences of measurement involving a combination of scientific inquiry and rational argument to justify test interpretation and use. The trend has recently been to view construct validity as the fundamental and all-inclusive validity concept that specifies what the test measures. This includes content validity and predictive validity as sources of information contributing to the understanding of the constructs assessed by a test (Anastasi & Urbina, 1997). Although this presents a more integrated evaluation of validity, the traditional concepts and terms, namely content validity and predictive validity have survived and are still used in the comprehensive evaluation of construct validity. Traditionally, validity has been divided into three separate types, namely content, criterion and construct validity:

6.2.1 Content validity: using content-description to evaluate content relevance

Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn. It involves the systematic examination of test content, and is usually judged by a panel of experts in the field. Content validity thus provides judgmental evidence in support of domain relevance and the representativeness of the content and does not provide evidence to sustain inferences made from test scores (Messick, 1994). According to Anastasi and Urbina (1997), content validity is built into the test from the outset by the choice of appropriate items and by using test specifications indicating the number of items of each kind to be prepared. Face validity, although strictly speaking not validity in the technical sense, is nevertheless also considered to be important and is related to content validity since it has to do with whether the test “looks valid” to examinees and other interested parties.

6.2.2 Criterion-related validity: using criterion-prediction procedures to evaluate predictive utility

Criterion-related validity is evaluated by comparing the test scores with one or more external variables or criteria which provide a direct measure of the characteristic or behaviour in question. This is generally done by using test scores to identify present performance on a criterion (concurrent validity) or to predict future performance on a criterion (predictive validity). A test may be validated against any criterion that is considered to be useful, depending on the intended use of the test. The correlation of a new test against previously available tests is also often included as evidence of criterion-related validity. The main aim of criterion-related validity is to assess the practical validity and utility of a test for a specified purpose.

6.2.3 Construct validity: using construct-identification procedures to evaluate the general meaning and utility of test scores

Construct validity involves the evaluation of the extent to which the test measures the psychological construct it purports to measure. It concerns the theoretical underpinning of the test and the accumulation of evidence from a variety of sources that help to illuminate what the test measures. Construct validity is evaluated by investigating what qualities a test measures by determining the degree to which certain explanatory concepts or constructs account for performance on the test. It places the focus on the role of psychological theory in test construction involving a process of gradual accumulation of information from a variety of sources, all throwing light on the nature of the trait being measured and helping to build up the concept of the behaviour domain sampled by the test (Anastasi & Urbina, 1997).

Typical methods used to evaluate construct validity include the assessment of

- the factor structure of the test
- internal consistency

- convergent and discriminant validity
- correlations with other tests
- developmental changes (age differentiation)

Messick (1994) warns that the evidence usually gathered in validity investigation of a test, may or may not include pertinent specific evidence of the relevance of the test for the particular purpose and the utility of the test in a particular applied setting. Validity coefficients are affected by conditions such as the nature of the group, sample heterogeneity, preselection, and the form of the relationship between test and criterion.

6.3 PLANNED VALIDITY EVALUATION OF THE LPCAT

The validity study of the LPCAT will be presented in four sections, namely:

(1) General LPCAT validity evaluation based on test construction information
(2) LPCAT-1 validity evaluation based on the information obtained from two Technikon samples and one school sample
(3) LPCAT-2 validity evaluation based on the information of an adult learner sample and a school sample
(4) Further LPCAT validity evaluation information based on specific investigations regarding selected groups or combination of groups

The first section on general validity involves face validity and content validity, involving the content relevance of both the LPCAT-1 and LPCAT-2. For the next two sections that deal with the specific validity of the LPCAT-1 and LPCAT-2 respectively, the samples, measures and procedures used will be described in the present chapter. These sections focus on the measures obtained to provide information on convergent validity, criterion validity, comparison of mean scores for important subgroups and regression analysis for the respective samples used. The LPCAT results were evaluated against academic and other criteria including standard cognitive test results. Results will be reported separately per sample group because the samples were from different training institutions and because of the difficulty of equating or even

comparing criterion scores from different institutions. The groups used were selected to represent different levels of academic attainment and were also included because results on other existing standard tests of cognitive ability could be obtained for most of them. In the last section, information from specific samples or combinations of samples is used to provide additional validity information for the LPCAT. The additional sample used for the investigation of the effect of different forms of training will also be described in this chapter.

6.4 GENERAL VALIDITY OF THE LPCAT

6.4.1 Evaluation of LPCAT face validity

Although face validity is somewhat subjective and not strictly a technical form of validity, as a concept, it is still important. It refers to the obvious and more superficial evaluation by users as to whether the test, on inspection of its content, seems relevant in terms of what it is supposed to measure. It involves a form of social acceptability of the test and concerns overall satisfaction with and acceptance of test results. Face validity relates to whether the test "looks valid" to examinees, administrative personnel who decide on its use as well as other technically untrained observers (Anastasi & Urbina, 1997).

On the basis of the following factors, the LPCAT can be expected to have good face validity:

(1) Only universally known figures and concepts (geometric figures, size, form, shading, etc) are used.
(2) The items are not related to language proficiency or school material and can therefore measure learning potential in a more culture-fair manner.
(3) The test includes relevant learning, providing an opportunity to show to what extent performance can be improved in a test-train-retest procedure.
(4) The answering procedure is simple and only two keys are used throughout to answer the multiple-choice format questions.
(5) The test instructions are available in all 11 South African official languages.

(6) There is no time limit for the test - that is, it is a power test and not a speeded test.

6.4.2 Evaluation of LPCAT content validity

Content validity is determined by the extent to which the tasks (content) of the test represent the universe of behaviour that the test has been designed to sample (Gregory, 1996).

The LPCAT items, which involve three different types of figural, nonverbal items, provide acceptable content validity for measuring g_f, similar to other culture-fair tests like Raven's Progressive Matrices (Raven et al, 1977) and Cattell's Culture-fair Intelligence Test (Cattell, 1963). The questions were evaluated by a panel of experts from the Human Sciences Research Council and approved for assessing general, nonverbal reasoning ability by means of figural content. The reasoning required to solve these types of items involves general processes of identification, comparison and completion of sequences while keeping track of certain basic features such as size, form, number, rotation, et cetera. The training that is provided involves these basic building blocks to solve similar questions.

6.4.3 Factorial validity

One method of evaluating construct validity is to do factor analysis, which investigates the factorial composition of the test for the total group as well as for different subgroups. The LPCAT factor analysis results were discussed in chapter 5 as part of the test construction process. The results indicate a one-dimensional factor structure for the total group as well as for specific subgroups. The factor analysis therefore indicates construct validity in that the same theoretical construct is indicated for the total group and for various important subgroups.

6.4.4 Internal consistency

A minimum requirement for construct validity is to demonstrate the internal consistency of the test to ensure that it measures a single construct. This method uses the total score on the test as a criterion measure for assessing item performance, providing a measure of the homogeneity of the test. The internal consistency of the LPCAT was discussed in chapter 5 where the development of the test was reported. The coefficient alpha indices for all the LPCAT items for the total group as well as for important subgroups are all in the 0,90s (see 5.6.1) and indicate high homogeneity of items, thereby providing further support for the one-dimensional nature of the LPCAT. After item selection, the test information functions for the pretest and post-test respectively, also provided estimated reliability indices based on internal consistency. These measures were also in the 0,90s for the pretest and post-test of the LPCAT (see 5.6.6).

6.5 VALIDITY OF THE LPCAT-1

Three different samples were used to obtain validity information for the LPCAT-1. Two Technikon first-year samples from two different Technikons were used as well as a group of grade 9 school pupils. For convenience, the two Technikon groups will be named Group 1 and Group 2 respectively, and the school grade 9 group, Group 3. The samples, measures and procedures are described in the following subsections.

6.5.1 Group 1 for LPCAT-1 validity investigation

6.5.1.1 Sample for Group 1

The first of the three groups used to investigate the validity of the LPCAT-1 consisted of 92 first-year Technikon students in the Science and Engineering faculties. Arrangements for testing the first-year Science and Engineering students were made with the person responsible for student guidance. Although a concerted effort was made to involve all eligible students, testing was nevertheless voluntary. Most of the

first-year students from the targeted faculties took part in the study, but the sample cannot be regarded as being statistically representative, since not all students were tested and random procedures were not used in the selection of participants.

The mean age of the group was 19,8 years. The language distribution of this group was 50 percent African home language and 50 percent English/Afrikaans home language. The gender distribution was 11 percent female (N=10) and 89 percent male (N=82). The home language by gender distribution of this group is provided in Table 6.1.

TABLE 6.1 GROUP 1: HOME LANGUAGE BY GENDER CROSS-TABULATION

Home language/ Gender	African home language	English/Afrikaans home language	Total
Male	36	46	82
Female	10	-	10
Total	46	46	92

6.5.1.2 Measures obtained for Group 1

The LPCAT-1 was administered to the examinees, each examinee choosing to receive the test instructions in either English or Afrikaans on the screen. Results for the LPCAT-1 included a pretest score, a post-test score, a difference score and a composite learning potential score. The following other measures were obtained to assess the validity of the LPCAT-1:

- The General Scholastic Aptitude Test (Senior) (Claassen, De Beer, Hugo & Meyer, 1991) is a standard cognitive test which provides a verbal, nonverbal and total score. The verbal section consists of three subtests namely Word Pairs, Word Analogies and Verbal Reasoning, while the subtests of the nonverbal section are Figure Analogies, Number Series and Pattern

Completion. The verbal, nonverbal and total scores provided are on a scale with a mean of 100 and standard deviation of 15. The test-retest reliability of the scores provided by the GSAT range from 0,84 to 0,95. The parallel form reliability ranges from 0,89 to 0,96. Correlations of the GSAT with other intelligence tests range from 0,73 to 0,86 while correlations of the GSAT with scholastic achievement range from 0,33 to 0,86.

- Matriculation (grade 12) results in English, Mathematics and Science were obtained for most of the examinees. These subjects can be taken at either higher or standard grade. The results were available in symbols only, and a transformed score was thus calculated, taking into account whether the subject was taken at higher or standard grade. According to Claassen et al (1991, p 53) "... insight into the subject content is more important for good achievement in a higher grade paper than in a standard grade paper ... A subject on the higher grade is marked out of 400, whereas a subject on the standard grade is marked out of 300." It is possible to use the actual marks that were obtained as a criterion and to include both higher grade and standard grade examinees in one group on the assumption that marks obtained on the higher grade and the standard grade are equivalent. It is therefore assumed, for example, that 180 out of 300 (60%) in the standard grade, equals 180 out of 400 (45%) in the higher grade. Schools use this principle as the basis for calculating a pupil's mean percentage. The mean score of a particular symbol (ie 85 for an A symbol, 75 for a B symbol, etc) was multiplied by either 4 for a higher grade total out of 400 or by 3 for a standard grade total out of 300 in order to obtain a single scale for all scholastic results.
- First-year academic results in the form of percentage scores obtained in the November end-of-year examination were also obtained.

6.5.1.3 Procedures followed for obtaining validity information for Group 1

LPCAT testing took place early in March 1996. For Group 1, the computerised tests were administered on laptop computers and a roster was set up in advance to accommodate the students in smaller groups for the testing. School results from the

November matriculation examinations of the previous year were obtained from the student records. The GSAT had been administered at the beginning of the academic year in January/February by Technikon personnel, approximately one month before the LPCAT was administered. The school academic results and GSAT results were provided by the student guidance personnel from their student records. First-year academic results from the November end-of-year examinations were obtained at the beginning of the following year.

6.5.2 Group 2 for LPCAT-1 validity investigation

6.5.2.1 Sample for Group 2

The second group used for the LPCAT-1 validity investigation was also a Technikon first-year sample (N=223), from another South African Technikon. On initial contact and provision of information about the research project, the person responsible for student guidance at this Technikon expressed a keen interest in participating in the research. The Technikon was experiencing problems with their selection procedures and was interested in investigating new measures that could possibly be included in their existing test battery. Although an effort was made to include all the first-year Science and Engineering students in the sample, testing was nevertheless voluntary. The sample size is adequate for the purpose of obtaining validation information and can be considered to be reasonably representative of the first-year population of Science and Engineering students at this Technikon.

The mean age for this group was 19,9 years. The language distribution of this sample was 48 percent African home language students and 52 percent English/Afrikaans home language students. Of the latter group, most were English-speaking Indian students. In terms of gender distribution, 55 percent of the sample were male and 45 percent female. The home language by gender distribution of the sample is provided in Table 6.2. Note that some biographical information was incomplete, so that the numbers in the various tables do not always add up to the sample total.

TABLE 6.2 GROUP 2: HOME LANGUAGE BY GENDER CROSS-TABULATION

Home language/ gender	African home language	English/Afrikaans home language	Total
Male	57	52	109
Female	37	51	88
Total	94	103	197

6.5.2.2 Measures obtained for Group 2

The LPCAT-1 was administered to the examinees, providing for each examinee a pretest score, a post-test score, a difference score and a composite learning potential score. Each student could choose to receive the test instructions of the LPCAT-1 on the screen in either English or Afrikaans. The following other measures were obtained to assess the validity of the LPCAT-1:

- The GSAT-CAT (Van Tonder & Claassen, 1992) is the computerised adaptive version of the General Scholastic Aptitude Test (Senior). The subtests included in the verbal and nonverbal sections of the GSAT-CAT are the same as those of the paper-and-pencil GSAT described for Group 1. Since the GSAT-CAT was constructed as an equivalent version of the paper-and-pencil GSAT, the reliability and validity of the paper-and-pencil GSAT are assumed to apply to the GSAT-CAT also (Van Tonder & Claassen, 1992).
- Revised but unnormed subscales of the SAT (Senior Aptitude Test) (Owen & Taljaard, 1989) were also administered. Only three subtests were used, namely calculations, spatial 3D and mechanical insight, each of which consists of a total of 25 multiple-choice questions. Because standardisation of these revised subtests had not been completed at the time of administration, raw scores were used.
  - "Calculations" measures the ability to work quickly and accurately with numbers by doing the four basic operations in mathematics.
  - "Spatial 3D" measures spatial perceptual ability.

- “Mechanical Insight” measures the ability to solve problems of a mechanical nature.

- Matriculation (grade 12) results in English, Mathematics and Science were obtained for some of the students. The school matriculation results were available as symbols only and these were converted to number scores in the same way as described for Group 1.
- For the first-year results, the percentage scores for subjects obtained in the November end-of-year examination were used.

6.5.2.3 Procedures followed for obtaining validity information for Group 2

For Group 2, testing took place in August 1996. The computerised LPCAT-1 and GSAT-CAT tests were administered on a network system which had 30 workstations available. This meant that 30 students at a time could be tested. Arrangements were made to test students in two separate venues, one for the paper-and-pencil SAT subtests and the other for the computerised LPCAT-1 and GSAT-CAT tests. To simplify practical arrangements, groups of 60 students were tested at a time, half with the paper-and-pencil tests and the other half with the computer tests. At the end of the test session, the two groups were exchanged to do the alternate session of testing. The school results from the November matriculation examinations of the previous year were obtained from the student records. First-year academic results from the November end-of-year examinations were obtained at the beginning of the following year.

6.5.3 Group 3 for LPCAT-1 validity investigation

6.5.3.1 Sample for Group 3

The third group used to investigate the validity of the LPCAT-1 consisted of 37 grade 9 high school pupils from an urban high school. Arrangements for testing were made with the school guidance teacher. This group was randomly selected from the total grade 9

school class.

The language distribution of this group was 48,6 percent African home language and 51,4 percent English/Afrikaans home language. The gender distribution was 54 percent female (N=20) and 46 percent male (N=17). The home language by gender distribution of this group is provided in Table 6.3.

TABLE 6.3 GROUP 3: HOME LANGUAGE BY GENDER CROSS-TABULATION

Home language/ Gender	African home language	English/Afrikaans home language	Total
Male	10	10	20
Female	8	9	17
Total	18	19	37

6.5.3.2 Measures obtained for Group 3

The LPCAT-1 was administered to the examinees, each examinee choosing to receive the test instructions in either English or Afrikaans on the screen. Results for the LPCAT-1 included a pretest score, a post-test score, a difference score and a composite learning potential score. The following other measures were obtained to assess the validity of the LPCAT-1:

- School academic results in the form of the average marks for the four terms as well as an overall year average percentage mark

6.5.3.3 Procedures followed for obtaining validity information for Group 3

LPCAT testing took place early November 1999. For Group 3, the computerised tests were administered on personal computers which were available in the school computer

room. A roster was set up in advance to accommodate the students for the testing. School results from the November examinations and the average term marks were obtained from the school.

6.6 VALIDITY OF THE LPCAT-2

Two groups were used to investigate the validity of the LPCAT-2 version of the LPCAT. For the sake of convenience, these two groups are named Group 4 and Group 5.

6.6.1 Group 4 for LPCAT-2 validity investigation

6.6.1.1 Sample for Group 4

Group 4 was an adult learner group (N=194) which consisted of a group of low-literacy adults who were mostly male (more than 95% of examinees in this group were male) and all African home language speakers. This group was involved in assessment for vocational training as part of a retrenchment package agreement. The mean age of this group was 29,7 years. Their level of education ranged from grade 1 to grade 12 with a mean of grade 8. Because of the composition of the sample, the results are reported for the total group only, since it can be regarded as reasonably homogeneous in terms of language and gender.

6.6.1.2 Measures obtained for Group 4

The LPCAT-2 was administered to the examinees, providing results in the form of a pretest score, a post-test score, a difference score and a composite learning potential score. The instructions per screen were read in English first, after which they were repeated by an instructor in the African language spoken by most examinees. The following other measures were obtained to assess the validity of the LPCAT-2:

- The Paper-and-Pencil Games (PPG), level 3 (Claassen, 1996) is a test which

measures figural, quantitative and verbal skills that are closely related to scholastic achievement and is suitable for the third and fourth school years. It is a group test that serves a screening function. Raw scores out of a total of 50 are provided for both the verbal and nonverbal sections respectively. The total score is the sum of the two scores. The Kuder Richardson formula 20 reliability for the level 3 form of the PPG test ranges between 0,78 and 0,95, while test-retest correlations range between 0,62 and 0,95. Correlations of the PPG with scholastic achievement scores range between 0,31 and 0,73.

- Level 1 literacy and numeracy scores were also obtained, each being scored out of a total of 50.

6.6.1.3 Procedures followed for obtaining validity information for Group 4

Testing of this group took place in October 1997. The LPCAT-2 was administered on a network system which had 40 workstations available, thereby allowing testing of up to 40 individuals in a single test session. Most of the examinees had a working knowledge of English. Oral instructions per screen were given in English first, after which they were repeated in the African language spoken by most of the examinees. An African assistant working for the organisation responsible for the overall testing programme gave the African language instructions and provided additional explanation when required. The paper-and-pencil tests - PPG, numeracy and literacy tests - were administered during the same time period.

6.6.2 Group 5 for LPCAT-2 validity investigation

6.6.2.1 Sample for Group 5

Group 5 consisted of 144 grade 8 pupils with a mean age of 13,2 years from an urban high school. With a few exceptions due to absenteeism on the days of testing, the entire group of grade 8 pupils of the specific school were included in the testing. The language distribution for this group was 41 percent African home language and 59

percent English/Afrikaans home language, while the gender distribution was 57 percent female and 43 percent male. The distribution in terms of home language and gender is provided in Table 6.4. Owing to some incomplete biographical data, the total in the table differs from the overall sample size given above.

TABLE 6.4 GROUP 5: HOME LANGUAGE BY GENDER CROSS-TABULATION

Home language/ gender	African home language	English/Afrikaans home language	Total
Male	23	32	55
Female	29	43	72
Total	52	75	127

6.6.2.2 Measures obtained for Group 5

The LPCAT-2 was administered to the examinees, providing a pretest score, a post-test score, a difference score and a composite learning potential score. The instructions were read in both English and Afrikaans per screen. Because the languages of instruction in the school are English and Afrikaans, it was felt that all pupils would be able to follow instructions for the LPCAT in these two languages. The following additional measures were obtained to assess the validity of the LPCAT-2:

- The GSAT-CAT (described for Group 2) was administered.
- The Learning Process Questionnaire (LPQ) (Biggs, 1987a, 1987b) was also administered. The LPQ is designed to assess the extent to which a secondary school student endorses different approaches to learning and the more important motives and strategies comprising those approaches. The LPQ is a 36-item, self-report questionnaire that yields scores on three basic motives for learning and three learning strategies (surface, deep and achieving), and on the approaches to learning that are formed by these motives and strategies. The three approaches lead to different kinds of learning outcome. The surface

approach leads to retention of factual detail at the expense of the structural relationships inherent in the data to be learned. The deep approach leads to an understanding of the structural complexity of the task and to positive feelings about it. The achieving approach, particularly in combination with the deep approach, leads to good performance in examinations, a good academic self-concept, and to feelings of satisfaction. Norms are provided separately for males and females at two age levels. For the present study, raw scores of the three approaches, comprising the sum of the respective motive and strategy scores, were used. Use of raw scores allows for more accurate assessment of correlations with other measures, which was the focus of the present study.

- English proficiency scores were obtained by means of the Proficiency Test English Second Language (Intermediate level) (Chamberlain & Reinecke, 1992). This test determines the examinee's knowledge and skill in language proficiency on the assumption that language proficiency levels are not attained solely as a result of curricular activities, but also as a result of extracurricular language contact and use. The test is meant to determine the proficiency level of English second language examinees in grade 7 to grade 9. The test consists of 40 multiple-choice questions with four options per item, and the results are provided in the form of T-scores with a mean of 50 and standard deviation 10. The reliability coefficient for this test is 0,89 and great care was taken to ensure its content validity (Chamberlain & Reinecke, 1992).
- Mathematics proficiency scores were obtained by means of the test of Basic Numerical Literacy (Venter, 1997). This test covers basic knowledge and comprehension of numbers as well as the application of basic numerical knowledge and concepts. The test consists of 35 multiple-choice questions for which the correct answer from four distractors must be indicated. The results are provided as a T-score with a mean of 50 and standard deviation of 10. The reliability of the test was determined by the Kuder Richardson formula 20 and is equal to 0,66 (Venter, 1997). No validity information is provided in the test manual, although the specification table reflecting content validity is provided.
- Two teacher rating scales, one for Mathematics and one for English were constructed and completed by the teachers involved in the teaching of those subjects to the examinees. These scales consisted of the same 13 questions

each, providing an assessment of the examinee's general performance, subject-specific performance and potential for improved performance in the specific subject. A copy of each of the rating scales is provided in Appendix C.

- For the school results, the average percentage marks for the four terms were obtained as well as the overall year percentage mark.

6.6.2.3 Procedures followed for obtaining validity information for Group 5

Testing of this group took place in January/February 1999. The grade 8 school guidance teacher was responsible for organising the schedule for getting the examinees to the testing venues at the agreed times. The paper-and-pencil tests (English and Mathematics proficiency tests and LPQ questionnaire) were administered to all the pupils in a single session on the first day of testing, while a roster was used to test smaller groups with the computer tests. Computer testing took place in the school's computer class, where 30 personal computers were available. The LPCAT-2 and GSAT-CAT computer tests were administered to groups of approximately 30 at a time. Because testing had to be scheduled at times when the pupils had free school sessions as well as when the computer classroom was available, the computer testing took place over three weeks. The teacher rating forms were completed at the end of November. School marks for the entire year were obtained early in December, following the final November end-of-year examinations.

6.7 ADDITIONAL VALIDITY INFORMATION

6.7.1 Group 6 for LPCAT further validity investigation

6.7.1.1 Sample for Group 6

Group 6 consisted of 109 grade 9 pupils from an urban high school. With a few exceptions due to absenteeism on the days of testing, the entire group of grade 9 pupils of the specific school were included in the testing. The language distribution for this group was 46 percent African home language and 54 percent English/Afrikaans home language, while the gender distribution was 44 percent female and 56 percent male. The distribution in terms of home language and gender is provided in Table 6.5.

TABLE 6.5 GROUP 6: HOME LANGUAGE BY GENDER CROSS-TABULATION

Home language/ gender	African home language	English/Afrikaans home language	Total
Male	30	31	61
Female	20	28	48
Total	50	59	109

6.7.1.2 Measures obtained for Group 6

The purpose for the inclusion of Group 6 was to investigate specifically the effect of various types of training on LPCAT-1 results. For this reason, the pupils from this group were randomly assigned to three different groups, after which the three specific procedures used were randomly allocated to the three groups. These three procedures involved the following:

- administration of the LPCAT-1 in its standard form (N=37) (this subgroup was the group that was reported on earlier regarding school validity information for the LPCAT-1,) namely Group 3).
- administration of the LPCAT-1 with provision of additional training - that is, working through 18 additional examples of the typical items used in the LPCAT - to further identify the key strategies available to solve these kinds of problems

(N=35)

- administration of the LPCAT-1 without any training between the pretest and post-test (N=37)

For all three the above groups, the LPCAT-1 (in various forms) was administered, providing a pretest score, a post-test score, a difference score and a composite learning potential score. Examinees chose to receive their instructions in either English or Afrikaans. For the group that did the additional training between the pretest and the post-test, instructions for the additional examples were provided in both English and Afrikaans. The group who did no training between the pretest and the post-test, could nevertheless choose to receive the initial explanation and examples prior to the pretest in either English or Afrikaans. Because the languages of instruction in the school are English and Afrikaans, it was felt that all pupils would be able to follow instructions for the LPCAT-1 in these two languages.

The school average term marks for the four terms as well as the overall average year mark were obtained for this group. Term 4 results and the average year mark are of most interest, since the LPCAT was administered at the end of the academic year - during the same time that the pupils were writing their end-of-year examinations.

6.7.1.3 Procedures followed for obtaining validity information for Group 6

Testing of this group took place in November 1999. The school guidance teacher was responsible for organising the schedule for getting the examinees to the testing venues at the agreed times. A list of the three groups to which the pupils had been randomly allocated was supplied to the teacher. For practical reasons, each of these three groups were further divided into two separate groups to accommodate the computer administration of the LPCAT using the different training procedures. The LPCAT was administered to groups of approximately 20 at a time. Because testing had to be scheduled at times when the pupils had free school sessions as well as when the computer classroom was available, the computer testing took place in six separate sessions over two weeks. School marks for the entire year were obtained early in

December, following the final November end-of-year examinations.

6.7.2 Combination groups for LPCAT further validity investigation

For additional validity information for the LPCAT, certain groups were evaluated together. Issues that are specifically investigated in this regard are the significance of difference scores for the different groups and the developmental changes indicated by the LPCAT.

6.7.2.1 Samples used for combined groups

All the samples (Groups 1 to 6) were used in these investigations.

6.7.2.2 Measures obtained for combined groups

For the evaluation of the significance of the difference scores, the mean LPCAT difference scores were used. For the evaluation of the effect of different training on the improvement in LPCAT post-test performance, the LPCAT was administered in three different forms, namely with standard training, additional training and no training respectively. The evaluation of the correlations of difference scores with other measures was done for Group 5 only, since this is the only group for which useful data in this regard are available. Regarding the developmental changes, all groups were used and their mean LPCAT scores compared.

6.7.2.3 Procedures followed for combined groups

No specific procedures are used - the data from the groups already described in the present chapter are used.

6.8 DATA CAPTURING AND STATISTICAL ANALYSIS

The data were at first captured in ASCII data files. These data files were then incorporated into the Statistical Packages for Social Sciences (SPSS) (Norusis/SPSS Inc, 1993) statistical analysis system for data analysis. Data analysis included descriptive statistics; comparison of means; comparison of frequency distributions; correlations for construct validity; correlations for criterion-related validity; and regression for the prediction and cross-cultural validity of test scores. These analyses will be discussed in the next chapter.

CHAPTER 7

EMPIRICAL VALIDITY RESULTS FOR THE LPCAT

7.1 INTRODUCTION

It is generally accepted that the validity of a test can be better evaluated if several types of validity evidence from different contexts can be supplied. In the LPCAT validity evaluations, an attempt was made to gather information from different samples and in various contexts, using different criteria to obtain information that would shed light on the meaning and utility of LPCAT scores. The samples used were from specific training institutions, and although the results described in this chapter provide empirical evidence of the criterion-related and construct validity of the LPCAT-1 and LPCAT-2, these samples are generally not very large and the results therefore have limited generalisability.

In the previous chapter, the content and face validity of the LPCAT were discussed. The samples, measures obtained and procedures followed to gather criterion-related validity information for the two versions of the LPCAT were described. Throughout this chapter on results, it should be borne in mind that the LPCAT makes use of the figural nonverbal reasoning domain and combines it with the dynamic testing strategy to measure learning potential. LPCAT results are reported for all four scores, namely the pretest, post-test, difference and composite scores. In terms of score interpretation, different options - that is, with a focus on specific scores in specific situations each time - were discussed in chapter 5 (see 5.7.6). In general, when reference is made to the measurement of learning potential with the LPCAT, all the measures obtained are implied (pretest, post-test, difference score and composite score) - each of which contributes to the measurement of learning potential in its own way as previously explained. For the sake of brevity, the LPCAT pretest, post-test and composite scores will be referred to as the LPCAT (PPC) scores throughout this chapter. In cases where a single score representing learning potential is required, preference is given to the composite score, since it represents a reasoned combination of the other scores and, within the learning potential framework, may

therefore be regarded as the most suitable *single* measure of learning potential.

Criterion-related validity, which is the main focus of this chapter, concerns evidence of the relationships between performance on the test and other independently obtained scores which also reflect the behaviour concerned. The evaluation of test validity is a continuous process and the accumulation of research results continues to add to the validity evidence of a test, even after publication. Thus the evidence provided in this chapter should be seen as initial information, which will in future be supplemented with further applications and research results for the LPCAT in different contexts.

The results for the different groups involved will be given, as far as possible, in a standard format in an attempt to simplify the presentation. Since one of the main aims of the LPCAT is to function as a screening instrument that is cross-culturally fair, specific attention will be given to cross-cultural utility throughout the presentation of the results by comparing cultural (ie home language) groups.

The different groups for which validity information is supplied in this chapter are described briefly below.

LPCAT-1 validity information

- Group 1: Technikon first-year students from Science and Engineering courses (N=92)
- Group 2: Another group of first-year Technikon students, also from Science and Engineering courses (N=223)
- Group 3: A group of grade 9 high school pupils from an urban high school (N=37)

LPCAT-2 validity information

- Group 4: A group of adult learners all retrenched from the same governmental organisation (N=194)
- Group 5: A group of grade 8 high school pupils from an urban high school (N=144)

Additional LPCAT validity information

- Group 6: A group of grade 9 high school pupils from an urban

high school (N=109)

- Groups 1-6: A combination of the above groups to investigate specific features of the LPCAT

Although these groups are not representative samples of South African culture groups, most of them were multicultural, and a comparison of performance of subgroups based on home language allows for the investigation of the cross-cultural functioning and utility of the LPCAT. For four of the groups, results on one of two other (standard) cognitive tests were also available. In the remainder of this introductory section, the framework for the format of the presentation of the results will be discussed.

7.1.1 Comparison of mean scores

The first information provided for each group is the descriptive statistics for the total group, and where possible, also for the language and gender subgroups. The mean scores of the subgroups are statistically compared by means of t-tests for independent samples. Comparison of the means of the subgroups based on language, namely the African home language group and the English/Afrikaans home language group, was considered to be practical, since it distinguishes between people who receive most of their education in their mother tongue, as opposed to those who do not. Furthermore, the African group is the most disadvantaged, hence it was considered necessary to assess the value of the LPCAT for different socioeconomic and language groups. Where practical, the gender groups are also compared.

7.1.2 Distribution of scores

While the descriptive statistics and comparison of mean scores provide useful information, the utility and value of the LPCAT can be further illustrated by comparing the frequency distributions of test scores. In addition to the comparison of mean group scores, the frequency distributions can give an indication of possible cross-cultural differences in performance patterns. Where possible, the frequency distribution of scores for the two language groups are compared for the LPCAT, a

standard cognitive test and (academic) criterion measures.

7.1.3 Correlations with other cognitive tests

One method of evaluating construct validity is to correlate a test with other tests that measure the same or a similar construct. According to Anastasi and Urbina (1997), the correlations between tests that measure approximately the same general area of behaviour should be moderately high, but not too high, since too much overlapping without added advantage, implies needless duplication. In the case of the LPCAT, the focus on dynamic measurement of learning potential by using only nonverbal, figural item content, together with the test-train-test computerised adaptive test administration, makes a unique contribution. Since the LPCAT learning potential measures are in the domain of general nonverbal reasoning ability, correlation with existing standard cognitive tests, in particular their nonverbal scores, provides useful information about its construct validity.

7.1.4 Correlations with criterion measures

Criterion-related validity refers to the effectiveness of a test to estimate performance on some other outcome measure of the same construct. The two types of criterion-related validity of interest are concurrent validity and predictive validity. A criterion should itself be reliable if it is to be a useful index of what the test measures. The validity coefficient will be diminished to the extent that the reliability of the test or the criterion is low, since the validity coefficient is always less than or equal to the square root of the test reliability multiplied by the criterion reliability (Gregory, 1996). There is no general answer to the question of how high a validity coefficient should be, and the less overlapping there is in content between the test and the criterion, the lower the expected validity coefficient will be. The latter is important to keep in mind when interpreting the LPCAT results, where performance on a test of nonverbal, figural content is compared to different measures of academic performance. Where possible, these correlations are presented for the total group as well as for the two language groups separately.

Two of the groups used were Technikon first-year students. Academic results for validity studies at tertiary level are notoriously problematic, because students come from extremely diverse academic backgrounds, take different combinations of subjects where different marking standards are employed, and furthermore, are not all equally proficient in the language of instruction (Huysamen, 1999). Although many problems exist with regard to the use of grade 12 results as well as tertiary academic results, especially for cross-cultural comparisons (Huysamen, 1999), these often are the only real-life criteria available.

7.1.5 Regression analysis and comparison of regression lines

If a test is to be used for the purpose of prediction, a regression equation can be obtained, which describes the best-fitting straight line for estimating the criterion from the test. The primary aim of the LPCAT is *not* to predict academic performance but to assess learning potential in the general reasoning ability domain. Regression analysis is nevertheless used for inspection and comparison of the regression lines for different subgroups - mainly to investigate possible differences between groups. The samples used for the present research are not very large and interpretation of and generalisation from regression results will therefore be cautious. For this reason, an in-depth statistical analysis of the regression results was not performed. Instead, a practical angle is taken whereby results are interpreted overall and described in terms of the possibility for over- or underprediction of criterion results, should the total group regression line be used.

The results for the different samples described in chapter 6 will be presented next according to the sequence and format discussed above.

7.2 EMPIRICAL VALIDITY RESULTS FOR THE LPCAT-1

Two Technikon first-year groups from two different Technikons and a grade 9 school group were used to investigate the empirical validity of the LPCAT-1. Because

Technikon students have at least a grade 12 qualification, they are a preselected group who have already shown a certain level of academic proficiency. This may represent some restriction of range in ability level. Restriction of range usually results in smaller correlation coefficients and should be kept in mind when interpreting the Technikon results provided for Group 1 and Group 2. No adjustments were made to correct for restriction of range. Another factor that could affect the validity results for these two groups is the reliability of the criterion measure. For the Technikon students used in the studies involving the LPCAT-1, academic results in the form of Mathematics I results and an average first-year score were used. Since examinees in the Technikon student samples do not all take the same combination of subjects and subjects may differ in difficulty or in marking standard, the resulting average score does not necessarily reflect a reliable comparative score for the examinees. The school group (Group 3), represents a group with a wider range of ability levels. Since they mostly take the same subjects, the academic criteria for this group in the form of average results of the four terms and an overall average year mark, are more reliable as criterion measures.

7.2.1 LPCAT-1 validity results for Group 1

As criterion measures, the results of the GSAT (Claassen et al, 1991) as well as academic results (school grade 12 and Technikon first year) were used to shed light on the construct and criterion-related validity of the LPCAT. The multicultural composition of the sample provided the opportunity to investigate the cross-cultural functioning of the LPCAT for these students.

7.2.1.1 Group 1: Comparison of mean scores

The descriptive statistics for Group 1 are provided in Table 7.1, together with the results of the comparison of the mean scores of the two language groups.

There are statistically highly significant differences between the mean scores of the two language groups on all but three of the measures. The three measures for which the differences between the two groups are not significant are the LPCAT difference score, Mathematics I and grade 12 English. The fact that there is no significant difference between the two language groups on the LPCAT difference score indicates that, despite general differences in their performance on cognitive ability measures (GSAT and LPCAT), it could not be shown that the possibility for improvement - ZPD or difference score - is different for the two groups.

TABLE 7.1 GROUP 1: DESCRIPTIVE STATISTICS AND COMPARISON OF LANGUAGE GROUP MEAN SCORES

	Total group			African language group			Eng/Afr language group			Comparison of means	
	N	Mean	SD	N	Mean	SD	N	Mean	SD	Mean diff #	p-value
LPCAT pretest	92	57,78	6,27	46	54,43	6,39	46	61,13	3,98	6,70	,000**
LPCAT post-test	92	59,15	5,22	46	55,74	3,92	46	62,57	3,99	6,83	,000**
LPCAT composite score	92	58,28	5,99	46	54,87	5,75	46	61,70	3,97	6,83	,000**
LPCAT difference score	92	1,37	3,27	46	1,30	3,65	46	1,43	2,87	0,13	,849
GSAT verbal	76	105,14	15,09	35	92,49	8,53	41	115,95	10,26	23,47	,000**
GSAT nonverbal	76	109,63	15,59	35	98,43	9,40	41	119,20	13,31	20,77	,000**
GSAT total	76	108,28	15,57	35	95,31	8,54	41	119,34	10,96	24,03	,000**
Mathematics I	77	55,60	12,11	35	53,79	9,22	42	57,11	14,00	3,32	,217
First-year average	89	53,11	6,92	45	50,63	6,07	44	55,64	6,88	5,01	,000**
Grade 12 English	92	225,54	37,16	46	219,89	41,53	46	231,20	31,66	11,30	,146
Grade 12 Mathematics	90	170,89	50,00	45	147,00	43,97	45	194,78	44,22	47,78	,000**
Grade 12 Science	90	176,89	44,61	44	152,61	34,98	46	200,11	40,49	47,50	,000**

absolute value of the difference between the mean scores

** p < ,01 * p < ,05 (p-values for a nondirectional t-test for independent samples between the two language groups)

The LPCAT scores are of the expected magnitude, considering the educational level of the examinees. The mean LPCAT (PPC) scores for both language groups are above the average of 50, the latter being commensurate with a grade 10 level of education. The means of the African language group are significantly lower than those of the English/Afrikaans group. The mean GSAT scores of the English/Afrikaans group are all above the average of 100, while those of the African language group are all below the average.

To further interpret the mean score differences between the two language groups, the cognitive test (LPCAT and GSAT) differences are also evaluated as a proportion of different standard deviation scores. The difference between the two language groups considered as a proportion respectively of the theoretical standard deviation, the standard deviation of the total group and the unweighted mean (common) standard deviation, are reported in Table 7.2. with most emphasis in interpretation on the values obtained by use of the common standard deviation. The size of these differences between the two language groups on the LPCAT and GSAT are thus standardised by expressing them in terms of the unweighted mean standard deviation of the two language groups (Hugo & Claassen, 1991; Jensen, 1980). In this way it becomes possible to compare the size of the respective mean differences between the two language groups on the different measures. Smaller values indicate that the scores for the two groups are more similar and that the measures can thus be viewed as more cross-culturally fair.

For the LPCAT, the differences in terms of the common standard deviation varied between 1,291 and 1,725, while for the GSAT it varied between 1,828 and 2,497. The sizes of the proportional differences are smaller for the LPCAT scores and the GSAT nonverbal scores than for the GSAT verbal and total scores. This emphasises the importance of language proficiency on verbal test performance in cross-cultural testing. This could also be interpreted to mean that, since the two language groups show larger differences on verbal performance, a focus on verbal performance could lead to an underestimation of the general reasoning ability of African language examinees.

TABLE 7.2 GROUP 1: MEAN DIFFERENCES BETWEEN LANGUAGE GROUPS AS A PROPORTION OF DIFFERENT STANDARD DEVIATION UNITS

Variable	Mean difference	Proportion of theoretical SD*	Proportion of total group SD*	Proportion of common SD*
LPCAT pretest	6,70	0,670 [10]	1,069 [6,27]	1,291 [5,19]
LPCAT post-test	6,83	0,683 [10]	1,308 [5,22]	1,725 [3,96]
LPCAT composite score	6,83	0,683 [10]	1,140 [5,99]	1,405 [4,86]
GSAT verbal score	23,47	1,565 [15]	1,553 [15,09]	2,497 [9,40]
GSAT non-verbal score	20,77	1,385 [15]	1,332 [15,59]	1,828 [11,36]
GSAT total score	24,03	1,602 [15]	1,543 [15,57]	2,465 [9,75]

* SD value provided in square brackets below the proportional value each time.

For comparative purposes, the mean scores of the gender groups are also compared, although these results cannot be generalised, owing to the small number of females (N=10) in this sample. The descriptive statistics for the gender groups on different measures as well as the results of comparing the mean scores by means of a t-test for independent samples, are nevertheless reported in Table 7.3 for the sake of completeness.

The gender groups differ significantly on all the LPCAT scores except the LPCAT difference score. Significant differences are also found for the GSAT verbal and total scores. While the Mathematics first-year results do not differ significantly, the mean first-year average differs significantly - with the male group obtaining the higher mean score. The differences between the two gender groups are not significant in the

grade 12 school results. Although these results indicate some gender differences, the female subgroup of Group 1 was too small to allow generalisation from these results.

TABLE 7.3 GROUP 1: DESCRIPTIVE STATISTICS AND COMPARISON OF GENDER GROUP MEAN SCORES

Variable	Females			Males			Mean diff #	p-value
	N	Mean	SD	N	Mean	SD		
LPCAT pretest	10	53,90	8,75	82	58,26	5,80	4,36	,037*
LPCAT post-test	10	55,60	4,43	82	59,59	5,16	3,99	,022*
LPCAT composite score	10	54,58	7,61	82	58,74	5,66	4,15	,038*
LPCAT difference score	10	1,70	5,08	82	1,33	3,02	0,37	,737
GSAT verbal	9	95,11	10,49	67	106,49	15,16	11,38	,033*
GSAT nonverbal	9	100,33	12,45	67	110,88	15,62	10,55	,056
GSAT total	9	97,67	11,92	67	109,70	15,52	12,03	,028*
Mathematics I	8	49,44	8,49	69	56,31	12,30	6,87	,129
First-year average	10	47,30	4,75	79	53,84	6,82	6,54	,004**
Grade 12 English	10	236,00	33,73	82	224,27	37,55	11,73	,349
Grade 12 Mathematics	10	147,00	43,54	80	173,88	50,19	26,88	,109
Grade 12 Science	10	151,00	39,92	80	180,13	44,33	29,13	,051

absolute value of the difference between the mean scores

** p < ,01 * p < ,05

(p-values for a nondirectional t-test for independent samples between the two gender groups)

7.2.1.2 Group 1: Distribution of scores

Figures 7.1 to 7.6 indicate the distributions of scores on the GSAT, LPCAT and academic results for the two language groups respectively.

FIGURE 7.1 GROUP 1: DISTRIBUTION OF GSAT VERBAL SCORES

FIGURE 7.2 GROUP 1: DISTRIBUTION OF GSAT NONVERBAL SCORES

FIGURE 7.3 GROUP 1: DISTRIBUTION OF AVERAGE FIRST-YEAR

The distributions of GSAT scores (Figures 7.1 and 7.2) indicate noticeable differences between the two language groups, with the African home language group showing a positively skewed distribution. The distributions of average academic scores of the two language groups are more similar. For the LPCAT, the distributions of scores are similar to those of the academic scores, with a smaller difference between the language groups than for the GSAT standard cognitive test and a less positively skewed distribution of scores for the African language group.

FIGURE 7.4 GROUP 1: DISTRIBUTION OF LPCAT PRETEST SCORES

FIGURE 7.5 GROUP 1: DISTRIBUTION OF LPCAT POST-TEST SCORES

FIGURE 7.6 GROUP 1: DISTRIBUTION OF LPCAT COMPOSITE SCORES

The LPCAT demonstrates a reasonable range of scores for both language groups, indicating that it can also distinguish between members within each group. As a measure of learning potential in the domain of general nonverbal reasoning ability, the LPCAT seems to provide a somewhat more equitable distribution of scores for the African language subgroup of Group 1 than do the GSAT standard cognitive measures.

7.2.1.3 Group 1: LPCAT correlations with the GSAT

In the case of Group 1, the GSAT (Senior) paper-and-pencil test (Claassen et al, 1991) was used as a criterion measure to investigate the functioning of the LPCAT. The correlations between the pretest, post-test, composite and difference scores of the LPCAT and the verbal, nonverbal and total scores of the GSAT respectively are reported in Table 7.4.

TABLE 7.4 GROUP 1: CORRELATIONS OF LPCAT WITH GSAT(SENIOR) PAPER-AND-PENCIL TEST (N=76)

		GSAT verbal	GSAT nonverbal	GSAT total
LPCAT pretest	r	,498	,533	,550
	p	,000**	,000**	,000**

LPCAT post-test	r	,653	,693	,713
	p	,000**	,000**	,000**
LPCAT composite score	r	,542	,574	,594
	p	,000**	,000**	,000**
LPCAT difference score	r	,064	,060	,059
	p	,584	,608	,615

** p < ,01

The correlations between the two tests show that they measure a similar construct, with the LPCAT post-test score correlating the highest with the GSAT scores. The LPCAT difference score shows negligible correlation with all the GSAT scores, as expected from the way that learning potential has been defined for the LPCAT. These results provide support for the construct validity of the LPCAT as a test that measures learning potential in the general reasoning ability domain. In particular, the correlations of the LPCAT post-test and LPCAT composite scores respectively with the GSAT scores are higher than that of the LPCAT pretest score with the GSAT scores and thus provide support for the dynamic measurement of learning potential.

7.2.1.4 Group 1: LPCAT correlations with criterion measures

For predictive validity using real-life criterion measures, the end-of-year results in Mathematics I as well as the first-year average were used. Correlations of LPCAT scores with grade 12 results in English, Mathematics and Science are also reported, although the latter results were obtained approximately four months before the LPCAT was administered. These school results were considered to be important, because school academic results are often used for selection purposes. The LPCAT correlations with first-year academic results and with grade 12 academic results for the total group are provided in Table 7.5

The LPCAT results generally show low correlation with first-year academic results and although the LPCAT post-test score shows a statistically significant correlation with the first-year academic average score, no clear pattern of correlations emerges.

In general, the correlations of the LPCAT with first-year academic performance are probably too low to be practically useful. For this group, correlations of the LPCAT post-test with the academic results are higher in each case than those of the LPCAT pretest with the academic results. Although not to the same degree, the LPCAT composite score correlations with academic results are also higher than those of the LPCAT pretest with the different academic results. This provides support for the use of learning potential measures that include measures of present (pretest ability) as well as the results following training (difference score). The LPCAT difference score shows higher correlations with the Mathematics first-year results than do the other LPCAT scores, and the LPCAT difference score correlation with first-year average is the second highest of the LPCAT correlations with the first-year average. The reason for these high correlations of the LPCAT difference score with the academic criteria could be because the group can be regarded as preselected (all having a minimum of grade 12 education), and therefore of comparable ability level. This aspect will be investigated next by means of a scatter diagram. The LPCAT correlations with grade 12 results in English, Mathematics and Science are noticeably higher and, except for the LPCAT difference score correlations with them, are all statistically significant. Some construct of general reasoning measured by the LPCAT nonverbal item content seems to overlap with a domain of reasoning required for and measured by these grade 12 school subjects.

TABLE 7.5 GROUP 1: CORRELATIONS OF LPCAT WITH FIRST-YEAR ACADEMIC AND GRADE 12 RESULTS

		First-year Mathematics	First-year average	Grade 12 English	Grade 12 Maths	Grade 12 Science
LPCAT pretest	r	,004	,098	,207	,365	,380
	p	,971	,363	,047*	,000**	,000**
	N	77	89	92	90	90
LPCAT post-test	r	,138	,230	,263	,419	,450
	p	,230	,030*	,047*	,000**	,000**
	N	77	89	92	90	90
LPCAT composite score	r	,018	,117	,222	,373	,394
	p	,877	,277	,033**	,000**	,000**
	N	77	89	92	90	90
LPCAT difference score	r	,209	,186	,022	-,038	-,007
	p	,068	,081	,833	,722	,947
	N	77	89	92	90	90

** p < ,01 * p < ,05

To investigate whether the low magnitude of correlations of the LPCAT (PPC) scores compared to the relatively higher correlations of the LPCAT difference score respectively with average first-year academic performance could be caused by restriction of range, a scatter diagram of the LPCAT composite scores and academic first-year average scores is presented in Figure 7.7.

FIGURE 7.7 GROUP 1: SCATTER DIAGRAM OF LPCAT COMPOSITE SCORES AND FIRST-YEAR AVERAGE ACADEMIC RESULTS PER LANGUAGE GROUP

The scatter diagram indicates a noticeable restriction of range for Group 1 for both the LPCAT composite score as well as for the first-year academic results. This restriction of range is even more distinct within each language group and is likely to lead to lower correlation between the scores concerned.

In Table 7.6, the correlations of LPCAT, GSAT and school academic results with first-year Mathematics and first-year average results, are provided separately for the two language groups.

TABLE 7.6 GROUP 1: CORRELATIONS OF LPCAT, GSAT AND GRADE 12 RESULTS WITH FIRST-YEAR ACADEMIC RESULTS FOR THE TWO LANGUAGE GROUPS

Predictor variable		African home language group		English/Afrikaans home language group	
		Maths I	First-year average	Maths I	First-year average
LPCAT pretest score	r	-,142	-,037	-,038	-,264
	p	,416	,811	,812	,083
	N	35	45	42	44
LPCAT post-test score	r	,153	,079	,019	-,115
	p	,380	,608	,904	,457
	N	35	45	42	44
LPCAT composite score	r	-,109	-,028	-,049	-,255
	p	,531	,853	,760	,095
	N	35	45	42	44
LPCAT difference score	r	,395	,151	,077	,208
	p	,019*	,323	,628	,176
	N	35	45	42	44
GSAT verbal score	r	,281	,326	,206	,052
	p	,155	,060	,208	,750
	N	27	34	39	40
GSAT nonverbal score	r	,304	,287	,285	,181
	p	,123	,100	,078	,263
	N	27	34	39	40
GSAT total score	r	,269	,305	,276	,083
	p	,175	,080	,089	,609
	N	27	34	39	40
Grade 12 English	r	,179	,390	,086	,210
	p	,303	,008**	,589	,171
	N	35	45	42	44
Grade 12 Mathematics	r	,459	,456	,585	,407
	p	,006**	,002**	,000**	,006**
	N	35	44	42	44
Grade 12 Science	r	,445	,406	,563	,569
	p	,008**	,007**	,000**	,000**
	N	34	43	42	44

** p < ,01 * p < ,05

For both language groups, neither the LPCAT (PPC) scores nor the standard cognitive

test scores correlate statistically significantly with academic results. The correlations of the GSAT with the first-year academic results are generally larger than those of the LPCAT (PPC) scores with first-year results. Considering the content of the standard cognitive tests, (ie number series and verbal subtests respectively), a higher correlation with Mathematics I and general academic performance - which is generally known to be related to language proficiency, could be expected. Correlations of the GSAT scores with first-year average - although not statistically significant - are generally higher for the African language group than for the English/Afrikaans group.

In Group 1, with the restricted ability range indicated by the scatter diagram of Figure 7.7, the LPCAT difference scores correlate significantly with first-year Mathematics for the African-language group, while its correlations with the first-year average for that group are also larger than those of the LPCAT (PPC) scores. For the English/Afrikaans language group, LPCAT correlations with Mathematics I are all close to zero. For this group, the LPCAT (PPC) score correlations with first-year average are negative, while the LPCAT difference score shows a positive correlation with the first-year average. Regarding the LPCAT and GSAT correlations with first-year academic results for Group 1, no clear pattern emerges and the overall results are insignificant.

The remaining results in Table 7.6 indicate that for both language groups, higher correlations are obtained between grade 12 results and first-year academic results, for grade 12 Mathematics and Science in particular. Grade 12 English correlates highly significantly with first-year average performance for the African language group, while the same correlation for the English/Afrikaans group is not significant. This provides support for the contention that language proficiency should be regarded as an important predictor of academic results for the African home language group (Huysamen, 1999).

Prior academic performance and performance on standard cognitive tests generally seem to correlate more highly with first-year academic performance. However, considering the specific problem in South Africa regarding inequalities in schooling, the use of only school results or measures that rely on them for selection purposes, would not constitute fair selection for multicultural groups. Use of only these

measures for selection and placement purposes, is likely to perpetuate existing inequalities in our society, because they are linked to prior educational and socioeconomic opportunities.
For all the measures concerned, restriction of range is likely to have affected the magnitude of correlations found.

7.2.1.5 Group 1: Regression analysis and comparison of regression lines

Since the correlations between the LPCAT composite score and the average first-year mark are not significant for either language group, no meaningful comparison of the regression lines between these variables for the two groups is possible for Group 1.

7.2.1.6 Group 1: Overview and summary

For this particular group, there are significant mean differences between the two language groups on most measures. Comparison of the standardised mean differences between the two language groups by expressing the differences in terms of the common standard deviation, indicates that differences between the two language groups are smaller on the LPCAT and on the nonverbal standard cognitive measure than on the verbal and total standard cognitive measures. Differences between the gender groups can only be taken as a rough indication, owing to the small number of females included in this sample. The frequency distributions of scores indicate that the distributions of LPCAT (PPC) scores of the two language groups are more similar than the distributions of the GSAT scores, although language group differences on the LPCAT distribution of scores were also evident.

Correlations of the LPCAT with the GSAT indicate that both tests measure the same construct. In particular, the measures that include pretest (present) performance as well as some component of learning, namely the LPCAT post-test and LPCAT composite scores respectively, show higher correlations with all the GSAT scores - thereby providing support for dynamic assessment of learning potential. LPCAT correlations with Technikon first-year academic results were generally low, although

some were statistically significant. Restriction of range seems to have reduced the correlations found. Grade 12 results gave the highest correlations with first-year academic results for both language groups.

For multicultural groups, the LPCAT, seems to be a reasonably equitable measure of learning potential within the nonverbal reasoning ability domain. The LPCAT measures are not as reliant on prior formal learning experiences as most standard tests, while allowing for improvement in test performance following relevant training. It can provide useful *additional* information, despite the fact that it may not correlate as highly with first-year academic results as either previous academic results or standard cognitive tests. For academic selection purposes, the best practice would probably be to use learning potential scores together with previous academic results and standard cognitive test results.

7.2.2 LPCAT-1 validity results for Group 2

A second sample of first-year students from another Technikon - also from the faculties of Science and Engineering - was used to obtain further empirical validity information for the LPCAT-1. For this group, the results of the GSAT-CAT (Van Tonder & Claassen, 1992) as well as academic results (grade 12 and first-year Technikon) were used to provide further information on the construct and criterion-related validity of the LPCAT. Three subtests of the Senior Aptitude Test (SAT) (Owen & Taljaard, 1989) were also administered to provide additional information. The multicultural composition of the sample, consisting largely of African (African home language) and Indian (English home language) students, also provided the opportunity to investigate the cross-cultural functioning of the LPCAT for these students.

7.2.2.1 Group 2: Comparison of mean scores

The descriptive statistics for the Group 2 Technikon first-year sample for the LPCAT-1 scores as well as for the other measures obtained are provided in Table 7.7 together with the results for comparing the mean scores of the two language groups by means of t-tests for independent samples. Owing to missing values for certain variables (ie language group), the subsample sizes and total sample sizes given do not always add up.

For group 2, there are significant differences between the two language groups on many of the measures. Regarding the psychometric tests, the two language groups differ significantly on all of the GSAT and SAT scores. For the LPCAT, there is a significant difference between the two language groups on the post-test score only. Apart from the LPCAT pretest, composite score and difference score where there is no significant difference, the only other two scores for which there is no significant difference between the two language groups, are Mathematics I and grade 12 Mathematics. First-year average scores show a highly significant difference between the two language groups, as do grade 12 English and grade 12 Science.

The differences between the two language groups on the GSAT-CAT and LPCAT are also evaluated as a proportion of the theoretical, total group and common standard deviation scores respectively. These results are provided in Table 7.8.

TABLE 7.7 GROUP 2: DESCRIPTIVE STATISTICS AND COMPARISON OF LANGUAGE GROUP MEAN SCORES

	Total group			African language group			Eng/Afr language group			Comparison of means	
	N	Mean	SD	N	Mean	SD	N	Mean	SD	Mean diff #	p-value
LPCAT pretest	159	55,21	6,13	69	54,14	6,04	86	56,01	6,21	1,87	,062
LPCAT post-test	159	56,47	5,39	69	55,42	4,76	86	57,30	5,81	1,88	,032*
LPCAT composite score	159	55,59	5,99	69	54,53	5,83	86	56,40	6,12	1,87	,055
LPCAT difference score	159	1,26	3,26	69	1,28	3,26	86	1,29	3,26	0,015	,977
GSAT-CAT verbal	159	100,11	13,39	68	91,19	9,91	87	107,31	11,57	16,12	,000**
GSAT-CAT nonverbal	159	107,25	12,70	68	103,34	12,29	87	110,18	12,23	6,85	,001**
GSAT-CAT total	159	103,92	11,90	68	97,53	9,47	87	108,95	11,29	11,42	.000**
Mathematics I	62	48,87	16,21	31	48,90	15,73	22	53,09	18,07	4,19	,373
First-year average	165	56,74	11,11	64	53,43	9,46	81	60,48	11,67	7,04	,000**
SAT-calculations	198	21,27	3,29	85	19,85	3,40	93	22,70	2,40	2,85	,000**
SAT-3-dimensional	206	18,77	4,20	87	17,26	4,49	95	20,16	3,35	2,89	,000**
SAT-mechanical	208	18,47	3,44	87	16,92	3,68	96	19,81	2,64	2,89	,000**
Grade 12 English	191	231,52	40,55	78	211,35	36,64	90	249,28	36,62	37,93	,000**
Grade 12 Mathematics	185	195,08	50,62	73	193,63	51,37	89	195,39	49,56	1,763	,825

Grade 12 Science	183	194,78	46,23	73	171,85	36,80	86	211,34	45,05	39,49	,000**

absolute value of the difference between the mean scores

** p < ,01 * p < ,05 (p-values for a nondirectional t-test for independent samples between the two language groups)

TABLE 7.8 GROUP 2: MEAN DIFFERENCES BETWEEN LANGUAGE GROUPS AS A PROPORTION OF DIFFERENT STANDARD DEVIATION UNITS

Variable	Mean difference	Proportion of theoretical SD*	Proportion of total group SD*	Proportion of common SD*
LPCAT pretest	1,87	0,187 [10]	0,305 [6,13]	0,305 [6,13]
LPCAT post-test	1,88	0,188 [10]	0,349 [5,39]	0,355 [5,29]
LPCAT composite score	1,87	0,187 [10]	0,312 [5,99]	0,313 [5,98]
GSAT-CAT verbal score	16,12	1,075 [15]	1,204 [13,39]	1,501 [10,74]
GSAT-CAT non-verbal score	6,85	0,457 [15]	0,539 [12,70]	0,559 [12,26]
GSAT-CAT total score	11,42	0,761 [15]	0,960 [11,90]	1,101 [10,38]

* Standard deviation in square brackets below the proportional value each time

For this group, the differences between the mean scores as a proportion of different standard deviation scores are again in all cases smaller for the LPCAT than for the GSAT. This indicates that, for cross-cultural use, the LPCAT seems to provide more equitable measures.

The mean scores of the two gender subgroups were also compared. The descriptive statistics for the two gender groups are provided in Table 7.9 together with the results of the comparison of the mean scores by means of a t-test for independent samples.

TABLE 7.9 GROUP 2: DESCRIPTIVE STATISTICS AND COMPARISON OF GENDER GROUP MEAN SCORES

Variable	Females			Males			Mean	p-value
	N	Mean	SD	N	Mean	SD	diffe-re nce #	
LPCAT pretest	69	54,64	5,18	88	55,69	6,83	1,06	,288
LPCAT post-test	69	55,68	4,52	88	57,13	5,97	1,44	,097
LPCAT composite score	69	54,98	5,09	88	56,12	6,64	1,14	,241
LPCAT difference score	69	1,04	3,47	88	1,43	3,13	0,39	,463
GSAT-CAT verbal	68	100,10	10,98	89	100,36	15,07	0,26	,902
GSAT-CAT nonverbal	68	105,51	11,33	89	108,65	13,48	3,14	,124
GSAT-CAT total	68	103,00	10,02	89	104,78	13,15	1,78	,338
Mathematics I	5	45,40	13,04	55	49,22	16,81	3,82	,624
First-year average	71	59,84	9,83	91	54,54	11,60	5,30	,002**
SAT calculations	88	21,86	2,84	107	20,90	3,51	0,97	,035*
SAT 3-dimensional	92	18,30	3,91	111	19,23	4,37	0,93	,115
SAT mechanical	92	17,70	3,29	113	19,13	3,43	1,44	,003**
Grade 12 English	83	248,92	37,64	105	218,86	37,41	30,06	,000**
Grade 12 Mathematics	81	197,53	49,71	101	194,01	51,88	3,52	,644
Grade 12 Science	78	204,17	42,70	102	188,09	47,99	16,08	,021*

absolute value of the difference between the mean scores

** p < ,01 * p < ,05

(p-values for a nondirectional t-test for independent groups between the two gender groups)

Overall, there are far fewer significant differences between the mean scores of the gender groups than there were for the two language groups. No significant differences are found for the LPCAT or the GSAT scores. For the academic results,

significant differences between the gender groups are found for grade 12 English and grade 12 Science as well as for the first-year average performance, with the female mean score larger than the male mean score for all these measures. Significant differences were also found on two of the three SAT subtests, namely Calculations and Mechanical Reasoning, with the female group obtaining a higher mean in the Calculations subtest, and the males obtaining a higher mean in the Mechanical Reasoning subtest.

7.2.2.2 Group 2: Distribution of scores

In Figures 7.8 to 7.13 the frequency distribution of the GSAT-CAT, academic and LPCAT scores for the two language groups are provided.

FIGURE 7.8 GROUP 2: DISTRIBUTION OF GSAT VERBAL SCORES

For Group 2, the distribution of GSAT-CAT verbal scores for the African home language group shows a positively skewed distribution. For the nonverbal scores, the distributions of the two language groups are more similar and both negatively skewed.

FIGURE 7.9 **GROUP 2: DISTRIBUTION OF GSAT NONVERBAL SCORES**

FIGURE 7.10 **GROUP 2: DISTRIBUTION OF AVERAGE FIRST-YEAR ACADEMIC SCORE**

FIGURE 7.11 **GROUP 2: DISTRIBUTION OF LPCAT PRETEST SCORES**

FIGURE 7.12 GROUP 2: DISTRIBUTION OF LPCAT POST-TEST SCORES

FIGURE 7.13 GROUP 2: DISTRIBUTION OF LPCAT COMPOSITE SCORES

The academic score distributions of the two language groups are also reasonably similar, possibly because the examinees represent a preselected group with proven (minimum grade 12) academic attainment. The LPCAT score distributions are reasonably similar for the two language groups and therefore seem to provide equitable learning potential measures for the two language groups in the domain of general reasoning ability. For Group 2, except for the GSAT-CAT verbal scores, the score distributions for the two language groups are generally reasonably similar.

7.2.2.3 Group 2: LPCAT correlations with the GSAT-CAT

In the case of Group 2, the results of the GSAT-CAT (Van Tonder & Claassen, 1992) were used as one of the criterion measures. The correlations of the LPCAT scores with the verbal, nonverbal and total scores of the GSAT-CAT are reported in Table 7.10.

TABLE 7.10 GROUP 2: CORRELATIONS OF LPCAT WITH GSAT-CAT
(N = 158)

	GSAT verbal	GSAT nonverbal	GSAT total

LPCAT pretest	r	,563	,627	,653
	p	,000**	,000**	,000**
LPCAT post-test	r	,571	,645	,668
	p	,000**	,000**	,000**
LPCAT composite score	r	,569	,638	,663
	p	,000**	,000**	,000**
LPCAT difference score	r	-,113	-,110	-,123
	p	,159	,167	,125

** p < ,01

The magnitude and significance of correlations between the LPCAT (PPC) scores and the GSAT-CAT scores for Group 2 indicate that the two scales measure a common construct. These results provide support for the construct validity of the LPCAT as a learning potential test in the domain of general reasoning ability. The GSAT-CAT scores can also be regarded as criterion measures for evaluating the concurrent validity of the LPCAT, since the two tests were administered during the same time period. The low negative correlations of the LPCAT difference scores are in line with the LPCAT definition of learning potential, with the difference score on its own not expected to provide meaningful correlations with external criteria, *unless* the group that is used is homogeneous with regard to ability level.

7.2.2.4 Group 2: LPCAT correlations with criterion measures

For criterion-related (predictive) validity, the end-of-year results in Mathematics I and first-year average results were obtained as criterion measures. Correlations of LPCAT scores with grade 12 results in English, Mathematics and Science are also reported. The sample included students taking general science courses, such as Food Technology and Environmental Health for which Mathematics is not a compulsory subject, with the result that relatively few students in the sample had Mathematics I as a subject. In Table 7.11, the correlations of the LPCAT with Mathematics I and first-year average scores are reported for the total group. The

limitations regarding these tertiary-level academic measures as discussed initially (see 7.1.4) and for Group 1 (see 7.2.1.4) also apply.

TABLE 7.11 GROUP 2: CORRELATIONS OF LPCAT WITH FIRST-YEAR ACADEMIC AND GRADE 12 RESULTS

		First-year Mathematics	First-year average	Grade 12 English	Grade 12 Maths	Grade 12 Science
LPCAT pretest	r	,165	,213	,218	,210	,287
	p	,257	,020*	,011*	,016*	,001**
	N	49	120	137	132	128
LPCAT post-test	r	,191	,158	,183	,124	,251
	p	,188	,084	,032*	,158	,004**
	N	49	120	137	132	128
LPCAT composite score	r	,169	,209	,221	,204	,293
	p	,247	,022*	,009**	,019*	,001**
	N	49	120	137	132	128
LPCAT difference score	r	,015	-,139	-,120	-,194	-,138
	p	,921	,131	,161	,026*	,119
	N	49	120	137	132	128

** p < ,01 * p < ,05

For Group 2, the correlations of the LPCAT (PPC) scores with Mathematics I and average first-year results are relatively low, although some are statistically significant. Whereas for Group 1 the LPCAT post-test was the only LPCAT measure that correlated significantly with first-year average marks, for the present group the LPCAT pretest and composite scores correlate significantly with the first-year average. As for Group 1, the LPCAT (PPC) scores of Group 2 generally correlate significantly with

grade 12 results, the only correlation that is not significant being that of the LPCAT post-test with grade 12 Mathematics. The low correlations of the LPCAT difference score with the academic scores are not significant, except for grade 12 mathematics, which indicates that larger difference scores on the LPCAT are associated with slightly lower performance in grade 12 mathematics. To investigate whether restriction of range may have affected the correlations found, a scatter diagram of the LPCAT composite scores and academic first-year average scores is presented in Figure 7.14.

FIGURE 7.14 GROUP 2: SCATTER DIAGRAM OF LPCAT COMPOSITE SCORES AND FIRST-YEAR AVERAGE ACADEMIC SCORES PER LANGUAGE GROUP

Figure 7.14 indicates that there is less restriction in range for both LPCAT and first-year average academic results for this group than was the case for Group 1, suggesting that the correlations were probably not affected to the same extent that they were for Group 1. This may explain the generally slightly higher correlations of the LPCAT (PPC) scores with first-year results. These correlations are, however, still of such a small magnitude that they are unlikely to be practically useful. In Table 7.12, the correlations of various cognitive and academic scores with first-year results are reported separately for the two language groups.

For the African language subgroup, the only scores that provide statistically significant

correlations with either Mathematics 1 or first-year average performance, are the grade 12 academic results. For the African language group, grade 12 English correlates highly significantly with first-year average, while both grade 12 Mathematics and grade 12 Science correlate highly significantly with Mathematics I. For the English/Afrikaans language subgroup, the correlations of the LPCAT pretest and LPCAT composite score with first-year average are significant while the GSAT-CAT nonverbal score, which contains number series items, correlates significantly with Mathematics I. For this group, grade 12 Mathematics correlates significantly with Mathematics I and highly significantly with the first-year average. Grade 12 Science also correlates significantly with first-year average for the English/Afrikaans group. It is important to note that, while grade 12 English correlates the highest with first-year average for the African language group, the same correlation for the English/Afrikaans language group is close to zero. This again emphasises the importance of language proficiency when the language of instruction is not the learner's first language.

TABLE 7.12 GROUP 2: CORRELATIONS OF PSYCHOMETRIC AND ACADEMIC MEASURES WITH MATHEMATICS 1 AND FIRST-YEAR AVERAGE FOR THE TWO LANGUAGE GROUPS

		African language		English / Afrikaans	
		Mathematics I	First-year average	Mathematics I	First-year average
LPCAT pretest	r	-,047	,043	,435	,291
	p	,820	,775	,055	,015*
	N	26	47	20	70
LPCAT post-test	r	-,058	-,016	,443	,183
	p	,778	,916	,051	,129
	N	26	47	20	70
LPCAT composite score	r	-,042	,049	,436	,277
	p	,840	,743	,054	,020*
	N	26	47	20	70
LPCAT difference score	r	-,002	-,110	-,072	-,232
	p	,993	,464	,762	,054
	N	26	47	20	70
GSAT-CAT verbal	r	,017	,280	,310	,171
	p	,934	,059	,183	,158
	N	26	46	20	70
GSAT-CAT nonverbal	r	,169	,001	,465	,145
	p	,408	,995	,039*	,230
	N	26	46	20	70
GSAT-CAT total	r	,122	,147	,428	,168
	p	,553	,329	,060	,163
	N	26	46	20	70
SAT calculations	r	,138	,065	,440	,214
	p	,510	,633	,052	,067
	N	25	56	20	74
SAT 3-dimensional	r	-,016	,129	,356	,132
	p	,940	,336	,134	,260
	N	25	58	19	75
SAT mechanical	r	-,272	-,025	-,035	,033
	p	,189	,854	,884	,777
	N	25	58	20	76
Grade 12 English	r	,319	,534	,058	,082
	p	,112	,000**	,803	,485
	N	26	56	21	75
Grade 12 Mathematics	r	,526	,268	,506	,342
	p	,006**	,052	,027*	,003**
	N	26	53	19	73
Grade 12 Science	r	,454	,325	,330	,293
	p	,029*	,020*	,167	,014*
	N	23	51	19	70

** p < ,01 * p < ,05

7.2.2.5 Group 2: Regression analysis and comparison of regression lines

Despite the low correlations of the LPCAT (PPC) scores with first-year average performance, and although the sample size for Group 2 is relatively small and the group cannot be regarded as representative of first-year Technikon students, regression analysis was nevertheless performed using the average first-year score as the dependent variable (Y) and the LPCAT composite score as the independent variable (X). The regression analysis was done solely to compare the regression lines of subgroups. The results are discussed in terms of possible over- or underprediction of academic results, should the total group regression line be used . The regression equations obtained for the total group and different subgroups are as follows:

Total group:	Y = 0,403 (X) + 34,767 (N = 120)
Males:	Y = 0,542 (X) + 24,797 (N = 66)
Females:	Y = 0,218 (X) + 47,781 (N = 53)
African language:	Y = 0,07199 (X) + 49,071 (N = 47)
English/Afrikaans:	Y = 0,565 (X) + 28,856 (N = 70)

The graphic representation of these regression lines is given in Figure 7.15.

The regression lines for most of the groups are reasonably similar, except for that of the African language group. If the regression line for the total group is used to predict average first-year academic performance, the performance of male students will

generally be overestimated, while that of female students will generally be underestimated, in particular those who score low on the LPCAT. For the English/Afrikaans language group, if the total group regression line is used for prediction of academic performance, the average first-year academic performance will be slightly underestimated for those who score above 30 on the LPCAT composite score, that is, for practically all of this group. The regression line for the African language group is distinctly different from the others and has a very flat slope. This indicates that fairly large differences in LPCAT composite scores do not result in distinctly different predicted average academic scores. A smaller range of average first-year academic scores for this group could have explained such a slope. The mean average first-year academic score for the African language group is 53,43 with a standard deviation of 9,46, compared to the English/Afrikaans language group, where the mean first-year academic percentage is 60,48 with a standard deviation of 11,67. The differences between the two groups do not seem pronounced enough to warrant the flat slope of the African home language group regression line (see scatter diagram). If the total group regression line is used for predicting average first-year results for the African language group, this would result in an underestimation of academic results for African language examinees who obtain scores of below 40 on the LPCAT composite score, while the average academic performance of those African language candidates with LPCAT composite scores above 40 will generally be overestimated. The problems generally associated with prediction of tertiary academic performance of African (disadvantaged) students (Huysamen, 1999) probably contributed to these results.

FIGURE 7.15 GROUP 2: REGRESSION LINES FOR THE TOTAL, GENDER AND LANGUAGE GROUPS

7.2.2.6 Group 2: Overview and summary

For this particular group, there are also significant mean differences between the two language groups. On both the GSAT and the SAT, the groups differ on all measures. For the LPCAT, the language groups differ only on the post-test score. In terms of the academic measures, the language groups differ significantly on grade 12 English, grade 12 Science and first-year average. Comparison of the standardised mean differences between the two language groups by expressing the differences in terms of the common standard deviation, indicates that differences between the two language groups are smaller on the LPCAT than on the nonverbal score of the GSAT, and that both of these are much smaller than those of the verbal and total GSAT scores. The two gender groups differ only on SAT calculations, SAT mechanical, grade 12 English, grade 12 Science and first-year average. Except for the GSAT verbal scores, the frequency distributions of scores for the two language subgroups of Group 2 are reasonably similar.

Correlations of the LPCAT with the GSAT indicate that both tests measure the same construct. LPCAT correlations with Technikon first-year academic results are generally low, although some were statistically significant. The pattern of significant correlations for Group 2 is different from that of Group 1, with the result that it is not possible to provide a general pattern for Technikon first-year results. Although there is

less restriction of range in Group 2 than in Group 1, it is still a select group and correlations may be affected. Once again, grade 12 results have the highest correlations with first-year academic results for both language groups. Similar to what was found for Group 1, grade 12 English correlated highly significantly with average first-year performance for the African language group, but showed no significant correlations for the English/Afrikaans language group. This underscores the importance of language proficiency in the language of teaching for academic performance. Except for the African language group, regression lines were reasonably similar.

Taking all the information provided in this section into account, in the general reasoning ability domain, the LPCAT seems to provide equitable measures of learning potential for Technikon students.

The low and generally insignificant correlations of LPCAT (PPC) scores with first-year academic performance can probably be ascribed to a combination of

the restriction of range on both the predictor (LPCAT) and criterion (academic performance) scores, as well as

the problematic nature of the tertiary level academic criterion scores - academic average being calculated for different subjects for different students, and lastly

the effect of language proficiency in the language of teaching on academic performance

In order to investigate this hypothesis, the LPCAT-1 was administered to a grade 9 school group (Group 3).

7.2.3 LPCAT-1 validity results for Group 3

Group 3 consisted of 37 grade 9 pupils from an urban high school. This group was used to obtain empirical validity information for the LPCAT-1 for a group with a wider range of ability and where the academic criterion measures are more reliable. Junior levels of high-school generally represent a reasonably wide range of ability levels. Furthermore, since grade 9 pupils (with a few exceptions) all take the same core subjects, the average marks attained are more directly comparable than first-year average results. As criterion measures, the school results for the four terms as well

as the average year mark were used to shed light on the criterion-related (predictive) validity of the LPCAT-1. The multicultural composition of the sample provided the opportunity to investigate the cross-cultural functioning of the LPCAT-1 for these pupils.

7.2.3.1 Group 3: Comparison of mean scores

The descriptive statistics for Group 3 are provided in Table 7.13, together with the results of the comparisons of the mean scores of the two language groups by means of a t-test for independent samples.

TABLE 7.13 GROUP 3: DESCRIPTIVE STATISTICS AND COMPARISON OF LANGUAGE GROUP MEAN SCORES

	Total group			African language group			Eng/Afr language group			Comparison of means	
	N	Mean	SD	N	Mean	SD	N	Mean	SD	Mean diff #	p-value
LPCAT pretest	37	49,65	6,79	18	48,39	7,01	19	50,84	6,53	2,45	,278
LPCAT post-test	37	49,81	6,48	18	48,94	5,98	19	50,63	6,99	1,69	,436
LPCAT composite score	37	49,79	6,76	18	48,58	6,93	19	50,93	6,56	2,35	,297
LPCAT difference score	37	0,1622	2,90	18	0,56	3,13	19	-0,21	2,70	0,77	,430
Term 1 average	37	51,57	13,27	18	47,56	10,98	19	55,37	14,39	7,81	,073
Term 2 average	37	42,76	15,02	18	37,56	10,68	19	47,68	17,04	10,13	,037*
Term 3 average	37	46,97	13,90	18	44,28	11,48	19	49,53	15,74	5,25	,257
Term 4 average	37	42,22	14,36	18	38,28	11,27	19	45,95	16,19	7,67	,105
Year average	37	46,16	13,91	18	41,83	10,66	19	50,26	15,60	8,43	,065

absolute value of the difference between the mean scores

** p < ,01 * p < ,05 (p-values for a nondirectional t-test for independent samples between the two language groups)

The only score on which the two language groups show a significant difference is the second-term academic average. On all the LPCAT scores as well as on most of the academic scores, there are no significant differences between the two groups.

The descriptive statistics for the gender groups on the LPCAT and academic results as well as the results of comparing the mean scores by means of a t-test for independent samples, are reported in Table 7.14.

TABLE 7.14 GROUP 3: DESCRIPTIVE STATISTICS AND COMPARISON OF GENDER GROUP MEAN SCORES

Variable	Females			Males			Mean diff #	p-value
	N	Mean	SD	N	Mean	SD		
LPCAT pretest	17	49,18	7,32	20	50,05	6,48	0,87	,702
LPCAT post-test	17	49,12	6,98	20	50,40	6,14	1,28	,556
LPCAT composite score	17	49,24	7,30	20	50,25	6,41	1,00	,659
LPCAT difference score	17	-0,059	2,41	20	0,350	3,31	0,409	,675
Term 1 average	17	54,12	16,89	20	49,40	9,10	4,72	,313
Term 2 average	17	45,94	17,60	20	40,05	12,23	5,89	,240
Term 3 average	17	51,82	16,38	20	42,85	10,05	8,97	,049*
Term 4 average	17	45,24	17,18	20	39,65	11,27	5,59	,244
Year average	17	49,59	16,95	20	43,25	10,26	6,34	,170

absolute value of the difference between the mean scores

** p < ,01 * p < ,05

(p-value for a nondirectional t-test for independent samples between the two gender groups)

The only score on which the two gender groups differ significantly is the third-term average. On all the LPCAT scores as well as on the other academic scores, there is no significant difference between the two gender groups.

7.2.3.2 Group 3: Distribution of scores

Figures 7.16 to 7.19 indicate the distributions of scores on the LPCAT and academic results for the two language groups respectively.

FIGURE 7.16 GROUP 3: DISTRIBUTION OF GRADE 9 AVERAGE YEAR MARKS

FIGURE 7.17 GROUP 3: DISTRIBUTION OF LPCAT PRETEST SCORES

FIGURE 7.18 GROUP 3: DISTRIBUTION OF LPCAT POST-TEST SCORES

FIGURE 7.19 GROUP 3: DISTRIBUTION OF LPCAT COMPOSITE SCORES

The frequency distribution of the academic scores has a slight positive skewness for both groups with that of the African language group being more pronounced. Although some differences between the two language groups are also evident for the LPCAT distributions of scores, these are less pronounced than those for the academic results.

7.2.3.3 Group 3: LPCAT correlations with criterion measures

It was hypothesised that the low correlations found between LPCAT-1 (PPC) results and academic criteria at Technikon level, were largely a result of restriction of range and the unreliability of criterion scores. It was also hypothesised that, for groups with a wide range of ability, the correlations of the LPCAT difference score with criterion measures will tend towards zero. The correlations of the LPCAT scores with the academic results for Group 3 are provided in Table 7.15.

TABLE 7.15 GROUP 3: CORRELATIONS OF LPCAT-1 WITH GRADE 9 ACADEMIC RESULTS (N=37)

LPCAT score		Term 1	Term 2	Term 3	Term 4	Year mark
LPCAT pretest	r	,474	,632	,557	,550	,591

	p	,003**	,000**	,000**	,000**	,000**
LPCAT post-test	r	,536	,659	,558	,588	,619
	p	,001**	,000**	,000**	,000**	,000**
LPCAT composite	r	,477	,635	,599	,555	,594
score	p	,003**	,000**	,000**	,000**	,000**
LPCAT difference	r	,088	-,007	-,058	,026	,000
score	p	,603	,966	,731	,876	1,000

* p < ,05 ** p < ,01

These results provide support for the hypotheses, since the LPCAT (PPC) scores all correlate highly significantly with all the academic scores, while the LPCAT difference score shows no significant correlation with any of the academic scores. The term 4 results provide concurrent validity information for the LPCAT-1, since these results were obtained during the same time.

To investigate whether restriction of range is present, a scatter diagram of the LPCAT composite scores and grade 9 average scores is presented in Figure 7.20.

FIGURE 7.20 GROUP 3: SCATTER DIAGRAM OF LPCAT COMPOSITE SCORES AND GRADE 9 AVERAGE MARKS PER LANGUAGE GROUP

Despite the small sample size (N=37), the scatter diagram shows a reasonably wide distribution of scores on both the LPCAT and the year average results.

Notwithstanding the small sample size, it was decided to investigate the criterion-related validity for the two language groups separately. In Table 7.16, the correlations of the LPCAT with academic results are provided separately for the two language groups.

TABLE 7.16 GROUP 3: CORRELATIONS OF LPCAT WITH GRADE 9 RESULTS FOR THE TWO LANGUAGE GROUPS

Predictor variable		Term 1	Term 2	Term 3	Term 4	Year average
			African language group			
LPCAT pretest score	r	,243	,393	,226	,278	,311
	p	,331	,106	,368	,264	,209
	N	18	18	18	18	18
LPCAT post-test score	r	,289	,430	,224	,327	,342
	p	,245	,075	,371	,185	,165
	N	18	18	18	18	18
LPCAT composite score	r	,243	,397	,223	,283	,312
	p	,330	,103	,374	,256	,207
	N	18	18	18	18	18
LPCAT difference score	r	,008	-,061	-,078	,002	-,045
	p	,976	,811	,758	,994	,860
	N	18	18	18	18	18
			English/Afrikaans group			
LPCAT pretest score	r	,616	,793	,791	,727	,776
	p	,005**	,000**	,000**	,000**	,000**
	N	19	19	19	19	19
LPCAT post-test score	r	,672	,786	,740	,725	,769
	p	,002**	,000**	,000**	,000**	,000**
	N	19	19	19	19	19
LPCAT composite score	r	,621	,795	,793	,730	,779
	p	,005**	,000**	,000**	,000**	,000**
	N	19	19	19	19	19
LPCAT difference score	r	,248	,114	,000	,118	,112
	p	,306	,641	1,000	,631	,647
	N	19	19	19	19	19

* p < ,01 * p < ,05

Keeping in mind the small sizes of the two groups, a distinct pattern nevertheless seems to emerge. For the African language group, none of the LPCAT (PPC) scores correlate significantly with any of the academic results, although these correlations are consistently higher than those found for the Technikon groups. According to expectations, the LPCAT difference score correlations with academic performance for

this group are low. For the English/Afrikaans group, however, the results look quite different with all of the LPCAT (PPC) scores correlating numerically highly and statistically highly significantly with all the academic results. The correlations of the LPCAT difference score with academic results, although not all close to zero, are nevertheless of a much smaller magnitude than those of the other LPCAT scores.

7.2.3.4 Group 3: Regression analysis and comparison of regression lines

Regression analysis was not performed for this group, owing to the small size of the sample and subgroups.

7.2.3.5 Group 3: Overview and summary

Less information is available for this group than for the two preceding groups. For this particular group, only one mean academic measure differed significantly between the two language groups, namely the term 2 average, indicating general similarity in academic performance for the two groups. Similar results were found for the two gender groups. The distributions of scores seem to indicate a more similar pattern for the two language groups on the LPCAT than for the average grade 9 year mark, although differences on the LPCAT were evident.

The correlations of the LPCAT with academic criterion measures indicate a distinct difference between the two language groups. For the English/Afrikaans group, all the LPCAT (PPC) scores correlate highly significantly with all academic measures, while for the African language group, no significant correlations are found. Unfortunately, no other results were available for this particular group, which precluded investigation of the hypothesis regarding the importance of language proficiency for academic performance, for the African language group in particular. This aspect will be investigated in the remaining groups.

The wider range of ability levels and better academic criterion measures seem to have resulted in higher correlations of the LPCAT (PPC) scores with academic results.

However, the distinct differences between the two language groups regarding criterion-related validity require further investigation. Based on the results of the two Technikon groups where other information was also available, English proficiency seems to play an important role in the academic performance of the African home language examinees in particular. For multicultural groups, the LPCAT nevertheless seems to provide an equitable measure of learning potential in the domain of general nonverbal reasoning ability, that is not as reliant on language proficiency or prior formal learning experiences as most standard tests. Despite the fact that it may not correlate as highly with academic results for the African language group as for the English/Afrikaans group, it can provide useful additional information for selection or evaluation purposes.

7.3 EMPIRICAL VALIDITY RESULTS FOR THE LPCAT-2

For the empirical validity evaluation of the LPCAT-2, two quite different groups were used. The first group consisted of low-literacy adult learners (N=194), with levels of education ranging from grade 1 to grade 12, while the second group consisted of 144 grade 8 school pupils. Both of these groups include a broad range of ability levels.

7.3.1 LPCAT-2 validity results for Group 4

The adult learner sample (Group 4) was being evaluated for vocational training after their retrenchment from a governmental organisation. The LPCAT-2 was included in the battery of tests administered to them as part of their general evaluation. As criterion measures, the Paper-and-Pencil Games (PPG) (Claassen, 1996) as well as literacy and numeracy assessment were used to shed light on the construct and criterion-related validity of the LPCAT-2. The sample consisted of African home language participants only, most of whom were male. The sample is therefore relatively homogeneous in terms of language group and gender. Consequently no comparisons between either language or gender subgroups were possible.

7.3.1.1 Group 4: Mean scores

The descriptive statistics for Group 4 on the LPCAT and other measures are provided in Table 7.17.

TABLE 7.17 GROUP 4: DESCRIPTIVE STATISTICS FOR MEASURES

	N	Mean	SD	Minimum	Maximum
LPCAT pretest	194	36,19	7,94	26	58
LPCAT post-test	194	37,76	9,00	23	60
LPCAT composite score	194	36,64	7,97	26	58
LPCAT difference score	194	1,57	4,03	-8	14
PPG verbal	110	36,50	7,12	4	49
PPG nonverbal	110	36,54	12,46	3	50
PPG total	110	72,79	17,70	2	98
Literacy level 1 total	182	47,79	15,81	9	73
Numeracy level 1 total	182	9,92	5,33	1	34
Literacy level 3 total	111	43,02	15,24	14	95
Numeracy level 3 total	26	13,85	6,53	3	29

The results indicate a reasonably wide range of ability (LPCAT and PPG) and proficiency (literacy and numeracy) scores for Group 4. The mean LPCAT scores are below the grade 10 average of 50. This was to be expected, considering that the average education was grade 8. For the PPG, the mean for Group 4 is above the average of 25. This can be explained by the fact that the comparison group for the PPG comprises primary school pupils with three to four years of education.

7.3.1.2 Group 4: Distribution of scores

The distribution of verbal and nonverbal PPG scores for Group 4 is provided in Figures 7.21 and 7.22.

FIGURE 7.21 GROUP 4: DISTRIBUTION OF PPG VERBAL SCORES

The verbal and the nonverbal scores of the adult learner group on the PPG show a negatively skewed distribution. This indicates that the scores were generally high on this test, as also indicated by the means. For the nonverbal scores there seems to be a ceiling effect with a high frequency of scores in the top three score categories. The distribution of LPCAT pretest, post-test and composite scores for Group 4 is provided in Figures 7.23 to 7.25.

FIGURE 7.22 GROUP 4: DISTRIBUTION OF PPG NONVERBAL SCORES

FIGURE 7.23 GROUP 4: DISTRIBUTION OF LPCAT PRETEST SCORES

FIGURE 7.24 GROUP 4: DISTRIBUTION OF LPCAT POST-TEST SCORES

FIGURE 7.25 GROUP 4: DISTRIBUTION OF LPCAT COMPOSITE SCORES

Most of the LPCAT scores fall below 50, the mean T-score value of the grade 10 comparison group. Owing to the wide range and relatively low grade 8 average level of education, this distribution is to be expected for the group concerned. The LPCAT nevertheless provides a reasonable distribution and a wide range of scores, indicating that it can distinguish between individuals at this level.

7.3.1.3 Group 4: LPCAT correlations with the PPG

For the adult learner group, the construct validity of the LPCAT was evaluated by correlating the LPCAT results with those of the PPG (Claassen, 1996). These results are reported in Table 7.18.

TABLE 7.18 GROUP 4: CORRELATIONS OF LPCAT WITH PPG (N = 110)

		PPG verbal	PPG nonverbal	PPG total
LPCAT pretest	r	,400	,542	,530
	p	,000**	,000**	,000**
LPCAT post-test	r	,408	,645	,610
	p	,000**	,000**	,000**
LPCAT composite score	r	,413	,570	,556
	p	,000**	,000**	,000**
LPCAT difference score	r	,121	,371	,315
	p	,207	,000**	,001**

** p < ,01

The results indicate that the LPCAT (PPC) scores measure a construct that overlaps with that measured by the PPG. The correlations of the LPCAT (PPC) scores with the verbal scores of the PPG are slightly lower numerically, although still statistically highly significant. The correlations of the LPCAT post-test score and LPCAT composite score with the PPG scores are consistently higher than those of the LPCAT pretest score with the PPG scores, indicating support for learning potential scores that include both present level of performance as well as the effect of training. The LPCAT difference score correlates significantly with the PPG nonverbal and PPG total scores, indicating that larger increases from pretest to post-test on the LPCAT are associated with higher PPG nonverbal and total scores overall. Although statistically significant, these correlations are smaller than those of the other LPCAT (PPC) scores.

7.3.1.4 Group 4: LPCAT correlations with criterion measures

For the adult learner group the only criterion measures available are literacy and numeracy scores. Although these individuals subsequently underwent practical vocational training, the kinds of training offered were extremely diverse, and were also offered at different levels. Hence very few individuals took exactly the same vocational training courses at the same levels and the training results could therefore not be compared or used as criterion measures. The criterion-related validity results of the adult learner group using literacy and numeracy results are provided in Table 7.19.

All the LPCAT (PPC) score correlations with the literacy and numeracy scores are statistically significant. With the exception of the LPCAT pretest correlation with level 3 numeracy, they are all statistically highly significant. The LPCAT post-test and LPCAT composite score correlations with the literacy and numeracy scores are also consistently higher than the correlations of the LPCAT pretest with the criterion scores - indicating support for measures of learning potential that incorporate present level of performance as well as the effect of training. The LPCAT difference score also correlates highly significantly with level 1 literacy and level 3 numeracy and significantly with level 1 numeracy. The correlations of the LPCAT difference score with literacy and numeracy are smaller than those of the LPCAT (PPC) scores except for level 3 numeracy scores, where the correlation with the LPCAT difference score is higher. The group that completed the level 3 numeracy evaluation (N=26) represents a selected group within this sample. These results therefore support the hypothesis that the more homogeneous the group, the higher the correlation of the LPCAT difference score with the criterion will be.

TABLE 7.19 GROUP 4: CORRELATIONS OF LPCAT SCORES WITH LITERACY AND NUMERACY RESULTS

		Level 1 literacy	Level 1 numeracy	Level 3 literacy	Level 3 numeracy
LPCAT pretest	r	,398	,474	,434	,455
	p	,000**	,000**	,000**	,020*
	N	182	182	111	26

LPCAT post-test	r	,437	,491	,461	,610
	p	,000**	,000**	,000**	,001**
	N	182	182	111	26
LPCAT composite score	r	,418	,492	,456	,518
	p	,000**	,000**	,000**	,007**
	N	182	182	111	26
LPCAT difference score	r	,197	,168	,168	,556
	p	,000**	,024*	,078	,003**
	N	182	182	111	26

** p < ,01 * p < ,05

A scatter diagram of the LPCAT composite score and a score obtained by adding together the level 1 literacy and numeracy scores is provided in Figure 7.26.

Most of the LPCAT scores fall below 50, but there is a reasonable spread of scores below this point. Considering that Group 4 consisted of low- literacy adults, and that a score of 50 on the LPCAT represents the average performance of a person with approximately grade 10 education, the LPCAT score distribution for Group 4 is as expected. The wide range of education of this group is reflected in the distribution of the combined literacy/numeracy scores.

FIGURE 7.26 GROUP 4: SCATTER DIAGRAM OF LPCAT COMPOSITE SCORES AND COMBINED LEVEL 1 LITERACY AND NUMERACY SCORES

7.3.1.5 Group 4: Overview and summary

The information available for this group shows the LPCAT to be a useful measure of learning potential in the general reasoning ability domain for this level of examinees. The distribution of LPCAT (PPC) scores indicates that the LPCAT-2 can distinguish between examinees at this (lower) level. Correlations with the PPG indicate that the two tests measure overlapping constructs, which supports the validity of the LPCAT as a measure of learning potential in the general reasoning ability domain. All the LPCAT (PPC) scores correlate highly significantly with the literacy and numeracy results, thereby supporting the criterion-related validity of the LPCAT-2. The higher correlations of the LPCAT post-test and LPCAT composite scores with criterion measures when compared to the LPCAT pretest, provide support for measures of learning potential that include both present level of performance as well as the effect of training (ie difference score). The numerically high and statistically highly significant correlations of the LPCAT difference score with level 3 numeracy can probably be ascribed to the selectiveness of the group involved. The indications are that the more similar the group, the higher the correlation of the LPCAT difference score with the criterion measures will be, compared to LPCAT (PPC) score correlations with the

same criterion measures.

7.3.2 LPCAT-2 validity results for Group 5

Group 5 consisted of 144 grade 8 high school pupils from an urban high school. As criterion measures, the results of the GSAT-CAT (Van Tonder & Claassen, 1992), academic results, proficiency test results in English and Mathematics, results of the Learning Process Questionnaire (LPQ) (Biggs, 1987a) as well as two teacher ratings were used to obtain further information on the construct and criterion-related validity of the LPCAT-2. At junior high school level, the group should represent a wide range of ability levels. The multicultural composition of the sample also provided an opportunity to further investigate the cross-cultural functioning of the LPCAT.

7.3.2.1 Group 5: Comparison of mean scores

The descriptive statistics for Group 5 of the LPCAT-2 and other measures are provided in Table 7.20, together with the results of the comparison of the mean scores for the two language groups by means of a t-test for independent samples. The p-values provided are for a nondirectional t-test for independent samples between the two language groups.

TABLE 7.20 GROUP 5: DESCRIPTIVE STATISTICS AND COMPARISON OF LANGUAGE GROUP MEAN SCORES

	Total group			African language group			Eng/Afr language group			Comparison of means	
	N	Mean	SD	N	Mean	SD	N	Mean	SD	Mean diff #	p-value
LPCAT pretest	128	45,67	8,38	44	40,16	8,73	82	48,57	6,67	8,41	,000**
LPCAT post-test	128	47,83	7,50	44	42,07	8,24	82	50,83	4,96	8,76	,000**
LPCAT composite score	128	46,03	8,25	44	40,46	8,59	82	48,97	6,46	8,51	,000**
LPCAT difference score	128	2,16	3,40	44	1,91	3,39	82	2,26	3,46	0,35	,590
GSAT-CAT verbal	133	94,32	11,21	49	87,33	7,77	83	98,58	10,86	11,25	,000**
GSAT-CAT nonverbal	133	91,67	13,09	49	85,14	10,00	83	95,51	13,30	10,36	,000**
GSAT-CAT total	133	93,23	11,25	49	86,49	7,77	83	97,28	11,15	10,79	,000**
English proficiency 1	144	47,83	10,90	53	41,64	9,11	89	51,43	10,36	9,79	,000**
Mathematics proficiency 1	144	44,31	15,05	53	35,40	8,56	89	49,70	15,71	14,30	,000**
LPQ surface	121	36,66	7,23	41	34,51	6,23	79	37,84	7,51	3,32	,017*
LPQ deep	121	39,50	7,50	41	39,00	7,54	79	39,82	7,53	0,82	,571
LPQ achieving	121	39,02	7,11	41	39,83	6,84	79	38,61	7,30	1,22	,376
English teacher rating	134	33,69	11,99	48	25,88	8,24	85	38,12	11,61	12,24	,000**
Mathematics teacher rating	134	35,14	10,93	48	31,90	9,95	85	36,94	11,15	5,05	,010*
First term average	131	55,99	16,95	47	45,21	10,42	83	62,20	16,97	16,99	,000**

Second term average	133	47,57	17,67	47	36,23	11,00	83	54,02	17,64	17,79	,000**
Third term average	133	49,98	17,57	47	40,91	10,96	83	55,10	18,66	14,18	,000**
Fourth term average	133	47,21	18,11	47	37,13	11,58	83	52,82	18,81	15,69	,000**
Average year mark	133	50,37	17,08	47	39,79	10,39	83	56,40	17,35	16,61	,000**

The only scores for which there are no statistically highly significant differences between the mean scores of the two language groups are the LPCAT difference score, the LPQ deep score and the LPQ achieving score. Regarding the LPCAT difference score, this means that the potential to improve upon LPCAT pretest performance following relevant training seems to be similar for the two language groups, despite the fact that their mean performance on all other measures differs significantly. The fact that the mean LPQ deep and achieving scores are not significantly different for the two language groups, indicates similarities in the learning attitude of the two groups.

To further interpret the mean score differences between the two language groups, the differences on the LPCAT and GSAT-CAT as general cognitive measures are expressed as a proportion of the theoretical standard deviation, the standard deviation of the total group and the common standard deviation (see Table 7.21). In this way the differences on different measures become directly comparable.

TABLE 7.21 GROUP 5: MEAN DIFFERENCES BETWEEN LANGUAGE GROUPS AS A PROPORTION OF DIFFERENT STANDARD DEVIATION UNITS*

Variable	Mean difference	Proportion of theoretical SD*	Proportion of total group SD*	Proportion of common SD*
LPCAT pretest	8,41	0,841 [10]	1,004 [8,38]	1,092 [7,70]
LPCAT post-test	8,76	0,876 [10]	1,168 [7,50]	1,327 [6,60]
LPCAT composite score	8,51	0,851 [10]	1,032 [8,25]	1,130 [7,53]
GSAT verbal score	11,25	0,75 [15]	1,004 [11,21]	1,207 [9,32]
GSAT non-verbal score	10,36	0,691 [15]	0,791 [13,09]	0,889 [11,65]
GSAT total score	10,79	0,719 [15]	0,959 [11,25]	1,141 [9,46]

* Standard deviation scores in square brackets below the proportional value

For Group 5, the mean differences as a proportion of different standard deviation

scores for the LPCAT and the GSAT-CAT are of a similar magnitude, although the proportional value of the GSAT-CAT nonverbal score is generally smaller.

The descriptive statistics for the two gender groups are provided in Table 7.22 together with the result of comparing the mean scores of the two groups by means of a t-test for independent samples.

TABLE 7.22 GROUP 5: DESCRIPTIVE STATISTICS AND COMPARISON OF GENDER GROUP MEAN SCORES

Variable	Females			Males			Mean diff #	p-value
	N	Mean	SD	N	Mean	SD		
LPCAT pretest	69	45,20	8,27	59	46,22	8,54	1,02	,496
LPCAT post-test	69	47,36	7,50	59	48,37	7,52	1,01	,449
LPCAT composite score	69	45,55	8,16	59	46,60	8,37	1,06	,472
LPCAT difference score	69	2,16	3,27	59	2,15	3,58	0,01	,991
GSAT verbal	74	95,19	11,81	59	93,22	10,42	1,97	,316
GSAT nonverbal	74	91,20	13,36	59	92,25	12,83	1,05	,647
GSAT total	74	93,41	11,78	59	93,02	10,63	0,39	,844
English proficiency	78	49,38	9,68	66	45,98	12,00	3,40	,067
Mathematics proficiency	78	44,21	15,63	66	44,42	14,45	0,22	,931
LPQ surface	67	36,82	7,16	54	36,46	7,38	0,36	,788
LPQ deep	67	39,46	8,14	54	39,54	6,69	0,07	,957
LPQ achieving	67	39,01	7,74	54	39,02	6,32	0,004	,998
English teacher rating	71	37,59	11,84	63	29,29	10,63	8,31	,000**
Mathematics teacher rating	71	37,03	10,99	63	33,02	10,55	4,01	,033*
First term average	68	62,35	16,05	63	49,13	15,23	13,23	,000**
Second term average	68	51,84	18,11	63	42,97	16,09	8,87	,004**
Third term average	68	55,34	17,30	63	44,21	16,06	11,13	,000**

Fourth term average	68	51,62	18,17	63	42,46	16,93	9,16	,003**
Year mark average	68	55,41	17,10	63	44,94	15,43	10,48	,000**

absolute value of the difference between the mean scores ** p < ,01 * p < ,05

(p-values for a nondirectional t-test for independent samples between the two gender groups)

There are no significant differences between the gender groups on any of the LPCAT, GSAT-CAT or proficiency test results. There are, however, significant differences between the two gender groups on all the measures of grade 8 academic performance, with the female group obtaining higher means. The female group was also rated significantly higher in both the English and the Mathematics teacher rating. No significant differences were found on any of the LPQ scores, indicating that the study attitudes of the two gender groups are generally the same.

7.3.2.2 Group 5: Distribution of scores

The frequency distributions of GSAT-CAT scores, average year marks for grade 8 and LPCAT (PPC) scores for the two language groups are provided in Figures 7.27 to 7.32.

FIGURE 7.27 GROUP 5: DISTRIBUTION OF GSAT VERBAL SCORES

FIGURE 7.28 GROUP 5: DISTRIBUTION OF GSAT NONVERBAL SCORES

FIGURE 7.29 GROUP 5: DISTRIBUTION OF AVERAGE YEAR MARKS

FIGURE 7.30 GROUP 5: DISTRIBUTION OF LPCAT PRETEST SCORES

FIGURE 7.31 GROUP 5: DISTRIBUTION OF LPCAT POST-TEST SCORES

FIGURE 7.32 GROUP 5: DISTRIBUTION OF LPCAT COMPOSITE SCORES

The GSAT-CAT and year mark distributions are somewhat more positively skewed for the African language group than for the English/Afrikaans group. The LPCAT (PPC) score distributions of the African language group are more symmetrical, while those of the English/Afrikaans group are slightly negatively skewed. Most scores for both groups are below the average score of 50, which is to be expected considering the educational level of Group 5.

7.3.2.3 Group 5: LPCAT correlations with the GSAT-CAT

Construct validity was evaluated by comparing the results of the LPCAT with results of the GSAT-CAT (Van Tonder & Claassen, 1992). See Table 7.23.

TABLE 7.23 GROUP 5: CORRELATIONS OF LPCAT WITH GSAT-CAT (N = 120)

		GSAT verbal	GSAT nonverbal	GSAT total
LPCAT pretest	r	,567	,639	,652
	p	,000**	,000**	,000**
LPCAT post-test	r	,613	,665	,691
	p	,000**	,000**	,000**

LPCAT composite	r	,575	,651	,663
score	p	,000**	,000**	,000**
LPCAT difference	r	-,042	-,103	-,080
score	p	,648	,264	,383

** p < ,01

The results are consistent with the results for the previous groups and indicate that the LPCAT learning potential measures cover a construct domain reasonably similar to that measured by the GSAT. As expected, the correlations of the LPCAT (PPC) scores with the GSAT nonverbal scores are somewhat higher in magnitude than those with the GSAT verbal scores. The LPCAT post-test and LPCAT composite score correlations with the GSAT scores are generally higher than those of the LPCAT pretest score with GSAT scores, providing support for learning potential measures that include both pretest performance as well as the effect of training. For Group 5, where a broad range of abilities is present, the LPCAT difference score shows negligible correlations with the GSAT.

7.3.2.4 Group 5: LPCAT correlations with criterion measures

For Group 5, the four term average percentage marks and year percentage mark were used as criterion scores. The results are presented in Table 7.24.

All the LPCAT (PPC) scores correlate statistically highly significantly with all academic results. Once again, the LPCAT post-test and LPCAT composite score correlations with academic criteria are consistently higher than the LPCAT pretest correlations with the same criterion scores - providing further support for the use of learning potential scores that contain measures of pretest performance as well as the effect of training. As expected, the LPCAT difference score correlations with the school grade 8 results all tend towards zero.

TABLE 7.24 GROUP 5: CORRELATIONS OF LPCAT, GSAT-CAT AND PROFICIENCY TEST RESULTS WITH ACADEMIC

RESULTS

Variable		Term 1	Term 2	Term 3	Term 4	Average year mark
LPCAT pretest score	r	,460	,479	,439	,454	,474
	p	,000**	,000**	,000**	,000**	,000**
	N	116	116	116	116	116
LPCAT post-test score	r	,530	,543	,489	,524	,538
	p	,000**	,000**	,000**	,000**	,000**
	N	116	116	116	116	116
LPCAT composite	r	,472	,489	,448	,467	,485
	p	,000**	,000**	,000**	,000**	,000**
	N	116	116	116	116	116
LPCAT difference score	r	,070	,053	,028	,074	,053
	p	,454	,572	,764	,427	,571
	N	116	116	116	116	116
GSAT-CAT verbal score	r	,679	,723	,669	,680	,715
	p	,000**	,000**	,000**	,000**	,000**
	N	122	122	122	122	122
GSAT-CAT nonverbal score	r	,558	,609	,574	,576	,603
	p	,000**	,000**	,000**	,000**	,000**
	N	122	122	122	122	122
GSAT-CAT total score	r	,666	,719	,672	,677	,711
	p	,000**	,000**	,000**	,000**	,000**
	N	122	122	122	122	122
Mathematics proficiency	r	,669	,724	,671	,712	,710
	p	,000**	,000**	,000**	,000**	,000**
	N	131	131	131	131	131
English proficiency	r	,677	,681	,632	,672	,678
	p	,000**	,000**	,000**	,000**	,000**
	N	131	131	131	131	131

** p < ,001

The correlations between the GSAT-CAT results and academic performance are also highly significant and of a higher magnitude than those of the LPCAT (PPC) scores. Since the overlap in content between the GSAT-CAT and school subjects is greater, larger correlations could be expected. Both the Mathematics proficiency and the English proficiency tests also correlate statistically highly significantly with all school results. The higher correlations found for this group compared to those for the Technikon first-year students are probably largely due to the broader range of abilities

found in this Grade 8 group as well as the better quality of the school academic criterion measures.

A scatter diagram of the LPCAT composite scores with the average year mark is provided in Figure 7.33, indicating reasonable ranges of scores for both measures. This may help explain why the correlations between these scores are higher for Group 5 than for the Technikon first-year groups. However, differences between the two language groups are evident in the scatter diagram, which indicates a more pronounced restriction of range within each language group. The correlations of the LPCAT, GSAT, and proficiency test scores with grade 8 academic performance are reported separately for the two language groups in Table 7.25.

TABLE 7.25 GROUP 5: CORRELATIONS OF LPCAT, GSAT AND PROFICIENCY TESTS WITH ACADEMIC RESULTS PER LANGUAGE GROUP

		African home language group					English / Afrikaans home language group				
		Term 1	Term 2	Term 3	Term 4	Year mark	Term 1	Term 2	Term 3	Term 4	Year mark
	N	38	38	38	38	38	77	77	77	77	77
LPCAT pretest	r	,337	,255	,280	,249	,298	,335	,397	,376	,387	,383
	p	,039*	,123	,089	,132	,069	,003**	,000**	,001**	,001**	,001**
LPCAT post-test	r	,315	,245	,241	,219	,281	,467	,524	,497	,546	,513
	p	,054	,138	,145	,187	,087	,000**	,000**	,000**	,000**	,000**
LPCAT composite score	r	,341	,255	,280	,250	,300	,347	,410	,388	,406	,396
	p	,036*	,122	,089	,131	,067	,002**	,000**	,000**	,000**	,000**
LPCAT difference score	r	-,088	-,047	-,129	-,103	-,072	,053	,025	,017	,069	,025
	p	,600	,781	,439	,538	,666	,646	,896	,887	,549	,831
GSAT-CAT verbal score	N	43	43	43	43	43	78	78	78	78	78
	r	,288	,334	,319	,282	,337	,664	,713	,665	,689	,701
	p	,061	,029*	,037*	,067	,027*	,000**	,000**	,000**	,000**	,000**
GSAT-CAT nonverbal	r	,229	,154	,177	,083	,192	,533	,633	,588	,613	,612
score	p	,139	,325	,258	,597	,217	,000**	,000**	,000**	,000**	,000**
GSAT-CAT total score	r	,292	,270	,275	,195	,295	,634	,729	,679	,705	,711
	p	,057	,080	,074	,209	,055	,000**	,000**	,000**	,000**	,000**
English proficiency	N	47	47	47	47	47	83	83	83	83	83
	r	,671	,470	,552	,498	,555	,570	,638	,567	,636	,613
	p	,000**	,001**	,000**	,000**	,000**	,000**	,000**	,000**	,000**	,000**
Mathematics proficiency	N	47	47	47	47	47	83	83	83	83	83
	r	,230	,201	,180	,202	,201	,634	,723	,670	,722	,701
	p	,121	,175	,225	,173	,176	,000**	,000**	,000**	,000**	,000**

** p < ,01 * p < ,05

FIGURE 7.33 GROUP 5: SCATTER DIAGRAM OF LPCAT COMPOSITE SCORES AND GRADE 8 AVERAGE YEAR MARK PER LANGUAGE GROUP

For the African home language group, the only significant correlations are those of the LPCAT pretest and composite scores with first-term results. For the English/Afrikaans group the LPCAT (PPC) scores all correlate highly significantly with all the academic results. For both groups, the LPCAT difference score correlations with academic performance are negligibly small. An interesting result is that for the African language group, the LPCAT composite score gives the highest correlation of all the LPCAT score correlations with academic results, while for the English/Afrikaans group, the post-test gives the highest correlation. Both these scores are nevertheless measures that include both present level of performance (pretest) as well as the effect of training (difference score), albeit not in exactly the same way. In the LPCAT post-test, the effect of training is included more directly than in the case of the LPCAT composite score.

For the African language group, on the GSAT, only the verbal score correlates significantly with some academic results, namely with the second and third term and with the year mark. For this group, the GSAT verbal score and the LPCAT (PPC) scores generally correlate more highly with academic scores than the GSAT nonverbal score. For the English/Afrikaans group, the GSAT scores all correlate highly significantly with all academic scores, and these correlations are all larger than those

of the LPCAT with academic results. For the Mathematics proficiency scores, the same pattern as in the GSAT emerges with no significant correlations with school results for the African home language group and all correlations with school results being numerically high and statistically highly significant for the English/Afrikaans group. For the English proficiency scores, a different pattern emerges. English proficiency correlations with academic results, for both language groups are statistically highly significant. For the African home language group, these correlations are numerically the highest of all correlations with academic results, and of the same magnitude as those for the English/Afrikaans group. While the magnitude of correlations of English and Mathematics proficiency with academic results is similar for the English/Afrikaans language group, for the African language group the English proficiency correlations with academic performance are much higher than those of the Mathematics proficiency. These results underscore the importance of language proficiency in academic performance - in particular for those for whom the language of training is not their first language.

Overall, the LPCAT (PPC) correlations with school academic performance are reasonably similar to those of the GSAT-CAT for each group respectively, although there are noticeable differences between the two language groups.

7.3.2.5 Group 5: Regression analysis and comparison of regression lines

Regression analysis was performed for Group 5, using the average year mark as the dependent variable (Y) and the LPCAT composite score as the independent variable (X). This was done solely to compare the regression lines of subgroups and to investigate possible over- or underprediction of academic results, should the total group regression line be used. The regression equations for the total group and different subgroups are as follows:

Total group: $Y = 1{,}027\ (X) + 3{,}517$ (N = 116)

Males: $Y = 0{,}946\ (X) +$

	1,832 (N = 56)
Females:	Y = 1,230 (X) - 0,625 (N = 60)
African language:	Y = 0,361 (X) + 25,369 (N = 77)
English/Afrikaans:	Y = 1,078 (X) + 3,710 (N = 38)

These regression lines are shown in Figure 7.34.

FIGURE 7.34 GROUP 5: REGRESSION LINES FOR THE TOTAL, GENDER AND LANGUAGE GROUPS

With the exception of the regression line for the African language group, the regression lines for the different groups are reasonably similar. Should the total group regression formula (line) be used, the scores for the English/Afrikaans group as well as those for the female group are likely to be slightly underpredicted over the entire LPCAT composite score range, while scores for males are likely to be slightly overpredicted. The slope of the regression line for the African language group is noticeably different from those of the other regression lines. In the case of the African language group, academic performance is likely to be underpredicted for those who obtain LPCAT

composite scores lower than 35, while academic performance is likely to be over predicted for those obtaining scores higher than or equal to 35, should the regression formula for the total group be used.

7.3.2.6 Group 5: Overview and summary

Group 5 shows statistically significant differences between the mean scores of the two language groups on most measures. The gender groups obtained very similar results on all the psychometric tests, but differ significantly on all the academic results, with the female group obtaining higher scores than the male group. Differences between the two language groups are evident from the frequency distributions of scores on the GSAT, academic and LPCAT measures. For the African language group, the frequency distribution of LPCAT (PPC) scores are somewhat more symmetrical than those for the standard cognitive and academic scores.

The LPCAT (PPC) scores correlate statistically highly significantly and numerically highly with the GSAT-CAT, indicating support for the LPCAT as a measure of learning potential within the general reasoning ability domain. The correlations of the LPCAT with criterion measures are noticeably higher for Group 5 than for the other groups, probably as a result of both the wider range of ability levels in grade 8 - indicated by the scatter diagram - and the higher reliability of the school academic results.

Distinct differences can be seen between the two language groups regarding the correlations of the LPCAT with academic performance. Whereas for the English/Afrikaans group, all LPCAT (PPC) scores correlated highly significantly with all the academic measures, for the African language group only two of the 15 correlations were at all significant - that of the LPCAT pretest and LPCAT composite score with first- term results. One similarity, however, was the almost zero correlations of the LPCAT difference scores with all academic results for both language groups. A crucial result is the relation between language proficiency and academic performance. For the English/Afrikaans group, these correlations were high and statistically highly significant and of a similar magnitude compared to the GSAT correlation with

academic results. For the African language group, the correlations between English proficiency and academic results were the highest of all the correlations. Also, for the African language group, the GSAT verbal score was the only GSAT score that showed significant correlation with academic performance. These results point to the undeniable importance of language proficiency for academic performance, in particular for those who do not receive education in their mother tongue.

7.4 INTEGRATED SUMMARY OF LPCAT VALIDITY RESULTS FOR GROUPS 1 TO 5

In this section, the results for Groups 1 to 5 are integrated and summarised under the same headings used for the presentation of the results. In an attempt to provide a broad overview of the validity results of the LPCAT, the results of the LPCAT-1 (Groups 1 to 3) and those for the LPCAT-2 (Groups 4 and 5) are combined. Since the two versions of the test use exactly the same practice examples, and items are selected from the exact same pretest and post-test item banks, the results for these two versions should be very similar. The only difference between the two versions is the initial entry level and the method of working through the practice examples and training section - either reading it independently from the screen for the LPCAT-1, or being provided with the verbal instructions at each screen for LPCAT-2.

7.4.1 Comparison of mean scores

Comparison of the mean scores of the language groups did not follow a particular pattern. However, in general, differences between the two groups on the LPCAT were similar to or smaller than those on the nonverbal scores of standard cognitive tests, and significantly smaller than those on the verbal scores of the standard cognitive tests. Generally, the LPCAT difference scores for the two language groups did not differ significantly. This indicates that, although the groups may differ in terms of their level of performance, no evidence of differences in their ability to improve upon

their present level of performance could be found (see Figure 7.35 for an example).

7.4.2 Distribution of scores

In general, the distributions of scores for Groups 1 to 5 indicate that performance on the LPCAT covers all ability levels, thus showing that the test can distinguish between individuals at various levels of general reasoning ability. Furthermore, although differences between the two language groups can also be seen in the distributions of LPCAT scores, these differences are generally smaller than those for academic performance and standard cognitive test results. The LPCAT (PPC) learning potenital scores therefore seems to provide a somewhat more equitable measure of general ability than the other measures.

7.4.3 Correlations with other cognitive tests

There is a consistent pattern of numerically high and statistically highly significant correlations between the LPCAT (PPC) scores and other cognitive instruments, such as the GSAT and PPG. This means that the LPCAT measures a construct similar to that measured by other tests of general cognitive ability, while specifically focusing on measurement of learning potential within the domain of general nonverbal reasoning. In general, the correlations of the LPCAT (PPC) scores with the nonverbal results of standard cognitive tests are higher than those with the verbal scores. The LPCAT post-test and composite score correlations with the standard test results are generally higher than those of the LPCAT pretest with the standard test results. This provides support for the use of learning potential measures that include both present (pretest) level of performance as well as the effect of training (difference score).

7.4.4 Correlations with criterion scores

For the two Technikon groups, correlations between the LPCAT (PPC) scores and

academic performance were generally low. The following are considered to be important factors contributing to these results:

preselection and the consequent restriction in range of ability
problems in measuring the average first-year performance, since not all students take the same combination of subjects.
strong indications that student's lack of proficiency in the language of training affects their academic performance

The LPCAT (PPC) scores showed much higher correlations with school academic results. This is probably partly because of less restriction of range on both the predictor and criterion scores, as well as the fact that the reliability of the average academic results as a criterion measure is much higher at school level than at the tertiary level, because the school groups of pupils generally take the same subjects.

Even for standard cognitive tests, the content of which (numbers and verbal material) overlaps more with academic material than the figural problems of the LPCAT, very low correlations with first-year academic results have been reported (Huysamen, 1998) - between 0,19 and 0,27 for verbal and between 0,12 and 0,13 for nonverbal subtests. The correlations for the present study of Technikon groups are similarly low.

Once again, the post-test and composite score learning potential measures which include present performance as well as the effect of learning (ie, pretest and difference scores), show higher correlations with the academic criteria than the pretest score.

7.4.5 Regression analysis and comparison of regression lines

The main purpose of the LPCAT is not to predict academic performance. Owing to its nonverbal figural content, which is related to basic reasoning skills, but not related to academic content, some correlation can nevertheless be expected in so far as both the LPCAT and academic performance rely on basic reasoning skills. The regression analyses were performed mainly to compare the total group and the different

subgroups with regard to the relation between LPCAT and criterion performance.

The results indicate that there are often differences between the regression lines of the different groups, and that the regression line for the African language group is often noticeably different from the others. This emphasises the problems involved in trying to predict *academic* results on the basis of general nonverbal figural reasoning ability. Standard cognitive tests have more in common with academic material and can therefore be expected to have higher correlations with academic performance. However, performance on standard cognitive tests necessarily relies to a large extent on prior learning experiences. They are therefore more likely to entrench present disparities, based on unequal prior learning opportunities. The main purpose of the learning potential measures of the LPCAT is to identify present and potential future levels of general reasoning performance, thereby indicating developmental possibilities. Its inclusion in assessment or selection batteries, together with standard cognitive measures, is likely to provide useful *additional* information.

7.5 ADDITIONAL EVIDENCE FOR THE VALIDITY OF THE LPCAT

In terms of LPCAT scores, learning potential is defined as a combination of the pretest score and the difference score. This learning potential score represents a composite overall estimate that takes both the present level of performance as well as the potential for future improvement in performance into account. The focus of the present chapter so far, has therefore been on the pretest, post-test and composite scores of the LPCAT. The difference score *on its own* is generally not considered to be a suitable measure of learning potential, primarily because an attempt has been made to measure learning potential over the entire range of ability levels. Nevertheless, some information about the difference score is of interest.

The first issue addressed below, is whether the difference scores reported for the groups included in the present research are significantly greater than zero - thereby demonstrating a real and significant difference between the LPCAT pretest and post-test scores.

A second issue, which follows once the significance of the difference scores has been verified, is whether the training provided in the LPCAT can account for the difference found between the pretest and post-test scores and that these differences are not simply the result of practice during the pretest.

A third issue that will be addressed, is the investigation of specific validity information for the LPCAT difference score - that is, whether the difference scores on the LPCAT correlate with other external measures that purport to measure "learning" or improvement in performance.

Finally, the developmental changes reflected by the LPCAT scores of the different groups are investigated.

7.5.1 The significance of LPCAT difference scores

The first issue that needs to be investigated is whether there is a significant difference score - in other words, whether there is significant improvement in the post-test (after training) compared with the pretest level of performance. Since only the very top level of performance precludes improvement, this is a logical assumption to check. The distribution of difference scores over the initial pretest level of performance is indicated in Figure 7.35. The scatter diagram indicates that various sizes of difference scores are found over the entire ability level range, with slightly smaller difference scores at the higher initial levels of performance, as expected.

FIGURE 7.35 SCATTER DIAGRAM OF LPCAT PRETEST SCORES AND LPCAT DIFFERENCE SCORES FOR THE GRADE 8 GROUP PER LANGUAGE GROUP

To investigate the hypothesis that, in general, the difference score is significantly greater than zero, the mean difference scores of the groups involved in the LPCAT validation are compared to zero by means of single-sample t-tests. The results are reported in Table 7.26.

TABLE 7.26 THE SIGNIFICANCE OF THE DIFFERENCE SCORES FOR THE TOTAL GROUPS AS WELL AS FOR THE LANGUAGE SUBGROUPS

Group	N	Mean	t-value	p-value
	TOTAL GROUP			
1	92	1,3696	4,021	,000**
2	159	1,2579	4,866	,000**
3	37	0,1622	0,340	,736
4	194	1,5722	5,440	,000**
5	128	2,1563	7,173	,000**
	AFRICAN LANGUAGE GROUP			

1	46	1,3043	2,423	,019*
2	69	1,2754	3,247	,002**
5	44	1,9091	3,738	,001**
		ENGLISH/AFRIKAANS GROUP		
1	46	1,4348	3,388	,001**
2	86	1,2907	3,667	,000**
5	82	2,2561	5,911	,000**

** p < ,01 * p < ,05

The results in Table 7.26 indicate that the difference scores for Groups 1, 2, 4 and 5 were all significantly larger than zero. This indicates that the differences found are not due to chance alone. Group 3 was the only group for which the difference score was not significantly larger than 0. Two factors that may contribute to this latter result are the small size of this particular sample and the fact that testing of this group took place during examination time, which may have affected their motivation and/or concentration. Since for most groups the difference score was found to be significantly greater than zero, the next step involves investigating of the reasons for these significant differences.

7.5.2 LPCAT difference scores and the training provided

The next issue that was investigated, was the specific effect of the training provided in the LPCAT to bring about the improvement in performance. In order to investigate the effect of training, a group (N=109) of grade 9 high school pupils in an urban high school was used. This group, which will be called Group 6, represented the entire grade 9 class. Pupils were randomly assigned to three groups, after which three different procedures were randomly allocated to the three groups. The three procedures were as follows:

The LPCAT-1 in its standard form was administered to the first group (N=37). This same group was described earlier as Group 3.

For the second group (N=35), the LPCAT-1 was administered, and additional training over and above the LPCAT training provided. This additional training was given at the start of the standard LPCAT-1 training session and involved working through 18 additional examples, providing more explanations and illustrations of the principles involved in solving LPCAT questions.

The LPCAT-1 pretest and post-test were administered to the third group (N=37) without any training at all between the two tests.

The mean difference score for the group that did the LPCAT pretest and post-test without any training (ie only practice effect) was 0,0541. For the group that did the standard LPCAT-1 pretest-train-post-test, the mean difference score was 0,1622, while the mean difference score for the group that did the additional training was 1,7429. Although the mean difference scores increase in the expected direction with the smallest difference score for the group that did no training and the largest difference for the group that did the additional training, the difference score for the subgroup that completed the standard LPCAT-1 was much smaller than those found for Groups 1, 2, 4 and 5 (see previous sections), and was found not to be significantly greater than zero. The mean difference score of the subgroup of Group 6 which received additional training (N=35) is much more in line with the typical difference scores found for the other groups and was also found to be significantly greater than zero.

This may indicate that the limited amount of training provided in the LPCAT does not necessarily reflect the examinee's full potential to improve on current performance, and that further training may lead to an additional improvement in performance.

7.5.3 Correlations of LPCAT difference scores with other measures

It was shown earlier that, while the LPCAT composite score (as well as the pretest and post-test scores) correlates highly significantly with school academic performance, the difference score showed virtually no correlation with academic performance. In Table

7.27, the correlation coefficients between LPCAT difference scores and various other measures are set out. For Group 5, the English and Mathematics proficiency tests were administered for the first time early in February and a second time during May 1999, three months after the first administration. It was decided to use the differences in performance on the proficiency tests between May and February as the criterion for the LPCAT difference score. These correlations are provided in Table 7.27.

Two scores were used for the English and Mathematics teacher rating scales. The total score is made up of the sum of the ratings on all 13 questions, while the "learning" score is made up of the sum of the ratings on specific questions that are focused on the learning attitude, adjustment and improvement in the pupil's performance (questions 4, 5, 8, 9, 11, 12 and 13) - see Appendix C for a copy of these rating questionnaires.

None of these scores correlate significantly with the LPCAT difference score. These results seem to provide support for the view that the difference score alone is not a suitable measure of learning potential.

TABLE 7.27 GROUP 5: CORRELATIONS OF LPCAT DIFFERENCE SCORES WITH OTHER NONACADEMIC MEASURES

	N	LPCAT difference	
		r	p
LPQ surface approach	109	-,001	,933
LPQ deep approach	109	-,160	,096
LPQ achieving approach	109	-,080	,409
Mathematics proficiency difference score	106	,030	,762

English proficiency difference score	103	,090	,365
English teacher rating total score	119	,033	,719
English teacher rating learning score	119	,028	,766
Mathematics teacher rating total score	119	-,067	,470
Mathematics teacher rating learning score	119	-,017	,851

* p < ,05 ** p < ,01

7.5.4 Developmental changes

To investigate whether the LPCAT differentiates successfully between people at different developmental levels, the mean LPCAT scores of Groups 1 to 5 are compared. These represent a wide range of education and ability levels. The means and standard deviations of the LPCAT scores for the groups are presented in Table 7.28.

The results in Table 7.28 indicate that the LPCAT pretest, post-test and composite scores do reflect developmental (educational) changes. The mean scores obtained by the groups increase with educational level.

TABLE 7.28 DESCRIPTIVE STATISTICS FOR LPCAT SCORES FOR THE DIFFERENT GROUPS

GROUP		LPCAT pretest	LPCAT post-test	LPCAT composite score	LPCAT difference score
Group 4 - adult learners	Mean	36,19	37,76	36,64	1,57

(N=194)	(SD)	(7,94)	(9,00)	(7,97)	(4,03)
(average education grade 8)					
Group 5 - grade 8 pupils	Mean	45,67	47,83	46,03	2,16
(N=128)	(SD)	(8,38)	(7,50)	(8,25)	(3,40)
(start of grade 8 school year)					
Group 3 - grade 9 pupils	Mean	49,65	49,81	49,79	0,1622
(N=37)	(SD)	6,79)	(6,48)	(6,76)	(2,9013)
(end of grade 9 school year)					
Group 1 - first year Technikon	Mean	57,78	59,15	58,28	1,37
(N=92)	(SD)	(6,27)	(5,22)	(5,99)	(3,27)
(testing in March)					
Group 2 - first-year Technikon	Mean	55,21	56,47	55,59	1,26
(N=159)	(SD)	(6,13)	(5,39)	(5,99)	(3,26)
(testing in August)					

This chapter has provided results of validity investigations involving five different groups and a variety of psychometric and academic criterion measures. These results will be discussed in the next chapter, with a view to integrating the results and providing an overview of the present project.

CHAPTER 8

DISCUSSION AND RECOMMENDATIONS

8.1 INTRODUCTION

The culture-fair measurement of cognitive ability has been a contentious issue for many years. In South Africa, political and social changes in recent times have brought new opportunities and challenges in many spheres - also in the field of psychometric testing and specifically cognitive assessment. The need for measures that can take the diversity of our population into account and that also make provision for differences in educational and socioeconomic background has been emphasised by researchers, the profession, and legislation (Claassen, 1997; Employment Equity Act, 1998; Foxcroft, 1997; Owen, 1998; Shuttleworth-Jordan, 1996; South African Professional Board of Psychology, 1998). The urgent need for the design and development of instruments for cross-cultural use and for which empirical studies are undertaken to investigate test bias and cultural appropriateness, is clear. Such instruments need to make allowance for the diversity of the population in terms of educational and socioeconomic background. At the same time they also need to ensure that the scarce resources available for training and development are utilised in such a way that opportunities can be provided to those who have been most disadvantaged while maintaining standards and success rates of training and development opportunities. With regard to practical considerations, ease of administration and test administration time should be optimised.

Recent developments in cognitive ability testing which allow for training within test administration to make possible the measurement of learning potential, seem particularly suitable for multicultural testing of cognitive ability. Such dynamic tests use a test-train-test strategy, which allows, in a manner of speaking, for the levelling of the playing field for people from diverse backgrounds to bring about more equitable testing. A test that includes some training, benefits examinees by allowing them to improve on their initial performance. Dynamic testing has increasingly received attention in international and national research. The basic idea that learning potential

tests should have the psychometric properties of standard tests, but that their administration procedure should differ in that a training phase is incorporated and improvement in performance is monitored, highlights the main characteristics pursued in dynamic testing. According to Grigorenko and Sternberg (1998, p 76) “notwithstanding the importance of the endeavour and of allocating significant resources to its realization, multiple attempts to quantify learning potential and to transform such testing into robust psychological diagnostic tools have not produced consistent results”. They claim that the target of prediction of dynamic testing is not always clear and that problems relating to the lack of standardisation of many of the procedures used have complicated its evaluation. Difficulties have been encountered with the quantification of learning potential and there has generally been very little published material on the reliability and validity of dynamic testing. According to Grigorenko and Sternberg (1998) there is a lack of evidence that learning potential assessment procedures contribute more to the prediction of school success than measures of nonverbal IQ. It needs to be shown that the dynamic testing approach proves its usefulness and shows distinct advantages over traditional static tests - relative to the resources that need to be expended.

The main aim of the present project was to construct, standardise and evaluate the LPCAT - a dynamic computerised adaptive test for the measurement of learning potential - for use in multicultural groups for the assessment of cognitive development/ability. Other aims of the project were to develop norms to allow for comparison between obtained and normative scores, to validate the LPCAT dynamic testing results against educational criteria and to show that results can be replicated. One of the most serious criticisms that has been levelled at the dynamic testing paradigm is that it lacks a sound psychometric foundation - particularly with regard to the measurement of change between pretest and post-test. The basic premise of the present project is the psychometric view of dynamic testing, in which greater emphasis is placed on standardisation of procedures and psychometric properties of the test. By means of IRT and CAT, many of the criticisms against dynamic testing can be addressed.

8.2 MEASUREMENT OF INTELLIGENCE

The first general test of intelligence was developed by Binet in the early 1900's. Many cognitive tests used today still resemble this test and Binet's ideas have continued to be of importance in the development of psychometric tests of cognitive ability. Not only has his basic test form stood the test of time, but some of his ideas - notably that of adaptive testing - were so far ahead of his time that it has only been possible to fully develop them with the advent of computer technology and modern IRT developments. Vygotsky's theory of the ZPD, which allows for improved performance following relevant training, thereby improving measurement of cognitive development, forms the theoretical base for the present project. His theoretical stance is in agreement with the views of Binet, who proposed the use of "mental orthopaedics" to improve and strengthen mental ability in the same way that physical exercises can improve physical strength. Both Binet's and Vygotsky's views subscribe to the modifiable view of intelligence. In the dynamic learning potential test developed for the present project, the aim is to identify the extent to which an individual can improve upon present level of performance when appropriate training is provided. Both present level of performance as well as the improvement indicated are used for the interpretation of the individual's level of development and learning potential. The test is aimed at addressing the need for effective cross-cultural testing of cognitive ability, by means of a culture-fair instrument for the measurement of learning potential that can allow for differences in socioeconomic and educational background of examinees in multicultural groups. To this end, the focus is on the general fluid ability domain, which is considered most culture-fair.

8.3 MEASUREMENT OF LEARNING POTENTIAL

Growing dissatisfaction with standard tests of cognitive ability in multicultural contexts or where individuals differ significantly in socioeconomic status, have led to the development of dynamic tests. Dynamic tests are aimed at providing learning experiences as part of the assessment, in order to evaluate to what extent the individual is able to improve upon present performance when relevant training is provided. The dynamic assessment approach is generally described as an innovative new direction in the measurement of intelligence, and although still

considered to be in its infancy, is receiving widespread attention in research. This approach provides a practical solution to measurement of cognitive ability in increasingly diverse test groups. The large educational and socioeconomic differences between cultural groups in South Africa with the African group in particular being distinctly disadvantaged - as shown by census and survey information - underscores the need for measures that can make allowance for the effect of these differences on test performance. Vygotsky's concept of the ZPD forms a natural theoretical base for the dynamic assessment of learning potential and is operationalised into three distinct measures, namely a pretest score, a post-test score and the difference between them. Although it is recommended that all three scores should be used for interpretation, as a practical measure, a combination of the pretest and difference scores can be used where a single measure of learning potential is required. The present project follows the psychometric approach to the measurement of learning potential with a focus on standardisation of procedures and accuracy of measurement. Key problem areas such as the difficulty experienced with accurate measurement of change, the extended times involved in administering dynamic tests and the limited psychometric properties of these tests have been highlighted by many researchers. Most of these problems have been addressed in the IRT and CAT framework and these methods were consequently used to construct the LPCAT, using only nonverbal figural item content in a further attempt to make the test culture-fair. Through the IRT and CAT procedures, measurement of difference scores is accurate, the testing time is comparable to that of standard cognitive tests and by using separate item banks for the pretest and post-test, a more accurate estimation of improvement of level of performance can be obtained.

8.4 IRT AND CAT

Item response theory represents new rules of psychometric measurement which improve effectiveness and, in particular with CAT, efficiency of testing. The three-parameter model, which is generally considered most appropriate for multiple-choice items, was used for the present project. It incorporates three parameters, namely the b-parameter (difficulty index), the a-parameter (index of

discrimination) and the c-parameter (pseudo-chance index) for each item. The central feature of IRT is the specification of a mathematical function relating the probability of an examinee's response on a test item to an underlying ability. This function is visually depicted by means of an item characteristic curve, which represents the item characteristics of the particular item. For the three-parameter model, large samples - in excess of 1 000 - are required to obtain stable and accurate estimation of the item parameters. For the present project, sample sizes complied with this requirement. IRT is particularly suited for dynamic assessment since it allows more accurate measurement of the difference score because the scores of the pretest and the post-test are on the same scale. One of the most powerful and exciting applications of IRT is CAT, which allows for interactive selection of suitable items during testing to match the estimated ability level of the examinee, thereby drastically reducing testing time without forfeiting measurement accuracy. No previous application of IRT-based CAT procedures for learning potential assessment was found in the literature, making this feature of the LPCAT a unique contribution of the present project. One of the major concerns regarding dynamic assessment has been the extended time needed for test administration. CAT makes possible dynamic tests that are comparable in testing time to standard tests, thereby improving its utility at a practical level.

8.5 CONSTRUCTION OF THE LPCAT

The construction of a dynamic computerised adaptive test for the measurement of learning potential, namely the LPCAT, forms the core of the present project, together with empirical investigation of its validity. The development and validation of the LPCAT took seven years from initial conceptualisation to completion. Its purpose was to provide a psychometric instrument that could provide useful information in South Africa's multicultural context by measuring learning potential in the nonverbal general reasoning domain. The following features were specifically built into the LPCAT:

- It uses only figural nonverbal items, thereby eliminating the effect of language proficiency or other school-related prior learning on test performance. The item types used are figure series, figure analogies and pattern completion.

- It makes allowance for examinees from disadvantaged backgrounds by providing a learning experience in the test, thereby allowing them to indicate at what level they may be able to perform, should better learning opportunities be provided. Both their initial level of performance and their improvement are used for interpreting overall level of cognitive development.
- Standard training is used to improve the comparability of test scores of individuals.
- IRT scoring allows for improved measurement of the difference score between pretest and post-test performance, thereby ensuring improved psychometric characteristics of the LPCAT.
- CAT is efficient, since items are selected from a precalibrated item bank during the testing session to continually match the examinee's estimated ability level. This saves administration time without forfeiting quality or accuracy of measurement.
- Multicultural groups were used for item analysis, standardisation and validation of the test to provide the required support for its psychometric properties and evidence in support of its validity for multicultural cognitive assessment.

In the construction of the LPCAT, ease of administration and standardisation of procedures were given priority. The test takes approximately one hour to administer and comprises five sections namely:

(1) Introduction and providing of information regarding the keys to be used and the general answering procedure.
(2) First set of examples before commencing with the pretest
(3) Pretest
(4) Training and additional examples
(5) Post-test

Both CTT and IRT procedures were used for item analysis. The sample used for item analysis was multicultural with adequate size to allow three-parameter IRT analysis. The three main IRT assumptions, namely one-dimensionality, item parameter invariance and ability parameter invariance, were empirically investigated. These results provide support for the use of the three-parameter IRT model for the LPCAT.

Subsequent to initial item analysis, extensive DIF analysis was performed on all the items by comparing the ICCs of selected contrast groups based on culture, language, gender and educational level respectively, and calculating the area between the two ICCs each time. Items that exceeded a certain cutoff value of DIF were discarded.

Of the initial item bank of 270 items - 90 items of each of the three item types - 47 items were discarded on the basis of IRT and CTT item analysis and a further 35 items were discarded on the basis of DIF. The remaining 188 items (65 of figure series, 58 of figure analogies and 65 of pattern completion) were allocated to the pretest and the post-test in a sequential 1:2 ratio per item type and in sequential order of difficulty, to ensure an even spread of item types and item difficulties in the pretest and the post-test respectively. This ratio was used since for more accurate measurement of performance in the post-test, more items at each difficulty level are required.

The mean difficulty of the items as well as the distribution of item difficulties in the pretest and the post-test respectively show that the items are reasonably easy for someone at approximately grade 10 level. This is in accordance with the original aim of the test as a measure of learning potential that is specifically aimed at people from disadvantaged backgrounds, who generally have lower levels of education. Nevertheless, sufficient numbers of very difficult and very easy items together with the adaptive testing process, allows for the measurement of learning potential at all ability levels.

For IRT-based CATs, test reliability is evaluated by means of a test information function, which graphically depicts the amount of information - and the resulting accuracy of measurement at that particular level - over the entire ability range. Both the pretest and post-test of the LPCAT show sufficiently high test information functions to allow for accurate and reliable measures over a wide range of ability.
In order to increase the utility of the LPCAT, a second version, named the LPCAT-2 was developed for those people who are not sufficiently proficient in English or Afrikaans to read the test instructions, explanations and feedback independently from the screen. For this version, the exact same practice examples and item banks were used as those used for the initial version (the LPCAT-1). The only difference

between the two versions is that in the LPCAT-1, all instructions, feedback and explanations appear on the screen and are read independently by the examinee in either English or Afrikaans, the particular language being chosen by the examinees themselves. In the LPCAT-2, only figures appear on the screen and the test instructions, examples and feedback are read to the examinee in any one of the 11 official South African languages.

Another difference between the LPCAT-1 and LPCAT-2 is that the initial entry point level of difficulty of the first item administered in the LPCAT-2 is slightly lower than that of the LPCAT-1. However, due to the adaptive testing procedure and the fact that the items for both test versions are selected from the exact same item banks, this does not place any limitation on the level of performance that can be attained with either of the two versions. The fact that CAT procedures allow for a variable entry point in testing to be specified, was further used by taking the exit level of performance in the pretest and making it the entry level of performance (ie estimated ability level) in the post-test. This further improves accuracy of measurement and streamlining of the adaptive testing procedure.

Termination of the pretest and the post-test is based on a minimum (predetermined) number of items having been administered and, either a maximum (predetermined) number of items being reached, or a (predetermined) level of accuracy of measurement of the estimated ability level being attained. This feature in particular improves equitable accuracy of measurement at all ability levels. Final scores for the LPCAT are given as T-scores, which, with a mean of 50 and standard deviation of 10, are comparable to percentage scores, thereby simplifying user interpretation. Because the measurement of learning potential involves multiple scores, the context may determine where the focus should be placed for interpretation of the results. Depending on the context, more focus may be placed on any one of the four possible scores, namely the pretest score, post-test score, difference score or an overall composite score.

Computerised adaptive testing based on IRT can address most of the problems that have been identified regarding dynamic testing. It not only provides improved accuracy of measurement, but also brings about much shorter testing times, making

dynamic testing comparable to standard psychometric tests in terms of effort needed to administer, while they provide very useful additional information that can be used for more culture-fair assessment of ability in multicultural contexts.

The construction of the LPCAT, its characteristics, and instructions for its use and interpretation of its scores are fully described in two comprehensive test manuals (De Beer, 2000a, 2000b, in press).

8.6 PROCEDURE FOR EVALUATING THE VALIDITY OF THE LPCAT

As stated earlier, multicultural groups were used for the investigation of the validity of the LPCAT. These groups were also selected to represent varying levels of educational attainment, to investigate the utility of learning potential measures at different educational levels.

Face validity and content validity were evaluated for the LPCAT as a whole, that is, for both versions of the LPCAT simultaneously, since the two versions use exactly the same practice examples and item banks from which items are selected during test administration. Using only universally known figures and concepts, not related to either language proficiency or prior educational content, provides support for the face validity of the LPCAT as a general culture-fair measure of learning potential. The easy answering procedure, together with the dynamic administration which allows for training within the test, further improves the face validity of the LPCAT. Lastly, having the LPCAT-2 available, with which a person can receive test instructions in any of the 11 official South African languages, also contributes to the face validity of the LPCAT. The item types used are similar to those used in widely accepted culture-fair tests such as the Raven's Progressive Matrices and Cattell's Culture-Fair Intelligence Test. This provides support for content validity of the LPCAT as a culture-fair measure of learning potential. Having had the items evaluated by an expert panel who approved their use for assessing general nonverbal reasoning ability, further provides support for the content validity of the LPCAT. In terms of the factor structure of the LPCAT items, evidence in support of one-dimensionality was found for the total group as well

as for important subgroups. This, together with the high internal consistency measures that were found once again for the total group as well as for important subgroups, further provide support for the content validity of the LPCAT.

For the empirical validity investigation of the LPCAT-1 and LPCAT-2, five different samples were used. For the LPCAT-1, two Technikon first-year samples were used as well as a sample of grade 9 school pupils. For the LPCAT-2, a sample of adult learners and a sample of grade 8 school pupils were used. Most of these groups were multicultural, with only the adult learner group consisting of African home language examinees only. For practical purposes, the multicultural utility of the LPCAT was investigated with subgroups based on home language. The examinees were identified as belonging either to the African home language group or to the English/Afrikaans home language group. Based on the fact that the first group receive most of their education in a language that is not their first language, while for the latter group this is not the case, this distinction was seen as practical. Where possible, results of other psychometric instruments were also obtained as criterion measures. For most of the groups, academic results were used as real-life criterion measures. Testing took place over a four-year period during different times of the year. The effect of different training procedures were also investigated, and groups of different educational levels were compared to investigate the extent to which the LPCAT can differentiate between developmental levels.

8.7 DISCUSSION AND INTERPRETATION OF THE RESULTS

With its exclusively nonverbal figural content, the LPCAT does not directly rely on language proficiency or prior education. This needs to be taken into account when interpreting validity results, since in most cases, the criterion measures used were academic performance. This means that a measure that has purposefully been chosen not to be dependent upon language proficiency or prior learning, is validated against criterion measures saturated with language and prior learning in particular. In so far as LPCAT performance is related to basic reasoning ability, the LPCAT and academic criterion scores can be expected to correlate with each other, but the

limitations of such correlations are clear. Nevertheless, while learning potential assessment is aimed at uncovering undeveloped latent capacity, the overall aim is still to obtain measures that can be used to predict future performance, emphasising the importance of both present level of performance as well as undeveloped capacity.

The results indicate that the LPCAT, as a measure of learning potential, is more culture-fair than standard tests of intelligence. It generally produces smaller differences in mean scores between the cultural groups than standard cognitive tests, as evidenced by the sizes of the proportional differences in terms of standard deviations. In terms of the distributions of scores, the LPCAT seems to provide a more equitable and less positively skewed distribution of scores for the African home language group compared with standard test results and academic results.

Comparison of the LPCAT with different standard cognitive tests indicates that the LPCAT measures a construct similar to that measured by standard tests of intelligence - in particular the nonverbal sections. Although the LPCAT and the standard cognitive tests seem to measure the same general reasoning ("*g*") construct, there are certain advantages to the LPCAT, namely that it uses only nonverbal figural items, that it incorporates learning within test administration and uses efficient CAT procedures based on IRT.

With regard to criterion-related validity, although the LPCAT results do not correlate highly with tertiary academic results, some of these correlations were statistically significant and the LPCAT does seem able to distinguish between examinees at this level. For the Technikon results, the problematic nature of tertiary and grade 12 academic results pointed out by Huysamen (1999) should also be kept in mind. He mentions, inter alia, restriction of range as one possible compounding problem in the issue of accurate prediction of tertiary academic results. Numerically higher and statistically highly significant correlations with academic criteria were found for the school groups where there is a wider range of ability and where the criterion data are more reliable. Generally, where there are indications of a limited range of ability, the LPCAT difference score correlations tend to be higher - sometimes even higher than the LPCAT (PPC) score correlations with the criteria. Where there is a wider range of

ability, the LPCAT (PPC) score correlations with the criterion data (psychometric or academic) are generally much higher, while the LPCAT difference score correlations tend towards zero. At school level, where there is both a wider range of ability levels and where the academic criterion measures are more reliable, this pattern can be seen clearly.

For the Technikon groups, the statistically highly significant correlation between grade 12 English and average first-year academic performance for the African language group, should be noted. For the English/Afrikaans home language group this correlation is very low. This shows that for the African language group in particular, language proficiency strongly affects academic results.

At school level, the LPCAT (PPC) scores generally correlate highly significantly with academic performance for the English\Afrikaans group, but not for the African language group. While the English- and Afrikaans-speaking pupils are taught in their home language, the African home language group receive tuition in a second (or third) language. The fact that for the grade 8 African language group, English proficiency had the highest correlation with academic performance, is again an indication that language proficiency acts as a filter for those who are not proficient, preventing them from realising their general reasoning ability into commensurate academic performance levels. It seems that when language proficiency is adequate, academic performance is generally commensurate with reasoning ability. This can explain the higher correlations between LPCAT (PPC) scores and academic performance for the English/Afrikaans language group. However, when lack of language proficiency or poor prior learning experiences act as a hurdle, those who are not proficient in the language of teaching or who do not have adequate prior learning experiences, are probably prevented by these factors from performing academically on a level commensurate with their reasoning ability. This can explain the lower correlations between LPCAT(PPC) scores and academic results for the African language group.

Differences in the performance of the two language groups seem in particular to be related to language proficiency and prior learning experiences. This can be seen by the large differences on measures that are language-dependent (ie standard verbal

tests or academic test performance) compared with smaller differences on the LPCAT, which is not as reliant on either language proficiency or prior learning. For the African language group, measures of language proficiency and educational performance provide higher correlations with academic performance than the LPCAT (PPC) scores but LPCAT correlations with academic performance are generally higher than those of the nonverbal parts of standard cognitive tests with academic performance. .

In multicultural groups it is generally considered unfair to use only measures that rely on either language proficiency or on previous educational opportunities to select people for training and development or for placement at training institutions. Although verbal ability and language proficiency generally provide the highest correlations with academic performance and can therefore be regarded as the best predictors of future academic results, they are related to previous educational opportunities and it would therefore not be fair to disadvantaged students if only they were used in selection procedures. The same would be true of previous academic performance, although it also generally correlates very highly and highly significantly with subsequent academic performance. However, where the aim is to provide training opportunities to those who may not have had the opportunities to develop to their full potential, use of *only* such measures would constitute an unfair practice. Since practical and financial considerations make it imperative that examinees with a real chance of success in the training that is to be provided should be selected, accurate prediction remains important. To obtain a balance between eventual success in training, while providing opportunities to those who may be disadvantaged at present but who show potential to perform at levels adequate for attaining eventual success, a combination of learning potential results, standard test results and language proficiency results are likely to provide the most useful information for predicting academic success. It is therefore recommended that, in addition to standard measures that have been shown to predict academic/training performance adequately, the LPCAT be used. In this way, individuals who may be hampered by poor language proficiency or poor educational history, will be able to show their level of reasoning ability. Investment in the form of either language proficiency training and/or basic educational skills training can be provided to those who show the required level(s) of reasoning ability, but who

may not presently meet all criteria for “success” in a particular training or educational environment. This would pave the way for the more impartial and equitable creation of opportunities for training and development.

- In cases where the focus is on affirmative action, learning potential measures can carry more weight. Where academic performance will be required or where academic training will be involved, additional language proficiency training is likely to improve the probability of success.
- In cases where limited funding or limited time for investment in language proficiency training is available, more emphasis may be placed on prior academic performance, existing language proficiency and/or performance on standard cognitive tests.

Large disparities between performance on the LPCAT and academic performance are indicative of persons most likely to benefit from additional training. When efforts are made to improve the language proficiency of such individuals, improved academic performance should follow - in particular for those individuals who already indicate the required level of general reasoning ability.

Grigorenko and Sternberg (1998) emphasised the need for conducting learning potential studies that involve larger participant populations where results are validated against educational criteria. They also noted the need for replication of results. These issues have been addressed in the present project. The LPCAT makes provision for people form a wide range of ability levels and the results indicate satisfactory predictive validity for predicting school academic achievement. In terms of replication, similar results were found for different sample groups - both at tertiary and junior high school levels. The LPCAT results in terms of developmental changes indicate that it can distinguish between examinees at various educational levels.

The two main contributions of the present project are as follows:

- The first contribution is the extended definition of learning potential as a combination of the present level of performance and improvement in performance, thereby allowing measurement of learning potential over the

wider ability spectrum. It allows comparison of individuals at different present ability levels and with varying improvements following training. This definition is based on Vygotsky's theoretical principles, but extends Vygotsky's special case to the broader ability range. The LPCAT definition of learning potential emphasises that the difference score alone is not an appropriate measure of learning potential when examinees have different ability levels. A combination of present level of performance (pretest score) as well as the improvement following training (difference score) is required to provide a useful measure of learning potential over the broader range of ability levels. In using multicultural samples for test construction and validation and by including only nonverbal figural test items, the general professional and legal requirements for psychological tests in South Africa were heeded, also taking into account the effect of language proficiency on test performance. It is furthermore recognised that learning potential measures need to specify the domain within which learning potential is assessed, based on the content of the tests used and the training provided.

- The second contribution lies in the use of IRT procedures not only for test development, but also for test administration in the form of two separate CATs for the pretest and the post-test. The training is standardised, computer-administered yet interactive, ensuring better comparability of results. Use of IRT and CAT ensures that the change measured is a direct manifestation of improved performance. Performance is not affected by memory because different test items are used, and measurement accuracy is improved by means of IRT-based comparison of pretest and post-test scores on the same scale. These features improve not only the psychometric characteristics of the LPCAT dynamic test, but also the test efficiency, making it comparable to standard cognitive tests in terms of standardisation as well as the ease with which it is administered.

The utility of the LPCAT can be evaluated in the following three areas:

(1) **The requirements of the South African Professional Board of Psychology**

With the development, standardisation and evaluation of the LPCAT, the call

of the South African Professional Board of Psychology for tests that take the diversity of the South African population into account has been heeded in that it is a test that has been designed and standardised for all South Africans. It has included empirical studies to investigate DIF and to assess its validity and cultural applicability.

(2) **Legislation on psychological testing**

By means of modern methods and procedures in test development and test administration, scientific support has been provided for the reliability and validity of the LPCAT. It is culture-fair in its content and incorporates training in the test administration, thereby allowing for differences with which people come to the testing situation, and can therefore be applied fairly to all employees. Lastly, having discarded items that indicated more than a certain amount of DIF, and using only nonverbal figural items, the LPCAT is not biased against any employee or group. These features which all point to the sound psychometric qualities of a test, were specifically addressed in the development of the LPCAT, thereby ensuring its compliance with the requirements of the Employment Equity Act of 1998.

(3) **Psychometric requirements for psychological tests**

In the development of the LPCAT IRT procedures were used together with extensive DIF analysis, to ensure the psychometric soundness of the test and accuracy of measures obtained. The inclusion of IRT-based CAT procedures further improves the testing efficiency and standard computerised training ensures comparability of test results.

The initial aims of the project that have been met are:

- DIF was investigated between language and culture groups as well as between the gender groups.
- The reliability and validity of the LPCAT was investigated.
- The predictive validity of the LPCAT for academic and other relevant results was investigated.
- Results of the LPCAT were compared to those of conventional cognitive test

results.

- The utility of the LPCAT for cross-cultural measurement was investigated.

The following specific outcomes resulted from these investigations:

- A dynamic instrument for the measurement of learning potential in the general nonverbal reasoning domain was constructed. This test is based on IRT principles and uses CAT administration procedures, thereby improving its psychometric properties. CAT computerised administration allows for interactive feedback to examinees during practice examples and the training session, while during test administration, items are selected to match the examinee's estimated ability level. The standardisation of test administration improves the comparability of test results.
- The LPCAT is based on Vygotsky's theory which provides a solid theoretical base for its development and its particular definition of learning potential over the wider range of ability levels.
- Recognising the effect of language proficiency on test performance, only nonverbal figural item content was used, avoiding material related to either language proficiency or prior education. Provision for people of lower educational levels was further made by constructing a second version of the LPCAT (LPCAT-2), in which all test instructions are read in the language best understood by the examinee and where no reading is therefore required by the examinee.
- Multicultural samples were used in its development and evaluation, and extensive DIF analysis was performed. Supporting evidence for use of the LPCAT in multicultural groups has been provided.
- The reliability of the LPCAT was investigated by means of the IRT-based test information function. Reliability of scores is enhanced by including the accuracy of ability estimation as one of the criteria for adaptive test termination in both the pretest and the post-test.
- Validity investigations for the LPCAT were performed in different contexts and at different educational levels, providing support for its construct validity for different levels of ability.

8.8 CRITICAL EVALUATION AND RECOMMENDATIONS

While the present project has made certain contributions to the field of dynamic assessment, there are aspects that can be improved upon. These may be addressed by future research.

In the present project, only three of the 10 provinces in South Africa were involved in the item analysis test administration. Although the sample sizes were adequate for three-parameter IRT item analysis purposes and while there is no reason to believe that the pupils from these provinces are any different in terms of their ability levels from pupils in other provinces, a more representative sample - both in terms of provincial representation and cultural representation - would add to the solidity of psychometric evidence for the utility of the LPCAT.

A more thorough investigation of the effect of training in LPCAT results, using larger and more representative samples, is also recommended. The samples used to investigate this particular aspect for the LPCAT, were small and testing took place during the school examination period, which could have affected the findings.
It is furthermore recommended that the LPCAT results should also be validated using practical training criteria or a combination of academic and practical training results. The use of academic criterion measures only does not address the need for measures to also identify people for practical training.

The effect of language proficiency in academic performance has been clearly indicated in the results of the present project. An investigation into the effect of providing language proficiency training prior to or concurrent with academic training needs to be investigated. In particular where more than 76 percent of the South African population have an African home language, but where the majority of training is provided in English and Afrikaans, such an investigation is important for the planning of future course content. For people who receive training in a language other than their first language - and in which they are not adequately proficient - language proficiency training may enable them to better receive the full benefit of the

training provided. It is vital to investigate whether the learning potential indicated in a nonverbal, general reasoning domain can be brought to full development in academic performance if the required level of language proficiency is attained.

On the whole, the results provided here indicate support for the psychometric soundness as well as the internal and external validity of the LPCAT. Addressing the issues mentioned above will provide valuable additional information in the field of dynamic assessment.

8.9 CONCLUSION

South Africa has been through tumultuous times and in the immediate future will need to focus on development both at individual and national level. Training and development have been identified as important priorities. Considering the cost of training, successful outcomes are important, which indicates a need for cognitive assessment to select people for such training. However, with large differences in terms of socioeconomic and educational background hampering test performance of disadvantaged individuals, and language proficiency further impacting negatively on their performance, the use of standard cognitive assessment instruments seems problematic - however well they may predict future performance. In the spirit of transformation and development, a need was identified to focus on learning potential, as a broader concept which includes present level of performance, but at the same time takes into consideration potential future level of performance if further training is provided. Learning potential measures are regarded as more equitable, since they take into account the differences in background with which examinees come to the testing situation.

Given the sociopolitical history of South Africa and the consequent differential impact of education and developmental opportunities on disadvantaged groups, differences between the culture groups can be expected to remain in the foreseeable future. However, with living conditions and educational opportunities improving for the disadvantaged, these differences can at the same time be expected to diminish over

time. A measuring instrument like the LPCAT, which makes provision not only for differences between culture groups, but also for ongoing changes within different groups, can provide useful information in the domain of general cognitive reasoning ability and future developmental potential for people of different cultures and at different developmental levels.

The emphasis of researchers such as Binet and Vygotsky on development and provision of opportunities to improve upon present performance levels is particularly relevant in the present South African context. Binet's focus was primarily on development - and not validity. He wanted a measure that could help indicate those who show the potential to be further developed. The operationalisation of such a focus on development in the form of dynamic testing procedures, combined with new test development theory and technology has provided the base for the construction of the LPCAT. Keeping to the original spirit of Binet's test and accommodating Vygotsky's theory by means of dynamic assessment, a psychometrically sound learning potential instrument in the domain of general nonverbal reasoning and for use over a wide range of ability levels was developed. Addressing some of the concerns about traditional standard cognitive assessment, the LPCAT - by using nonverbal figural content and a dynamic testing procedure - has been shown to be a culture-fair measure of learning potential. The LPCAT takes practical realities into account with a view to identifying people who can benefit from further training and development opportunities. It seems well suited to serve as a screening instrument that can counter inadvertent discrimination against disadvantaged groups, providing a measure of learning potential that is not dependent upon language proficiency or prior school learning and which complies with the legal, professional and psychometric standards for psychological tests.

Although the focus of the present project is on learning potential in the domain of general reasoning ability, this represents only one aspect of human behaviour. At this point it may be apt to refer to the selection requirements set by Cecil John Rhodes for scholars to apply for merit bursaries for their continued education. Rhodes's Will contains four standards by which prospective Rhodes Scholars are judged, namely:

(1) literary and scholastic attainments;

(2) fondness of and success in sports;

(3) truth, courage, devotion to duty, sympathy for and protection of the weak, kindliness, unselfishness and fellowship;

(4) moral force of character and instincts to lead, and to take an interest in one's fellow beings.

(Reference: http://rhodesscholar.org/info.html)

Only one of the four standards is directly related to cognitive ability. The remaining three refer to human characteristics which are found over a wide range of ability levels and are no less important in contributing to one's immediate environment and also to society. Although cognitive ability is important for success in certain fields, care should be taken in general not to overemphasise intellectual performance to the exclusion of or in isolation from other equally important human qualities.

REFERENCE LIST

Aiken, L.R. (1996). *Assessment of intellectual functioning (2nd Ed)*. New York: Plenum Press.

on. (1985). *Standards for educational and psychological testing*. Washington: APA.

Anastasi, A. & Urbina, S. (1997). *Psychological testing (7th Ed.).* New Jersey: Prentice Hall.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3 - 23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Armour-Thomas, E. (1992). Intellectual assessment of children from culturally diverse backgrounds. *School Psychology Review, 21(4)*, 552-565.

Assessment Systems Corporation. (1989). *User's manual for the MicroCAT testing system (Version 3.0)*. St Paul: Assessment Systems Corporation.

Assessment Systems Corporation. (1995). *User's manual for the MicroCAT testing system (Version 3.5).* St Paul: Assessment Systems Corporation.

Babad, E. & Budoff, M. (1974). Sensitivity and validity of learning potential measurement in three levels of ability. *Journal of Educational Psychology, 66*, 439-447.

Baker, F.B. (1985). *The basics of Item Response Theory.* Portsmouth, N.H.: Heinemann.

Bethge, H., Carlson, J.S. & Wiedl, K.H. (1982). The effects of dynamic assessment procedures on Raven Matrices performance, visual search behaviour, test anxiety and test orientation. *Intelligence, 6*, 89-97.

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

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

Biesheuvel, S. (1972a). An examination of Jensen's theory concerning educability, heritability, and population differences. *Psychologica Africana, 14*, 87-94.

Biesheuvel, S. (1972b). Adaptability: Its measurement and determinants. In L.J. Cronbach & P.J.D. Drenth (Eds.), *Mental tests and cultural adaptation*. The Hague: Mouton Publishers.

Biesheuvel, S. & Liddicoat, R. (1959). The effects of cultural factors on intelligence test performance. *Journal of the National Institute of Personnel Research, 8*, 3-14.

Biggs, J. (1987a). *Learning Process Questionnaire Manual*. Melbourne: Australian Council for Educational Research.

Biggs, J. (1987b). *Student approaches to learning and studying - research monograph*. Melbourne: Australian Council for Educational Research.

Binet, A. & Simon, T. (1905/1916). *The intelligence of the feeble-minded*. Baltimore: Williams and Wilkins.

Binet, A. & Simon, T. (1915). *A method of measuring the development of the intelligence of young children*. Chicago: Chicago Medical Book Co.

Blagg, N. (1991). *Can we teach intelligence? A comprehensive evaluation of Feuerstein's instrumental enrichment program*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Blake, R.H. (1972). Industrial application of tests developed for illiterate and semiliterate people. In L.J. Cronbach & P.J.D. Drenth (Eds.), *Mental tests and cultural adaptation* (pp. 64-74). The Hague: Mouton.

Boeyens, J. (1989a). *Learning potential: A theoretical perspective. Report PERS-432,* Human Sciences Research Council. Pretoria: Human Sciences Research Council.

Boeyens, J. (1989b). *Learning potential: an empirical investigation. Report PERS-435*, Human Sciences Researches Council. Pretoria: Human Sciences Research Council.

Boeyens, J.C.A. (1989c). *Learning potential and academic performance*. Unpublished Masters dissertation, University of South Africa.

Brown, A.L. & Campione, J.C. (1986). Academic intelligence and learning potential. In R.J. Sternberg & D.K. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Brown, A.L. & Ferrara, R.A. (1985). Diagnosing zones of proximal development. In J.V. Wertsch (Ed.), *Culture, communication, and cognition: Vygotskyan perspectives* (pp. 273-305). Cambridge: Cambridge University Press.

Brown, A.L. & French, L.A. (1979). The zone of potential development: Implications for intelligence testing in the year 2000. *Intelligence, 3*, 255-273.

Budoff, M. (1967). Learning potential among institutionalized young adult retardates. *American Journal of Mental Deficiency, 72*, 404-411.

Budoff, M. (1969). Learning potential: A supplementary procedure for assessing the ability to reason. *Seminars in Psychiatry, 1(3)*, 278-290.

Budoff, M. (1987a). The validity of learning potential assessment. In C.S. Lidz (Ed.),

Dynamic assessment: An interactional approach to evaluating learning potential (pp. 52-81). New York: The Guilford Press.

Budoff, M. (1987b). Measures for assessing learning potential. In C.S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 173-196). New York: Guilford Press.

Budoff, M. & Corman, L. (1974). Demographic and psychometric factors related to improved performance on the Kohs learning-potential procedure. *American Journal of Mental Deficiency, 78(5)*, 578-585.

Budoff, M. & Harrison, R.H. (1971). Educational tests of the learning-potential hypothesis. *American Journal of Mental Deficiency, 76(2)*, 159-169.

Budoff, M. & Pagell, W. (1968). Learning potential and rigidity in the adolescent mentally retarded. *Journal of Abnormal Psychology, 73(5)*, 479-486.

Campione, J.C. (1989). Assisted assessment: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities, 22*, 151-165.

Campione, J.C. & Brown, A.L. (1987). Linking dynamic assessment with school achievement. In C.S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82-115). New York: The Guilford Press.

Campione, J.C., Brown, A. & Bryant, N. (1985). Individual differences in learning and memory. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 392-490). New York: Cambridge Press.

Campione, J.C., Brown, A.L. & Ferrara, R.A. (1982). Mental retardation and intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 392-492). Cambridge: Cambridge University Press.

Campione, J.C., Brown, A.L., Ferrara, R.A. & Bryant, N.R. (1984). The zone of proximal development: Implications for individual differences in learning. In B. Rogoff & J.V. Wertsch (Eds.), *Children's learning in the 'zone of proximal development'* (pp. 77-91). San Francisco: Jossey-Bass.

Carlson, J.S. (1989). Advances in research on intelligence: The dynamic assessment approach. *The Mental Retardation and Learning Disability Bulletin, 17(1)*, 1-20.

Carlson, J.S. & Wiedl, K.H. (1978). Use of testing-the-limits procedures in the assessment of intellectual capabilities in children with learning difficulties. *American Journal of Mental Deficiency, 82*, 559-564.

Carlson, J.S. & Wiedl, K.H. (1979). Toward a differential testing approach: Testing-the-limits employing the Raven Matrices. *Intelligence, 3*, 323-344.

Carroll, J.B. (1997a). The three-stratum theory of cognitive abilities. In D.P.

Flanagan, J.L. Genshaft & P.L. Harrison (Eds.), *Contemporary Intellectual Assessment*. New York: The Guilford Press.

Carroll, J.B. (1997b). Psychometrics, intelligence and public perception. *Intelligence, 24(1)*, 25-52.

Cattell, R.B. (1950). *Handbook for the individual or group Culture Fair Intelligence Test (A measure of "g") Scale 1*. Illinois: Institute for Personality and Ability Testing.

Cattell, R.B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 18*, 165-244.

Central Statistical Service of South Africa. (1996a). *Living in South Africa: Selected findings of the 1995 October Household Survey*. Pretoria: Central Statistical Service.

Central Statistical Service of South Africa. (1996b). *October Household Survey 1995: Statistical release P0317*. Pretoria: Central Statistical Service.

Central Statistical Service of South Africa (1996c). *Census in Brief - Report number 1:03-01-11 (1996)*. Pretoria: Central Statistical Service.

Central Statistical Service of South Africa. (1998). *The people of South Africa Population Census, 1996: Census in Brief (Report No. 1:03-01-11(1996))*. Pretoria: Statistics South Africa.

Chamberlain, J.C. & Reinecke, S. (1992). *Manual - Proficiency Test English Second Language Intermediate Level*. Pretoria: Human Sciences Research Council.

Claassen, N.C.W. (1983). *Verslag oor die funksionering van die NSAG Intermediêr G in verskillende bevolkingsgroepe*. Pretoria: Human Sciences Research Council.

Claassen, N.C.W. (1990). *Die meting van intelligensie in verskillende groepe met die Algement Skolastiese Aanlegtoets (ASAT)*. Pretoria: Human Sciences Research Council.

Claassen, N.C.W. (1996). *Paper and pencil games (PPG): Manual*. Pretoria: Human Sciences Research Council.

Claassen, N.C.W. (1997). Cultural differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47(4)*, 297-307.

Claassen, N.C.W., De Beer, M., Hugo, H.L.E. & Meyer, H.M. (1991). *Manual for the General Scholastic Aptitude Test (GSAT) Senior Series*. Pretoria: Human Sciences Research Council.

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Coon, H., Carey, G. & Fulker, D.W. (1992). Community influences on cognitive ability. *Intelligence, 16*, 169-188.

Cronbach, L.J. & Drenth, P.J.D. (Eds.) (1972). *Mental tests and cultural adaptation*. The Hague: Mouton.

Cziko, G.A. (1989). Unpredictability and indeterminism in human behaviour: Arguments and implications for educational research. *Educational Researcher, 18(3)*, 17-25.

Dague, P. (1972). Development, application and interpretation of tests for use in French-speaking black Africa and Madagascar. In L.J. Cronbach & P.J.D. Drenth, (Eds.), Mental tests and cultural adaptation (pp. 64-74). The Hague: Mouton.

Das, J.P. (1987). Foreword. In C.S. Lidz (Ed.), *Dynamic Assessment: An interactional approach to Evaluating learning potential* (pp. vii-xi). New York: The Guilford Press.

De Beer, M. (1991). *Rekenarisering van die Algemene Skolastiese Aanlegtoets tot passingstoets* [Computerisation of the General Scholastic Aptitude test as adaptive test]. Unpublished Master's Dissertation, University of South Africa.

De Beer, M. (2000a). *Learning Potential Computerised Adaptive Test (LPCAT): User's Manual*. Pretoria: Unisa (in press)

De Beer, M. (2000b). *Learning Potential Computerised Adaptive Test (LPCAT): Technical Manual*. Pretoria: Unisa (in press)

De Beer, M. & Van Eeden, R. (1997). *Selection criteria for students in engineering and other science and technology courses at M.L. Sultan Technikon*. Unpublished research report.

De Weerdt, E.H. (1927). A study of the improvability of fifth grade school children in certain mental functions. *Journal of Educational Psychology, 18,* 547-557.

Dearborn, W.F. (1921). Intelligence and its measurement. *Journal of Educational Psychology, 12*, 210-212.

Embretson, S.E. (1987). Toward development of a psychometric approach. In C.S. Lidz (Ed.), *Dynamic assessment. An interactional approach to evaluating learning potential* (pp. 141-170). New York: Guilford Press.

Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56(3)*, 495-515.

Embretson, S.E. (1992). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement, 29(1)*, 25-50.

Embretson, S.E. (1995). A measurement model for linking individual learning to process and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement, 32(3)*, 277-294.

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8(4)*, 341-349.

Employment Equity Act, No 55 of 1998. *Government Gazette no 19370.*

Ertubey, C. & Russell, R.J.H. (1996). *Dealing with comparability problem of cross-cultural data*. Paper presented at the 26th International Congress of Psychology, Montreal, Canada, 16-21 August 1996.

Eysenck, H.J. (1971). *Race, Intelligence, and Education*. London: Temple Smith.

Eysenck, H.J. (1988). The concept of 'intelligence': Useful or useless. *Intelligence, 12*, 1-16.

Eysenck, H.J. (1994). A biological theory of intelligence. In D.K. Detterman (Ed.), *Current Topics in Human Intelligence (Volume 4 - Theories of intelligence)*. New Jersey: Ablex Publishing Corporation.

Eysenck, H.J. & Kamin, L. (1981). *The intelligence controversy*. New York: John Wiley and Sons.

Fancher, R.E. (1985). *The intelligence men: Makers of the IQ controversy*. New York: W.W. Norton & Company.

Ferrara, R., Brown, A. & Campione, J. (1986). Children's learning and transfer of inductive reasoning rules: Studies in proximal development. *Child Development, 57*, 1087-1099.

Feuerstein, R. (1972). Cognitive assessment of the socioculturally deprived child adolescent. In L.J. Cronbach & P.J.D. Drenth, (Eds.), *Mental tests and cultural adaptation* (pp. 266-275). The Hague: Mouton.

Feuerstein, R. (1979). *The dynamic assessment of retarded performance*. Baltimore, MD: University Park Press.

Feuerstein, R., Feuerstein, R. & Gross, S. (1997). The learning potential assessment device. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 297-313). New York: The Guilford Press.

Feuerstein, R., Rand, Y., Jensen, M.R., Kaniel, S. & Tzuriel, D. (1987). Prerequisites for testing of learning potential: The LPAD model. In C.S. Lidz (Ed.), *Dynamic testing* (pp. 35-51). New York: Guilford Press.

Fick, M.L. (1929). Intelligence test results of poor white, native (Zulu), coloured and

Indian school children and the educational and social implications. *South African Journal of Science, 26*, 904-920.

Fick, M.L. (1939). *The educability of the South African Native. South African Council for Educational and Social Research: Research Series No. 8*. Pretoria: South African Council for Educational and Social Research.

Foxcroft, C.D. (1997a). *The PASS theory of cognitive processes as an alternative model in intellectual assessment*. Paper delivered at the national symposium on the Standardisation of the WAIS III for South Africa, Human Sciences Research Council, Pretoria, 6 March 1997.

Foxcroft, C.D. (1997b). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment, 13(3)*, 229-235.

Fraser, B.J., Walberg, H.J., Welch, W.W. & Hattie, J.A. (1987). Synthesis of educational productivity research. *International Journal of Educational Research, 11*, 145-252.

Frisby, C.L. & Braden, J.P. (1992). Feuerstein's dynamic assessment approach: A semantic, logical and empirical critique. *The Journal of special Education, 26(3)*, 281-301.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences (2nd ed)*. London: Fontana Press, New York: Basic Books.

Garrett, H.E. (1941). *Great experiments in psychology*. New York: D. Appleton-Century Company.

Gierl, M.J. & Hanson, A. (1995). *Evaluating the goodness-of-fit between Alberta education achievement test data and model assumptions in unidimensional item response theory.* Unpublished research report prepared for the Alberta education, student evaluation branch.

Goldstein, H. (1989). *Equity in testing after Golden Rule*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, March 27-31, 1989.

Gottfredson, L.S. (1997a). Foreword to "Intelligence and social policy". *Intelligence, 24(1)*, 1-12.

Gottfredson, L.S. (1997b). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24(1)*, 13-23.

Gould, S.J. (1981). *The mismeasure of man*. New York: W.W. Norton.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21(4)*, 347-360.

Gregory, R.J. (1996). *Psychological testing: History, principles, and applications*. Boston: Allyn & Bacon.

Grigorenko, E.L, & Sternberg, R.J. (1998). Dynamic testing. *Psychological Bulletin, 124(1)*, 75-111.

Gupta, T.M. & Coxhead, P. (1988). Why assess learning potential? In R.M. Gupta & P. Coxhead (Eds.), *Cultural diversity and learning efficiency: Recent developments in assessment* (pp. 1 - 21). Hong Kong: Macmillan Press.

Guthke, J. (1992). Learning tests - the concept, main research findings, problems and trends. *Learning and Individual Differences, 4(2)*, 137-151.

Guthke, J. (1993a). Current trends in theories and assessment of intelligence. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 13-20). Amsterdam: Swets & Zeitlinger.

Guthke, J. (1993b). Developments in learning potential assessment. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 43-68). Amsterdam: Swets & Zeitlinger.

Guthke, J. (1998). *Validity of learning test versions of the Raven test*. Paper presented at the 24th International Congress of Applied Psychology, San Francisco, 9-14 August, 1998.

Guthke, J. & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment, 12(1)*, 1-13.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1995). *Multivariate data analysis (4th Ed.).* New Jersey: Englewood Cliffs.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10(3)*, 229-244.

Hambleton, R.K. & Slater, S.C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment, 13(1)*, 21-28.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhof Publishing.

Hambleton, R.K. & Zaal, J.N. (Eds.). (1991). *Advances in educational and psychological testing: Theory and applications*. Boston: Kluwer Academic Publishers.

Hambleton, R.K., Zaal, J.N. & Pieters, J.P.M. (1991). Computerized adaptive testing: Theory, applications and standards. In R.K. Hambleton & J.N. Zaal (Eds.),

Advances in educational and psychological testing: Theory and applications (pp. 341-366). Boston: Kluwer Academic Publishers.

Hamers, J.H.M., Hessels, G.P. & Pennings, A.H. (1996). Learning potential in ethnic minority children. *European Journal of Psychological Assessment, 12(3)*, 183-192.

Hamers, J.H.M. & Resing, W.C.M. (1993). Learning potential assessment: Introduction. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 23-42). Amsterdam: Swets & Zeitlinger.

Hankins, J.A. (1990). The effects of variable entry for a Bayesian adaptive test. *Educational and Psychological Measurement, 50*, 785-802.

Harris, C.W. (1967). *Problems in measuring change*. Madison: The University of Wisconsin Press.

Haywood, H.C. & Switzky, H.N. (1986). The malleability of intelligence: Cognitive processes as a function of polygenic-experiential interaction. *School Psychology Review, 15(2)*, 245-255.

Hegarty, S. (1988). Learning ability and psychometric practice. In R.M. Gupta & P. Coxhead (Eds.), *Cultural diversity and learning efficiency: Recent developments in assessment* (pp. 22 - 38). Hong Kong: Macmillan Press.

Herrnstein, R.J. & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Hessels, M.G.P. & Hamers, J.H.M. (1993). A learning potential test for ethnic minorities. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 285-312). Amsterdam: Swets & Zeitlinger.

Hetter, R.D., Segall, D.O. & Bloxom, B.M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W.A. Sands, B.K. Waters & J.R. McBride, *Computerized adaptive testing: From inquiry to operation* (pp. 161-168). Washington, DC: American Psychological Association.

Hilliard, A.G. III (1990). Back to Binet: The case against the use of IQ tests in the schools. *Contemporary Education, 61(4)*, 184-189.

Holland, P.W. & Wainer, H. (1993). Preface. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. xiii - xv). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P.W. & Wainer, H.. (Eds.) (1994). *Differential item functioning: Theory and Practice*. Hillsdale, NJ: Erlbaum.

Hubbard, E.A. (1996). The IQ caste and gifted education. *Roeper Review, 18(4)*,

258-260.

Hugo, H.L.E. & Claassen, N.C.W. (1991). *The functioning of the GSAT Senior for students of the Department of Education and Training*. Pretoria: Human Sciences Research Council.

Huysamen, G.K. (1998). Die voorspelling van akademiese prestasie na die eerste universiteitsjaar (The prediction of academic achievement after the first year at university). *Journal of Industrial Psychyology*, 24(1), 41-46.

Huysamen, G.K. (1999). Psychometric explanations for the poor predictability of the tertiary-academic performance of educationally disadvantaged students. *South African Journal of Higher Education, 13(1)*, pp. 132-138.

Jensen, A.R. (1963). Learning ability in retarded, average, and gifted children. *Merrill-Palmer Quarterly, 9(2)*, 123-140.

Jensen, A.R. (1969a). How much can we boost IQ and scholastic achievement? Harvard *Educational Review, 39*, 1-123.

Jensen, A.R. (1969b). Intelligence, learning ability and socioeconomic status. *Journal of Special Education, 3(1)*, 23-33.

Jensen, A.R. (1974). How biased are culture-loaded tests? *Genetic Psycholical Monographs, 90*, 185-245.

Jensen, A.R. (1980). *Bias in mental testing*. London: Methuen.

Jensen, A.R. (1981). *Straight talk about mental tests*. London: Methuen.

Jensen, A.R. (1994). Phlogiston, animal magnetism and intelligence. In D.K. Detterman (Ed.), *Current Topics in Human Intelligence (Volume 4 - Theories of Intelligence)*. New Jersey: Ablex Publishing Corporation.

Kline, P. (1991). *Intelligence: The psychometric view*. London: Routledge.

Kozulin, A. & Falik, L. (1995). Dynamic cognitive assessment of the child. *Current Directions in Psychological Science, 4(6)*, 192-196.

Langenhoven, H.P. (1957). *Vergelyking van prestasies van Afrikaanse en Engelse groepe wat gelyk gemaak is t.o.v. nie-verbale roupuntteling in die Nuwe Suid-Afrikaanse Groeptoets (NSAGT)*. Pretoria: Nasionale Buro vir Opvoedkundige en Maatskaplike Navorsing.

Laughon, P. (1990). The dynamic assessment of intelligence: A review of three approaches. *School Psychology Review, 19(4)*, 459-470.

Lidz, C.S. (1987a). *Dynamic assessment: An interactional approach to evaluating learning potential*. New York: The Guilford Press.

Lidz, C.S. (1987b). Historical perspectives. In C.S. Lidz (Ed.), *Dynamic Assessment:*

An interactional approach to Evaluating learning potential (pp. 3-32). New York: The Guilford Press.

Lidz, C.S. (1991). *Practitioner's guide to dynamic assessment*. New York: The Guilford Press.

Lidz, C.S. (1992). The extent of incorporation of dynamic assessment into cognitive assessment courses: a national survey of school psychology trainers. *The Journal of Special Education, 26(3)*, 325-331.

Lidz, C.S. (1997). Dynamic assessment approaches. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison, (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 281-297). New York: The Guilford Press.

Linn, R.L. (1989). *Intelligence: Measurement, theory, and public policy*. Chicago: University of Illinois Press.

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-366). Hillsdale, NJ: Lawrence Erlbaum Associates.

Locurto, C. (1991). *Sense and nonsense about IQ - the case for uniqueness*. New York: Praeger.

Loehlin, J.C. (1992). Should we do research on race differences in intelligence? *Intelligence, 16,* 1-4.

Lohman, D.F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher, 22*, 12-23.

Lord, F.M. (1967). Elementary models for measuring change. In C.W. Harris (Ed.), *Problems in measuring change* (pp. 21-38). London: The University of Wisconsin Press.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Menlo Park: Addison-Wesley.

Matarazzo, J.D. (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic and courtroom. *American Psychologist, 45(9)*, 999-1017.

Mauer, K. & Retief, A.I. (Eds.). (1987). *Psychology in context: Cross-cultural research trends in South Africa*. Pretoria: Human Sciences Research Council.

McBride, J.R. (1997). Technical Perspective. In W.A. Sands, B.K. Waters & J.R.

McBride, *Computerized adaptive testing: From inquiry to operation* (pp. 29-44). Washington, DC: American Psychological Association.

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment, 10*, 1-9.

Miller, R. (1992). *Psychology : Foundation Series*. Durban: University of Natal (TTT Programme)

Minick, N. (1987). Implications of Vygotsky's theories for dynamic assessment. In C.S. Lidz, (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 116-140). New York: The Guilford Press.

Muniz, J. (1998). Book Review: Handbook of modern item response theory, edited by W.J. van der Linden and R.K. Hambleton. New York: Springer-Verlag. *European Journal of Psychological Assessment, 14(1)*, 91-93.

Neman, B.S. (1989). *Writing effectively*. New York: Harper & Row.

Norusis, M.J. / SPSS Inc (1993). *SPSS for Windows - Base System User's Guide Release 6.0*. Chicago: SPSS Inc.

Osterlind, S.J. (1983). *Test item bias*. Beverly Hills: Sage.

Owen, K. (1998). *The role of psychological tests in education in South Africa: Issues, controversies and benefits*. Pretoria: Human Sciences Research Council.

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

Passow, A.H. & Frasier, M.M. (1996). Minority and disadvantaged students: Toward improving identification of talent potential among minority and disadvantaged students. *Roeper Review, 18(3)*, 198-202.

Plomin, R. (1997). Genetics and intelligence: What's New? *Intelligence, 24(1)*, 53-77.

Psychological Society of South Africa. (1998a). *Guidelines for the validation and use of assessment procedures for the workplace*. Pretoria: Society for Industrial Psychology.

Psychological Society of South Africa. (1998b). *Code of practice for psychological assessment for the work place in South Africa*. Pretoria: Society for Industrial Psychology.

Pyryt, M.C. (1996). IQ: Easy to bash, hard to replace. *Roeper Review, 18(4)*, 255-258.

Raven, J.C. (1958). *Standard Progressive Matrices Sets A, B, C, D and E.* London: H.K. Lewis & Co.

Raven, J.C., Court, J.H. & Raven, J. (1977). *Manual for the Raven's progressive matrices and vocabulary scales*. London: H.K. Lewis & Co.

Raven, J.C., Court, J.H. & Raven, J. (1985). *Manual for the Raven's progressive matrices and vocabulary scales (1985 Edition)*. London: H.K. Lewis & Co.

Reality check. (1999). *Sunday Independent*, 2 May 1999.

Reckase, M.D. (1988). *Computerized adaptive testing: a good idea waiting for the right technology*. Paper presented at the meeting of the American Educational Research Association, New Orleans, April 1988.

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8(3)*, 11-15.

Reckase, M.D. (1996). Test construction in the 1990's: Recent approaches every psychologist should know. *Psychological Assessment, 8(4)*, 354-359.

Ree, M.J. & Jensen, H.E. (1983). Effects of sample size on linear equating of item characteristic curve parameters. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 135-146). New York: Academic Press.

Reschly, D.J. & Wilson, M.S. (1990). Cognitive processing versus traditional intelligence: Diagnostic utility, intervention implications, and treatment validity. *School Psychology Review, 19(4)*, 443-458.

Retief, A. (1988). *Method and theory in cross-cultural psychological assessment*. Pretoria: Human Sciences Research Council.

Richardson, K. & Bynner, J.M. (1984). Intelligence: Past and future. *International Journal of Psychology, 19*, 499-526.

Rogoff, B. & Wertsch, J.V. (Eds.) (1984). *Children's learning in the 'zone of proximal development'*. San Francisco: Jossey-Bass.

Sands, W.A., Waters, B.K. & McBride, J.R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schepers, J.M. (1972). Critical issues which have to be resolved in the construction of tests for developing groups. *Humanitas, 2(4)*, 395-406.

Schepers, J.M. (1998). *The bell curve revisited: A South African perspective*. Paper presented at the XIVth international conference of the Association for Cross-cultural Psychology, Bellingham, USA, 3-8 August.

Schmitt, N. & Noe, R.A. (1986). Personnel selection and equal employment

opportunity (pp. 71-115). In C.L. Cooper & I. Robertson, (Eds.), *International Review of Industrial and Organizational Psychology*. Boston: John Wiley and Sons Ltd.

Schoonman, W. (1989). *An applied study on computerized adaptive testing*. Amsterdam: Swets & Zeitlinger.

Shayer, M. & Beasley, F. (1987). Does instrumental enrichment work? *British Educational Research Journal, 13(2)*, 101-119.

Shirley, D.W. (1992). *Psychometric testing and organisational culture: Implications for affirmative action*. Paper presented at the South African Psychometrics conference, June 1992.

Shochet, I.M. (1992). A dynamic testing for undergraduate admission: The inverse relationship between modifiability and predictability. In H.C. Haywood & D. Tzuriel (Eds.), *Interactive testing* (pp. 332-355). New York: Springer-Verlag.

Shochet, I.M. (1994). The moderator effect of cognitive modifiability on a traditional undergraduate admissions test for disadvantaged black students in South Africa. *South African Journal of Psychology, 24(4)*, 208-215.

Shuttleworth-Jordan, A.B. (1996). On not reinventing the wheel: a clinical perspective on culturally relevant test usage in South Africa. *South African Journal of Psychology, 26(2)*, 96-102.

Sijtsma, K. (1993a). Classical and modern test theory with an eye toward learning potential testing. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars, *Learning Potential Assessment: Theoretical, methodological and practical issues* (pp. 117-134). Amsterdan: Swets & Zeitlinger.

Sijtsma, K. (1993b). Psychometric issues in learning potential assessment. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars, *Learning Potential Assessment: Theoretical, methodological and practical issues* (pp. 175-194). Amsterdan: Swets & Zeitlinger.

Skuy, M.S., Kaniel, S. & Tzuriel, D. (1988). Dynamic assessment of intellectually superior Israeli children in a low socio-economic status community. *Gifted Education International, 5*, 90-96.

South African Professional Board for Psychology. (1998). *Policy on the classification of psychometric measuring devices, instruments, methods and techniques. (18/9/B)*. Pretoria: South African Professional Board for Psychology.

Spearman, C.E. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Spitz, H.H. (1989). Variations in Wechsler interscale IQ disparities at different levels of IQ. *Intelligence, 13*, 157-167.

Sternberg, R.J. (1984). A contextualist view of the nature of intelligence. *International Journal of Psychology, 19*, 307-334.

Sternberg, R.J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R.J. (1991). Death, taxes and bad intelligence tests. *Intelligence*, 15, 257-269.

Sternberg, R.J. (1997a). Intelligence and lifelong learning: What's new and how can we use it? *American Psychologist, 52(10)*, 1134-1139.

Sternberg, R.J. (1997b). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52(10)*, 1030-1037.

Suzuki, L.A. & Valencia, R.R. (1997). Race-ethnicity and measured intelligence. *American Psychologist, 52(10)*, 1103-1114.

Taylor, T.R. (1992). *Beyond competence: Measuring potential in a cross-cultural situation fairly: Potential in Psychometrics (part two).* Paper presented at the 1992 South African Psychometrics Congress, June 1992.

Taylor, T.R. (1994a). A review of three approaches to cognitive assessment, and a proposed integrated approach based on a unifying theoretical framework. *South African Journal of Psychology, 24(4)*, 184-193.

Taylor, T.R. (1994b). *Learning potential: Theory and practical assessment*. Paper presented at the Annual Department of Industrial Psychology Conference on Fairness in Testing and Assessment (30 August 1994). Pretoria: Holiday Inn.

Terman, L.M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale.* Boston: Houghton Mifflin.

Terman, L.M. & Merrill, M.A. (1937). *Measuring intelligence*. Boston: Houghton Mifflin.

Thorndike, R.L. (1922). Practice effects in intelligence tests. *Journal of Experimental Psychology, 5*, 101-107.

Thorndike, R.M. & Lohman, D.F. (1990). *A century of ability testing*. Chicago: The Riverside Publishing Company.

Thurstone, L.L. (1938). *Primary mental abilities.* Chicago: Chicago University Press.

Tzuriel, D. (1992). The dynamic assessment approach: A reply to Frisby and Braden. The *Journal of Special Education, 26(3)*, 302-324.

Tzuriel, D. (1997). A novel dynamic assessment approach for young children: Major dimensions and current research. *Educational and Child Psychology, 14(4)*, 83-108.

Tzuriel, D. & Haywood, H.C. (1992). The development of interactive-dynamic approaches to assessment of learning potential. In H.C. Haywood & D. Tzuriel (Eds.), *Interactive Assessment*. New York: Springer-Verlag.

Van de Vijver, F. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-cultural Psychology, 28(6)*, 678-709.

Van den Berg, A.R. (1989). *Basic item response theory*. Pretoria: Human Sciences Research Council.

Van der Linden, W.J. & Hambleton, R.K. (Eds). (1997). *Handbook of modern item response theory*. New York: Springer.

Van der Veer, R. & Valsiner, J. (1991). *Understanding Vygotsky: A quest for synthesis*. Cambridge, MA: Blackwell.

Van Eeden, R. (1993). *The validity of the Senior South African Individual Scale - Revised (SSAIS-R) for children whose mother tongue is an African language: Private schools.* Pretoria: Human Sciences Research Council.

Van Niekerk, H.A. (1991). *Evaluation of Feuerstein's instrumental enrichment programme for culturally different senior secondary students*. Unpublished Master's Thesis, University of Pretoria.

Van Tonder, M. & Claassen, N.C.W. (1992). *Manual for the General Scholastic Aptitude Test (Senior) - Computerized Adaptive Test*. Pretoria: Human Sciences Research Council.

Vapi, X. (1998). Equity act's benefits won't come soon. *The Star, 5 December 1998*, p11.

Venter, E.J. (1997). Manual: *Test of Basic Numerical Literacy*. Pretoria: Human Sciences Research Council.

Verster, J.M. (1987). Cross-cultural cognitive research: Some methodological problems and prospects. In K.F. Mauer & A.I. Retief (Eds.), *Psychology in context: Cross-cultural research trends in South Africa* (pp. 65-118). Pretoria: Human Sciences Research Council.

Verster, J.M. & Prinsloo, R.J. (1988). The diminishing test performance gap between English speakers and Afrikaans speakers in South Africa. In S.H. Irvine & J.W. Berry, (Eds.), *Human abilities in cultural context* (pp. 534-560). Cambridge: Cambridge University Press.

Vincent, K.R. (1991). Black/White IQ differences: Does age make the difference? *Journal of Clinical Psychology, 27(2)*, 266-270.

Visser, J.D. (1996). *Navorsingsgeleenthede vir gedragswetenskaplikes in die werkplek*. Professorale intreerede, 23 Oktober 1996.

Von Hirschfeld, S. (1992a). *Exploring the use of trainability testing in South Africa*. Paper presented at the South African Psychometrics Congress, June 1992.

Von Hirschfeld, S. (1992b). *The psychological concept of potential*. Paper presented at the South African Psychometrics Congress, June 1992.

Vygotsky, L.S. (1978) *Mind in society: The development of higher-order psychological processes*. Cambridge, MA: Harvard University Press.

Wainer, H. (1993). Model-based standardised measurement of an item's differential impact. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Warm, T.A. (1978). *A primer of item response theory (Technical Report 941078)*. Oklahoma city, OK: U.S. Coast Guard Institute.

Weiss, D.J. (Ed.) (1983a). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Weiss, D.J. (1983b). Introduction. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Weiss, D.J. (1983c). *Computer-based measurement of intellectual capabilities : Final report*. Minnesota: University of Minnesota.

Weiss, D.J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.

Weiss, D.J. & Vale, C.D. (1987). Adaptive testing: *Applied Psychology: an International review, 36(3&4)*, 249-262.

Weiss, D.J. & Yoes, M.E. (1991). Item response theory. In R.K. Hambleton & J.N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 69-95). Boston: Kluwer Academic Publishers

Wolf, T. (1973). *Alfred Binet*. Chicago: University of Chicago Press.

Zaaiman, H. (1995). *The UNIFY selection research project: Rationale, background, literature overview and 1994 student data*. Unpublished report: University of the North.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zolezzi, S.A. (1995). *The effectiveness of dynamic assessment as an alternative aptitude testing strategy*. Unpublished doctoral thesis. University of South Africa.

APPENDIX A

APPENDIX B

APPENDIX C