A PRELIMINARY INVESTIGATION INTO THE PATTERNS OF PERFORMANCE ON A COMPUTERIZED ADAPTIVE TEST BATTERY: IMPLICATIONS FOR ADMISSIONS AND PLACEMENT

MARLENE VORSTER

A PRELIMINARY INVESTIGATION INTO THE PATTERNS OF PERFORMANCE ON A COMPUTERIZED ADAPTIVE TEST BATTERY: IMPLICATIONS FOR ADMISSIONS AND PLACEMENT

MARLENE VORSTER

Submitted in partial fulfillment of the requirements for the degree of

MAGISTER ARTIUM

In the Faculty of Health Sciences at the

UNIVERSITY OF PORT ELIZABETH

January 2002

Supervisor: Professor C.D. Foxcroft

I dedicate the write-up of this research to my father, Adolph Joachim Kurt Uderstadt, who did not get to see the results of finalizing the project report but had a major influence on the process and conclusion of this dissertation.

ACKNOWLEDGEMENTS

I extend my sincere gratitude to the following, without whom this dissertation would never have been finished:

To my Heavenly Father who has been faithful in the trying times and who has strengthened my resolve when I felt like giving up.

To my supervisor, Cheryl Foxcroft, for the refining of the research topic, for including me in the practicalities of placement assessment, thereby allowing me an opportunity to learn from her experience, for the suggestions and input, and for helping me with rounding off the argument.

To Gayna Astbury, for being available when I needed to chat through ideas for the proposal.

To my husband, André, for his love, practical support and encouragement through the process.

To my mom, Gloria Uderstadt, for all her practical, emotional and spiritual support through my student years.

To my gran, Iris Robbie, for reminding me to "think positive".

To my friend, Joan Matthews, for listening to and understanding all my concerns about motivation, thereby encouraging me to endure and persevere.

To the NRF for their financial assistance. Opinions expressed and conclusions reached are those of the author and should not be considered a reflection of the opinions and conclusions of the National Research Foundation.

To the University of Port Elizabeth for the financial assistance provided over the years of studying.

To Elize Koch for helping me with my data and patiently explaining how to interpret cluster analysis.

To Greg Saunders, Andrea Watson, and Johan Cronje for their assistance with extracting data from the larger database which facilitated the finalization of the data used for this study.

To Robert Callahan for looking up references when it was impossible for me to do so.

To Annemarie Barnard, my classmate and friend, for her interest and for setting an example of the work ethic that kept me going.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The fallibility of human judgment in the making of decisions requires the use of tests to enhance decision-making processes. Although testing is surrounded with issues of bias and fairness, it remains the best means of facilitating decisions over more subjective alternatives. As a country in transition, all facets of South African society are being transformed. The changes taking place within the tertiary education system to redress the legacy of Apartheid, coincide with an international trend of transforming higher education. One important area that is being transformed relates to university entrance requirements and admissions procedures. In South Africa, these were traditionally based on matriculation performance, which has been found to be a more variable predictor of academic success for historically disadvantaged students. Alternative or revised admissions procedures have been implemented at universities throughout the country, in conjunction with academic development programmes. However, it is argued in this dissertation that a paradigm shift is necessary to conceptualise admissions and placement assessment in a developmentally oriented way. Furthermore, it is motivated that it is important to keep abreast of advances in theory, such as item response theory (IRT) and technology, such as computerized adaptive testing (CAT), in test development to enhance the effectiveness of selecting and placing learners in tertiary programmes.

This study focuses on investigating the use of the Accuplacer Computerized Placement Tests (CPTs), an adaptive test battery that was developed in the USA, to facilitate unbiased and fair admissions, placement and development decisions in the transforming South African context. The battery has been implemented at a university in the Eastern Cape and its usefulness was investigated for 193 participants, divided into two groups of degree programmes, depending on whether or not admission to the degree required mathematics as a matriculation subject. Mathematics based degree programme learners (n = 125) wrote three and non-mathematics based degree programme learners (n = 68) wrote two tests of the Accuplacer test battery. Correlations were computed between the Accuplacer scores and matriculation performance, and between the Accuplacer scores, matriculation performance and academic results. All yielded significant positive relationships excepting for the one subtest of the Accuplacer with academic performance for the non-mathematics based degree group. Multiple correlations for both groups indicated that the Accuplacer

scores and matriculation results contribute unique information about academic performance. Cluster analysis for both groups yielded three underlying patterns of performance in the data sets. An attempt was made to validate the cluster groups internally through a MANOVA and single-factor ANOVAs. It was found that Accuplacer subtests and matriculation results do discriminate to an extent among clusters of learners in both groups of degree programmes investigated. Clusters were described in terms of demographic information and it was determined that the factors of culture and home language and how they relate to cluster group membership need further investigation. The main suggestion flowing from these findings is that an attempt be made to confirm the results with a larger sample size and for different cultural and language groups.

KEY WORDS/PHRASES
Admissions programmes
Admissions and placement assessment
Predictors of academic performance
Item response theory (IRT)
Computerized adaptive testing (CAT)
Bias and fairness

CHAPTER ONE: INTRODUCTION

The broad context for this study will focus on the use of test information to aid decision-making when admitting and placing learners at university entrance. Human judgement is fallible where the making of decisions is concerned (Dahlstrom, 1993). Decision-making becomes more effective when one collects as much information as possible relating to an individual about whom a decision must be made (Foxcroft, 1994). Testing involves one source of information and “the collection of relevant information for making evaluative judgements” (Plug, 1996, p.6).

A complicating factor in measurement is that tests do not directly measure psychological attributes; they measure behaviour that is believed to be linked to the criterion in which there is an interest (Foxcroft, 1994; Murphy & Davidshofer, 1991).

Although there is no method in existence that can guarantee complete accuracy for decision-making relating to human psychological attributes, tests are widely used globally for the purposes of facilitating decision-making (Anastasi, 1988; Murphy & Davidshofer, 1991; Plug, 1996; Schoonman, 1989). They are usually included as part of a strategy for the making of effective decisions (Foxcroft, 1994) because they represent the best, most objective, most accurate and cost-effective method (Anastasi, 1988; Murphy & Davidshofer, 1991; Plug, 1996; Schoonman, 1989).

Decisions usually have consequences for individuals, groups, institutions and society. Tests are useful when there is concern for the development and nurturance of individual potential and the facilitation of productivity and high quality of life (Walsh & Betz, 1985). They benefit individuals and institutions when they contribute positively toward the achievement of their respective goals, and society benefits when the achievement of such goals has an impact for the general good (American Psychological Association, 1985; Canadian Psychological Association, 1987).

This research falls within the boundaries of educational measurement: which can be defined as "the process of specifying the position, or positions, for educational purposes, of persons, situations, or events on educationally relevant scales under stipulated conditions" (Bunderson, Inouye & Olsen, 1989, p. 368).

Educational purposes served through testing are usually varied but basically are concerned with the assistance of educational decision making through the provision of information about the position of a group or individual along educationally relevant

scales, thus serving institutions and individuals. Institutions historically use measurement to improve decisions relating to admissions (i.e., selection) and placement, assessing achievement of educational goals, evaluating staff, programmes and organizational entities, and to motivate students. Individuals tend to be served by such measurement being used for guidance, and counselling, based on achievement, ability, aptitude and interest, progress monitoring, and assisting decisions for instruction (Bunderson, Inouye & Olsen, 1989).

Although educational measurement serves many purposes, the focus of this study is on the use of test information for the purpose of improving admissions decisions for the benefit of institutions and individuals. Any admissions decision has four possible outcomes in that a learner who is admitted person may succeed (true positive), or fail (false positive), whereas someone not admitted may have succeeded if given the opportunity (false negative), or failed if given the opportunity (true negative). Interest is focussed on increasing true positives, that is those admitted learners who will succeed, and reducing the number of false positives, so as not to place people in situations where they are likely to experience failure. In addition, there exists a moral obligation to maintain as small a false positive component as possible because people should not be precluded from opportunities where they would meet with success because of an erroneous belief that they would fail (Brown, 1983; Foxcroft, 1994; Murphy & Davidshofer, 1991).

It is important to keep abreast of advances in test development theory, such as item response theory, and technological advances, such as computerized adaptive testing, when identifying the test battery to be used for the purposes of admitting and placing learners in tertiary programmes. Also, it is important that the tests administered are unbiased and the results used fairly. Testing is only a small aspect of a broader process because decisions reflect reasoning and value judgements on the basis of test results (Cronbach, 1990). Fairness issues around testing are related not only the notions of equal opportunity or equal outcomes, but also to the outcome of decisions made (i.e., distributive fairness), the process by which decisions are made in order to obtain a particular outcome (i.e., procedural fairness) and also the treatment of individuals about whom decisions are made during the process (i.e., interactional fairness) (Nunns & Ortlepp, 1994).

The history and cultural compilation of South Africa is such that bias and fairness are relevant issues within the context of selection and admissions testing,

especially in education, which is a gateway to training for future occupations, and thus ultimate occupational attainment (Spolsky, 1997). It is suggested that a paradigm shift be undertaken in terms of the perception of the purposes of assessment - assessment should not operate solely for admission but also for placement (Foxcroft, 1999). The developmentally oriented placement assessment paradigm constitutes the platform from which this research study was conducted.

The White Paper on Higher Education (1997) and the National Plan for Higher Education (2001) make it clear that universities are autonomous with regard to their admission and placement procedures. However, the goals of these procedures and other policies must be to increase equality and equity of access to universities by identifying students with the potential for success at tertiary-level, and should ultimately retain students of high calibre to increase graduate outputs. Also clarified through the National Qualifications Framework (NQF) is the goal of facilitating the tailoring of degree programmes to the needs of students and society by determining the strengths and weaknesses of learners upon university entrance (Foxcroft, 1999).

The application of well-researched advances in test theory and development and in technology in gathering information so that effective decisions can be made is integral in the process of transformation, not only in higher education, but also in the broader context of social reconstruction (White Paper on Higher Education, 1997).

Specifically, this study focuses on investigating whether a test developed in the USA, using advances in psychometric theory and technology, to enhance the placement of learners into tertiary education programmes can be utilized in the transforming South African context to facilitate unbiased and fair admissions and placement procedures.

Overview of Chapters

Chapter two describes admissions policies and programmes of tertiary institutions within South Africa during the transition period. Included are examples of international admissions methods and South African special admissions methods, also with the citation of a few examples. Furthermore, entry-level proficiencies that are considered to be important for tertiary programmes are outlined. Chapters three and four cover important advances in psychometrics that are relevant to this research. Chapter three on item response theory, which constitutes a modern method of test construction, includes a description and explanation of the more important aspects of and concepts within its development, and mentions how it has

expanded and areas of continued research. Chapter four on computerized adaptive testing describes how this type of testing has developed, how it differs from the conventional paper-and-pencil testing, and the issues involved in computerized adaptive administration of tests. Chapter five covers the issues surrounding testing for admission and placement purposes, namely, bias and fairness, and ultimately describes how these operate within multicultural contexts, with special reference to South Africa. Chapter six delineates the problem being investigated and specifies the aims of the study. Chapter seven describes the methods employed in conducting the study and the statistical analyses applied. Chapter eight contains a presentation of the results and their interpretation, and outlines how this study has contributed to the larger research project in progress. The limitations of this study and possible future research prospects are also addressed.

CHAPTER TWO: ADMISSIONS TO HIGHER EDUCATION INSTITUTIONS

South Africa is in transformation, and thus, all facets of society are undergoing changes in policy and practice. The transformation in higher education, however, coincides with an international trend. This chapter briefly describes admissions procedures to tertiary education institutions in certain overseas countries as well as the history of entrance requirements and current admissions programmes in South Africa. Information relevant for predicting academic success is outlined and the motivation for a paradigm shift concerning admissions and associated procedures is provided.

Transformation in Higher Education

Tertiary education is undergoing transformation internationally. Although in traditionally Western countries this transformation involves evolutionary change (i.e., incremental and gradual), in South Africa, the transformation has been more revolutionary (i.e., radical), at least in theory (Dlamini, 1995). Generally, new educational goals are emerging, resulting in a corresponding change in the goals and methods of assessment (Portes, 1996).

South African universities in transition are faced with the challenge of maintaining and striving for excellence (i.e., in terms of the quality of education) while working toward and implementing equity (i.e., adequate representation through diversification of staff and student profiles) (Dlamini, 1995; Jordaan, 1995; Nel, 1997; Nunns & Ortlepp, 1994; Pavlich, Orkin & Richardson, 1995; Saunders, 1992). The evils of the apartheid system are well documented and the effects on education have been pervasive (Dlamini, 1995; Pavlich, Orkin & Richardson, 1995).

Certain policies have been implemented in an attempt to redress these influences on higher education, especially in terms of the admission of learners from educationally disadvantaged backgrounds. Such policies relate to two broad efforts, namely access initiatives, where the emphasis is on increasing the admission of traditionally disadvantaged learners, and academic support activities, where the emphasis is on enhancing coping abilities of learners and enabling staff members to facilitate this (Nel, 1997; Pavlich, Orkin & Richardson, 1995; Van der Walt, 1995).

There are two stages in formulating policies about admission and academic support, and these are context-specific. First, priorities must be determined.

Specifically, this involves identifying the focus of the institution, namely, whether it concentrated primarily on academic excellence or service delivery, and evaluating how to achieve greater balance on this continuum. Second, decisions must be made for effective programmes to be developed for the achievement of the prioritized goals (Pavlich, Orkin & Richardson, 1995).

One aspect that is inherent in access initiatives and affects the development of academic support activities is the admission process utilized. Special alternative admissions procedures may be developed and refined, and within this realm is the identification of competent learners so that attention can be focussed not so much on remedial training to redress deficits, but the determination of individual strengths to develop capacities for learning through academic support programmes (Pavlich, Orkin & Richardson, 1995).

International Admissions Procedures

The United States of America

The USA has a history of established testing for the purposes of admission to and placement within tertiary educational institutions. Admissions decisions relate to whether and on what grounds learners should be allowed to study at a particular institution, and placement decisions relate to decisions about enrollments or credits for particular courses offered at an institution (Whitney, 1989).

Although admissions procedures are diverse, there are common elements. Generally, policies are influenced by enrollment limitations and projected student applications, are formulated in accordance with faculty requirements and implemented by an administration staff. Decisions themselves are moulded and restricted by value systems, educational bodies and laws (Manning & Jackson, 1984; Whitney, 1989).

There are three levels of admissions decisions:

1. Undergraduate admissions have established minimum criteria which usually include test scores of either the Scholastic Aptitude Test (SAT) or American College Testing Programme assessment (ACT), biodata, secondary school grades and class ranking (Manning & Jackson, 1984; Rutherford & Watson, 1990; Whitney, 1989);
2. Postgraduate admissions are often based on completed coursework at undergraduate level in conjunction with test scores such as the Graduate

Record Examination (GRE) and Graduate Management Admission Test (GMAT) and biodata (Whitney, 1989);

3. Professional programme admissions categorize individuals into three groups based on minimum secondary school grades and admissions test scores such as the Medical College Admission Test (MCAT) and the Law School Admission Test (LSAT), namely immediate admissions, immediate refusals, and hold. Final decisions are then based on additional information derived from biodata and/or interviews (Whitney, 1989).

The principal reason for the development of the tests utilized in the different levels of admission decisions is that relevant student achievement information is then presented on a common scale that is standard across educational institutions, in the absence of knowledge about independent secondary education institutions (Manning & Jackson, 1984; Whitney, 1989). Use of a combination of test scores and high school rank enhances predictive validity for university success (Ben-Shakhar, Kiderman & Beller, 1996).

In addition, the USA has had affirmative action as part of their admissions policies for decades. Furthermore, their education system incorporates two-year public community colleges that offer associate degrees and other qualifications as a stepping stone to four-year colleges. This allows the learners to demonstrate their academic performance over time if they desire to attend a four-year college (Saunders, 1992).

Another option available to learners in the USA is the Advanced Placement Programme (AP), which allows individuals to take credit-bearing courses while still at high school so that they can shorten the time it usually takes to complete college`(College Entrance Examination Board, 1994).

Israel

In Israel nation-wide achievement tests are administered at the end of high school, the successful completion of which results in a certificate of matriculation being achieved. The score for each subject constitutes an average of school assessment and the score on the external examination. The certificate is a requirement for university registration and was the only admission criterion until the 1980's, when the Psychometric Entrance Test (PET) was implemented by the National Institute for Testing and Evaluation (NITE) as an additional admission criterion (Beller, 1994, 1995).

Universities are able to determine their own admissions policy and decide on the selection devices to be used. Learners apply to specific departments rather than to a university or faculty and candidates are chosen in accordance with departmental requirements, the minimum criteria being the matriculation certificate and the PET. Candidates are ranked on the basis of their composite scores, the cut points for which are determined by ability levels of applicants and selection ratios for each field (Beller, 1994, 1995; Jones, 1994).

Sweden

Sweden has two criteria for admission to tertiary education institutions, namely the average grade obtained at senior secondary school, and the Swedish Scholastic Aptitude Testing Programme (SweSAT). In addition, there are formal requirements for different study programmes, usually expressed as a minimum level based on grades obtained in specific subjects (Wedman, 1994).

The SweSAT has been in operation since the late 1970's but applicants to universities have a choice as to whether or not they take the test because they are judged on the most favourable condition (i.e., average grade or SweSAT results) (Feuer & Fulton, 1994; Jones, 1994; Wedman, 1994).

Other Countries

China conducts provincial examinations at the end of nine years of compulsory education, and thereafter, national examinations are administered for admission to tertiary institutions (Feuer & Fulton, 1994).

Japan has entrance examinations for all public and certain private universities, namely the Test of the National Center for University Entrance Examinations (TNCUEE). Thereafter, the College Entrance Examinations are administered by individual tertiary education institutions, the faculties of which decide on admission (Feuer & Fulton, 1994; Saunders, 1992).

In France, a national examination is administered at the end of advanced school, which is the required criterion for university admission. However, regional tests are also administered (Feuer & Fulton, 1994).

In Germany, university entrance is based on the successful completion of the examination administered at state level after the final year of secondary schooling (Feuer & Fulton, 1994; Saunders, 1992).

In the United Kingdom, university entrance is dependent on two levels of examinations, the first being attainment levels and grades on the General Certificate

of Secondary Education, which is locally determined, and the second being success on advanced examinations offered in the upper grades of comprehensive school (Feuer & Fulton, 1994). More than two of the latter is often a minimum requirement for admission to university (Rutherford & Watson, 1990; Saunders, 1992).

Finally, Australia state testing determines admission. For example, Queensland utilizes the Australian Scholastic Aptitude test to rank learners, and from this derive a tertiary entrance score for university (Henry, 1988).

Admissions to South African Universities: Past and Present

Past Entrance Requirements

Prior to the radical revision of the Universities Act originally passed in 1955, the minimum requirement for admission to a university was a matriculation certificate with exemption. Exemption criteria included that the individual had successfully completed a first language and second language at a higher-grade level, and had successfully completed three distinct subject groups at higher-grade level (Le Roux, 1995). Admission was also possible for those who qualified on the grounds of an age criterion, and conditional admission could be obtained on the basis of scores on a battery of tests to determine whether an individual would cope with the academic demands of tertiary education (Smith & Beecham, 1994).

Thus, it was usually matriculation results, alone or transformed by the Swedish point system, that were utilized to determine admissions decisions (Foxcroft, 1999; Greyling & Calitz, 1997; Nunns & Ortlepp, 1994; Sharwood & Rutherford, 1994; Smith & Beecham, 1994) as it has repeatedly been demonstrated that these are at least a reasonable predictor of success in tertiary education (Badenhorst, Foster & Lea, 1990; Louw, 1992; Nunns & Ortlepp, 1994; Sharwood & Rutherford, 1994). However, it has been documented that matriculation results do not adequately predict the academic success of historically disadvantaged learners (Greyling & Calitz, 1997; Louw, 1992; Miller, 1992; Skuy, Zolezzi, Mentis, Fridjhon & Cockroft, 1996; Smit, n.d., Sharwood & Rutherford, 1994; Smith & Beecham, 1994). These findings, in combination with the transformation of tertiary education in the country, has led to the adoption of alternative or the revision of admissions programmes in conjunction with academic support programmes (Pavlich, Orkin & Richardson, 1995).

Early Proposals for Alternative Admissions Programmes

During the 1980s, a number of admissions procedures were suggested and derived from research so as to broaden access to tertiary studies. One was that a

two-tiered admissions system be adopted. Specifically, individuals would be selected either on the basis of their achieving a specified minimum average at secondary school, or by graduating in the top 50 percent of their matriculation class (Louw, 1994).

Another suggestion was that achievement in a certain combination of matriculation subjects could be used as predictors of academic success, and that differential requirements be implemented for different fields of study (Louw, 1994).

It was also proposed that academic potential be assessed with a test battery similar to that used in the USA, and that the results of these be used in conjunction with matriculation results to determine decisions. In addition, it was recommended that colleges be established for the purposes of remedial training, channeling and certification (Louw, 1994).

A fourth suggestion was that learners wishing to apply for tertiary education should undergo a thirteenth year of schooling in which they would be academically prepared for what would be expected of them at university level. In addition, during this year, their emotional stability and motivation could be monitored. Matriculation results would then be considered in combination with the results from this additional thirteenth year of schooling to determine university entrance (Louw, 1994).

Along similar lines was the suggestion that introductory courses be offered at university with performance assessment, and that this then determine university admission rather than matriculation results (Nunns & Ortlepp, 1994).

To some extent, this also overlaps with the idea that individuals should be tested for potential to learn, and this potential is not evaluated by standard cognitive assessment and aptitude tests, which actually only identify manifest potential (Louw, 1992; Shochet, 1994). The argument has been that tests tend to tap prior learning and reflect a background of education and experience that is tainted with SES factors (Miller, 1992). The operational definition of this learning potential or cognitive modifiability was translated into dynamic or interactive assessment, which was epitomized in the advocated test-teach-test (TTT) programme (Shochet, 1994). This involves testing individuals, exposing them to appropriate teaching over a time period that can vary, and then evaluating the individuals again. The learners who demonstrate the greatest improvement are those with the greatest academic potential (Louw, 1992). A variant of this, called the teach-test-teach programme, was implemented at the University of Natal (Skuy, Zolezzi, Mentis, Fridjhon & Cockroft,

1996; The TTT Programme, 1993) and was incorporated into the Regional Access Programme (RAP) (Delvare, 1996).

Alternative or Revised Admissions Programmes

Those learners who do not qualify for admission on the basis of their matriculation results may qualify through an alternative admissions procedure (Nelson & Rainier, n.d., Smith & Segall, 1994), which usually involves assessment for potential to succeed using psychometric tests and/or specialized admissions instruments (Bodibe, 1995; Jacobs, 1995). The results of the additional tests are then used in conjunction with matriculation performance and/or a scholastic record, and/or biographical information (Bodibe, 1995; Nelson & Rainier, n.d.). A few examples of alternative admissions programmes will be provided below.

Examples of Alternative Admissions Programmes

University of Cape Town. The Alternative Admissions Research Project (AARP) is a voluntary testing programme that was implemented in the late 1980s. The tests utilized have changed over time and now include an English proficiency test, which is used for placement purposes, and two tests of mathematical skills. First time applicants for undergraduate studies are invited to take the tests at one of 22 centres and tests are administered three times a year. Learners who would have been admitted on the basis of their matriculation results alone are not refused admission to the university on the basis of their assessment results. The aim of this project is to identify individuals with the potential to succeed at university, and who would not be recognized as such on the basis of matriculation results alone (AARP, 1996; Badsha & Yeld, 1991; Delvare, 1996; Yeld, 1992; Yeld, Haeck, Shall & Hiscock, 1994; Watson, 1997; Skuy, Zolezzi, Mentis, Fridjhon & Cockroft, 1996).

Member Universities of the Eastern Seaboard Association of Tertiary Institutions (esATI). The Regional Access Programme (RAP) is available to applicants who do not meet minimum requirements. This is available to learners at the universities of Natal, Durban-Westville and Zululand, UNISA, and certain technikons. The Distance Access Course is offered to applicants desiring a degree in Arts, Social Sciences and Law. Learners obtain relevant materials and are offered extensive academic support. RAP makes recommendations for admission on the basis of performance during the course and in the final examination, and this serves as a replacement for matriculation results (Watson, 1997; Williams, 1999).

University of the Witwatersrand. Applicants who do not meet minimum entrance requirements undergo psychometric testing to assess English proficiency, reasoning, basic mathematics and science, and aptitude for science. In addition, biographical information is required and interviews are conducted to finalize admissions decisions (Watson, 1997).

University of Pretoria. Testing is open to all first-time applicants to the university, and the tests administered were developed internally, and focus on language proficiency, reasoning and mathematical skills (Nel, 1996; Watson, 1997).

Rand Afrikaans University. Tests are administered for all first-time applicants who do not meet minimum entrance requirements and include language proficiency, learning potential and certain non-cognitive aspects (Kotze, Van der Merwe & Nel, 1996; Watson, 1997).

Potchefstroom University for Christian Higher Education. Applicants who desire first-time admission, but do not meet the minimum requirements undergo certain tests that focus on learning potential, language proficiency in English and Afrikaans, and proficiency in mathematics (Kotze, Van der Merwe & Nel, 1996; Watson, 1997).

University of Stellenbosch. Applicants who do not meet minimum requirements are obliged to write access tests in one of two groupings. Those who intend to study Science-related degrees write a test battery including Mathematics, Science, Afrikaans or English, and those who intend to study in the Humanities undergo a test battery that includes Cognitive Skills, English or Afrikaans. An extended route of study may be suggested as an available option (ADP Overview, 2001).

University of the Free State. Learners who are applying for the first time but do not meet minimum entrance requirements are assessed using standard psychometric instruments to investigate aptitude, personality and interests. Also incorporated is school performance and language proficiency. This is followed by an interview, if testing criteria are met. Career counselling is also offered for these applicants (Delvare, 1996; Watson, 1997).

Rhodes University. Applicants who do not meet minimum requirements undergo testing and admissions decisions are finalized on consideration of the test results in conjunction with biographical information (Watson, 1997).

University of Port Elizabeth. Applicants who do not meet minimum requirements undergo testing in order to make admissions decisions (Watson, Van Lingen & De Jager, 1997). In addition, the university offers UPEAP, which is a form of bridging

course for previously disadvantaged learners (Snyders, 1997). The university is currently developing a proposal to use test and secondary school results for all first year applicants that will improve decision-making to all tertiary programmes (Foxcroft, 1999).

The 1997 White Paper on Higher Education, with its focus on a learner-centered outcomes based approach, requires the determination of learners' strengths and weaknesses in order that programmes can be tailored to learners needs. Furthermore, matriculation exemption will fall away as a statutory requirement for university access, but tertiary institutions will need to determine entrance prerequisites that take into consideration educational backgrounds and prior learning. Thus, proficiencies and potential of learners will have to be identified and operationalised, and then evaluated, implying that assessment will play a vital role in higher education application procedures (Foxcroft, 1999).

The National Plan for Higher Education (2001) highlighted what government has identified as problem areas that need to be addressed in higher education in South Africa. These include increased access and enrollment of learners, bearing the concept of equity in mind and including those from other Southern African countries, retention of learners, broadening of the social base of learners, and changed enrollments in fields to move away from humanities and towards science and commerce. It was stated that universities need appropriate admissions procedures to ensure that only learners with potential are recruited and only those with the ability to succeed are enrolled. Universities are required to develop the necessary administrative, management and academic structures to achieve the outcomes that derive from the identified problem areas.

These points indicate the necessity for accurate and fair assessment procedures that measure attributes related to academic success and that would facilitate selection and placement of learners in appropriate degree programmes.

Information Relevant to the Prediction of Academic Success

It has been stated that establishing accurate predictors of academic performance are vital for developing a fair admissions process (Nunns & Ortlepp, 1994). Although it has been advocated that admission policy should be based on research indicators of what predicts academic success at tertiary level (Venter, 1993), it has already been mentioned that scholastic performance reasonably predicts future success at university, but this generally only holds true for historically

advantaged groups in South Africa (Greyling & Calitz, 1997; Louw, 1992; Miller, 1992; Skuy, Zolezzi, Mentis, Fridjhon & Cockroft, 1996; Smit, n.d., Sharwood & Rutherford, 1994; Smith & Beecham, 1994).

Predictive validity improves when scholastic performance is combined with results on standard cognitive tests. In addition, non-cognitive aspects are believed to influence future academic success (Kotze, 1994; Venter, 1993). Studies conducted relating to selection predictors of performance have incorporated high school results, and have utilized multiple regression analysis with cognitive and non-cognitive variables presumed to predict performance (Burke, 1982). This is true for the South African context as well (e.g., Badenhorst, Foster & Lea, 1990; Calitz, 1997).

Ultimately, a combination of cognitive attributes, non-cognitive attributes and biographical information is presumed to be better predictors of academic success than any of these alone or sub-combined.

Cognitive Attributes

Traditionally, cognitive abilities have been inferred by matriculation results or high school grades (Jacobs, 1995), and the results of cognitive tests in a psychometric battery. However, examination of the common elements of test batteries utilized by different tertiary education institutions both internationally and nationally reveals that there are certain cognitive skills regarded as important for academic success, and these include language proficiency, reasoning and numerical and mathematical skills (Jones, 1994; Nel, 1997; Watson, 1997).

The individual must be proficient in the language of instruction, and this is recognized in countries where the language of instruction is not English (e.g., Beller, 1994; Wedman, 1994), as well as in countries where English is the principal language of instruction (Nel, 1997). Language proficiency is a term that is usually applied to the use of a language as a second language whereas verbal ability is usually applied to the use of a language as a first language. Both proficiency and ability is reflected in four skills, namely, comprehension, speaking, reading and writing (Duran, 1989). These can be categorized into two dimensions, namely receptive and expressive language skills (Foxcroft, 1999). Also, both verbal and non-verbal reasoning and problem solving are regarded as important for contributing to academic success (Foxcroft, 1999).

One aspect that is of concern internationally is the translation of cognitive measures into languages other than the language of instruction (Jones, 1994). This is

an issue for Israel (Beller, 1994; Beller, Gafni & Hanani, 1999), Sweden (Wedman, 1994) and the USA (Jones, 1994).

Non-cognitive Attributes

Tertiary institutions have repeatedly expressed interest in the possible contribution of noncognitive measures to admissions processes but these have not been conclusively proven practical (Cronbach, 1990).

Studies conducted in America have revealed that non-cognitive factors are useful for identifying learners likely to experience difficulty in higher education and for predicting attrition or retention (Pickering, Calliotte & McAuliffe, 1992; Tracey & Sedlacek, 1987) especially for specially admitted learners (Richardson & Sullivan, 1994; Sedlacek & Webster, 1989; Ting, 1997; Tracey & Sedlacek, 1985), and for international learners (Boyer & Sedlacek, 1988).

South African research on non-cognitive aspects includes consideration of personality factors and environmental conditions that impact upon academic performance (Louw, 1994; Schutte, 1994; Mollendorf & Sauer, 1990). This incorporates interests, attitudes toward and habits of studying, and motivation (Jacobs, 1995; Louw, 1992; Mollendorf & Sauer, 1990; Schutte, 1994). In addition, aspects such as self-efficacy, academic self-concept and internal locus of control seem to impact upon academic performance (Schutte, 1994; Venter, 1993).

On the basis of the American results, and what is known about the relationship between non-cognitive aspects and academic performance in South Africa, it is conceivable that non-cognitive variables might be especially useful for predicting academic success for historically disadvantaged learners in the South African context. These factors can be tapped either by using a questionnaire or conducting an interview with prospective learners.

Biographical Information

Background information pertaining to an individual can be useful as an admissions tool, and has been utilized in many selection procedures for predicting performance, specifically in industry and education (Melamed, 1992; Mitchell, 1994; Nickels, 1994; Schmitt & Pulakos, 1998; Stokes, 1994). Such biodata includes the following (Nickels, 1994):

1. History, namely the past behaviours and experiences of the individual;
2. Methodology, which process which helps ensure the information is accurate by focussing on (a) external, observable events, (b) objective, factual

information, (c) first-handed accounts of the individual, and (d) discrete or specific events; and

3. Verifiability, which refers to the degree that the information can be confirmed, taking into consideration the legal and moral issues surrounding the content, including (a) controllability, which is the extent to which events were choices of the individual, (b) equal access, which is the degree to which the individual had opportunities available to all people, (c) relevance of the information for the situation, and (d) invasiveness, which refers to the extent to which content of the information constitutes an invasion of privacy.

The difficulty of using biodata lies in the decision as to which biographical information is necessary (Nickels, 1994; Schmitt & Pulakos, 1998) and legal considerations (Stokes, 1994), but recent research has indicated that the use of this information contributes to the validity and fairness of admission procedures (Mitchell, 1994; Schmitt & Pulakos, 1998; Stokes, 1994). However, it has also been indicated that such methods show differential validity for certain ethnic groups (Schmitt & Pulakos, 1998), and that the predictive validity decays over time when determining future academic success (Melamed, 1992).

Research in South Africa has indicated that background factors do influence performance at university level, and this incorporate aspects such as the secondary school attended, gender, age, first language, ethnic group membership, religion, place of residence, and parental education level (Jacobs, 1995). The validity of utilizing biographical information as part of admissions in South Africa therefore seems plausible (Nelson & Rainier, n.d.; Rainier, 1995). Biographical information can also be obtained either by using a questionnaire or conducting an interview with prospective learners.

A Paradigm Shift

Although research has revealed and confirmed the best predictors of academic success, admissions policies cannot simply be based on these because of the history of this country (Badenhorst, Foster & Lea, 1990). What is required is a change in mindset, a transformation in the perception of the goals of admission and the purposes of testing in admissions programmes.

It is obvious that there needs to be a move away from admissions testing programmes that inherently have a gate-keeping function, that being to determine qualifications for positions or training (Spolsky, 1997), and a move towards

placement assessment that is developmentally-focussed in terms of providing supplemental or remedial instruction or advancement, depending on the functional level of the individual's skills (Sawyer, 1996).

Admission has typically involved the grouping of learners of varying abilities into alternative programmes containing different educational content. Placement testing typically refers to the positioning of learners at an optimal point in an instructional sequence based on their degree of knowledge about a subject. Placement criteria tend to be dependent on test scores or a pattern of test scores in particular subject matter (The College Board and Educational Testing Service, n.d.).

Foxcroft (1999) mentions the following ways in which placement assessment fulfills an enabling function:

1. It assists with the appropriate placement of learners into degree programmes or programmes that offer intensive academic development by linking the individual's competencies, strengths, weaknesses and interests with programme entry requirements;
2. It helps to individualize a learner's programme that may serve the purpose of being either remedial (e.g., by suggesting that an individual spread their first year over two years and incorporate specific academic development modules during that time) or accelerative (e.g., by planning with talented individuals how they could complete their degree in a shorter period of time than usual). Even more appropriate career planning can be recommended if it becomes clear that the learner's choice of degree programme does not match up with their talents and interests. Such individualization also represents an excellent expression of a learner-centered, developmental, outcomes-based focus;
3. It incorporates the recognition of prior learning (RPL) of applicants for undergraduate degree programmes who do not meet the formal qualification requirements or of applicants who wish to enroll for undergraduate degree programmes and receive accreditation for modules completed at other tertiary education institutions. An assessment of whether the learner's profile on the battery matches up with the expected generic competencies for a particular programme contributes to one of the aspects of evidence that must be collected in the recognition of prior learning process; and
4. It helps to set benchmarks for secondary school programmes by providing

feeder high schools with feedback about the match of their pupils with entry-level benchmarks set by the tertiary institution. This feedback could be used in a positive way to revise and improve school curriculums, with additional assistance from tertiary institutions on the form of workshops to guide the alignment of curricula with higher education requirements. A high school version of the placement assessment battery could be administered to pupils from Grade Nine onwards, which would enable them to gauge whether their learning development is at an appropriate level for entering tertiary studies, and the areas they might need to develop their competencies further. This would also foster the opportunity for exceptionally talented secondary school pupils to register for degree modules while they are still at high school. Thus placement assessment fulfils a developmental purpose in the wider community.

The paradigm shift from perceiving admissions decisions as selection processes rather than as placement systems is in line with the policies that are already in place in tertiary institutions in South Africa. Placement systems are typically comprised of two components: (a) assessment assists the estimation of a student's probability of succeeding first-year university, and (b) instruction in which underprepared individuals are given an opportunity to master the skills and knowledge required for success in standard courses, and advanced learners are provided with the opportunity of enrolling in a higher level course (Sawyer, 1996).

Most South African universities have implemented either some form of bridging or foundation courses or community college programme that is related to their alternative or revised admissions procedure and forms part of their academic support for educationally disadvantaged learners or applicants who do not meet the minimum requirements for admission. Such courses either provide learners with the opportunity to improve their basic skills in identified areas of cognitive ability (English and Mathematics) or to improve in areas directly related to their chosen field (Watson, 1997).

Admissions policies that are aimed at increasing access and diversity within tertiary education institutions while maintaining quality standards should be essentially geared toward placement, and are in line with the outcomes delineated by the National Plan for Higher Education of February 2001. Having mentioned this, it

should be borne in mind that this study was approached from a placement paradigm rather than that of the more narrowly focused selection or admission paradigm.

The White Paper on Higher Education (1997) and the National Plan for Higher Education (2001) stipulate that innovations and technological improvements should be incorporated in higher education research and training, and it is acknowledged that the efficiency and accuracy of admission and placement procedures are largely dependent on the application of advances in theory and technology in psychometrics and edumetrics. The following two chapters describe the main developments in theory and technology within these disciplines.

CHAPTER THREE: ADVANCES IN PSYCHOMETRICS: ITEM RESPONSE THEORY

Psychological and educational testing has advanced considerably and undergone many changes during the past few decades (Hambleton, Zaal, & Pieters, 1991). One of the main changes has been the transition from the use of classical to modern models and methods in test theory and development (Hambleton & Slater, 1997).

It is important to understand how these advances in theory have been applied in test development and how they are utilized to improve the efficiency and effectiveness of testing procedures for decision-making, especially in the context of entrance to university education.

This chapter contains a description and explanation of Item Response Theory (IRT). It begins with a brief description of the reasons for and the historical development of IRT, covers an explanation of concepts and aspects of IRT and the expansion of the theory, and ends with a description of its areas of application and critique.

Item Response Theory

A fundamental assumption of measurement theory and practice is that there exists an attribute that, although not directly observable, can be measured. This attribute is thus commonly referred to as a latent trait or ability (Hambleton & Slater, 1997; Hashway, 1998; Hulin, Drasgow & Parsons, 1983; Huysamen, 1979). Any instrument used for the measurement of a latent trait or ability can only obtain a sample of behaviour, presumed to reflect the attribute, that is somehow quantified in order to obtain a numerical score (Lord, 1980).

Classical Test Theory (CCT) constitutes the origins of measurement theory in practice, but it contained inherent limitations in its application, which resulted in the formulation of a new theory, namely, Item Response Theory (IRT).

Classical Test Theory (CTT) and It's Shortcomings

CTT, which was pioneered by Spearman in his work related to intelligence in the early 1900's (Baker, 1992; Embretson, 1996; Van der Linden & Hambleton, 1996), utilizes correlational concepts (Baker, 1992; Hambleton & Slater, 1997; Lord, 1980), which were dominant at that stage as a result of Pearsonian statistics (Baker, 1992). Gulliksen's classic work, published originally in 1950, presents a

comprehensive outline of CTT (Baker, 1992; Embretson, 1996). Traditionally, CTT was the only practically viable measurement model available to behavioural scientists (Weiss & Yoes, 1991) and it has served psychometrics well over many decades (Embretson, 1996; Hambleton & Slater, 1997).

CTT assumes a linear model in which an observed test score is comprised of two major components, namely, a true score and an error score. This is typically represented as $X = T + E$, where X is the observed test score, T is the true score, and E is an error component (Barnard, n.d.; Lord, 1980; Van der Linden & Hambleton, 1997).

The true score is defined as being the examinee's expected score across infinite replications of parallel forms of a test measuring a particular construct. The error score is the difference between the true score and the observed test score (Barnard, n.d.; Hambleton & Slater, 1997; Lord, 1980; Van der Linden & Hambleton, 1997). Every effort is made to reduce both random and systematic errors during the process of test construction and administration in order that test and true scores may be close, and reliability and validity increased (Hambleton & Slater, 1997).

CTT is based on weak assumptions (i.e., those that can be easily met) (Hambleton & Swaminathan, 1985; Huysamen, 1979; Lord, 1980). In addition, there are a number of well-documented shortcomings in this traditionally applied model. One limitation relates to the sample dependence of item statistics, such as difficulty and discrimination indices. Samples comprised of individuals with higher than average levels of population ability will thus yield higher difficulty levels. Also, discrimination indices tend to be higher when estimated from a sample that is heterogeneous in ability and lower when the sample is homogeneous in ability, due to the effect of group heterogeneity on correlation coefficients. Ultimately, such item statistics are useful in test development only when the examinee population is similar to the sample of examinees for which they were obtained (Barnard, n.d.; Hambleton, 1989, 1990, 1995, 1996; Hambleton & Slater, 1997; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Weiss & Yoes, 1991; Weiss, 1995). This impacts negatively upon test reliability and validity as well (Hambleton, 1990, 1995; Hambleton & Slater, 1997; Hambleton & Swaminathan, 1985).

Another shortcoming of CTT concerns the fact that scores are test dependent. Observed and true scores vary according to changes in difficulty levels (Barnard, n.d.; Hambleton, 1990; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan

& Rogers, 1991; Weiss, 1995). Comparisons among examinees on some ability are usually dependent on test scores that are derived from the administration of the same or parallel forms of an instrument. However, many achievement and aptitude tests are geared toward average level ability individuals, and do not provide very precise ability estimates for either very low or very high ability level individuals (Hambleton, 1990; Hambleton & Swaminathan, 1985). Slight or major variations in difficulty therefore have to be adjusted for by means of complex equating procedures (Hambleton, 1989, 1990, 1995; Hambleton & Slater, 1997; Hambleton & Swaminathan, 1985).

The definition of parallel forms, which is fundamental to the CTT concept of reliability, is questionable. Parallel measures are difficult to attain in practice for a number of reasons. Usually, parallel versions of instruments are not even attempted, and the use of nonparallel tests that are assumed to be equivalent result in inaccurate reliability estimates, standard errors of measurement, and test length required for achievement of the desired reliability (Hambleton, 1989, 1995; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991).

The assumption of equal errors of measurement for all examinees is problematic in that the consistency with which individuals perform tasks tends to vary with ability (Barnard, n.d.; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Weiss, 1995). It can be conceded that errors of measurement on a difficult test are greater for examinees with low ability than for those with average and high ability, but violations of the assumption of equal errors of measurement are the rule in the classical model. Although such violations do not necessarily detract from the general utility of CTT, and there is a solution provided within the framework, numerous shortcomings are not addressed, and models where this assumption is not made, are preferable (Hambleton 1989, 1995).

A fifth and final shortcoming of this model, concerns the fact that it is test-based, and thus offers no basis for determining how an individual might perform when confronted with a particular item. An estimate of the probability of an examinee's response is quite valuable in adaptive testing and when wanting to predict test score characteristics in a population or to design tests with certain characteristics for a special target population (Barnard, n.d.; Hambleton, 1990; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991).

There are other aspects of measurement relating to test design, equating, and identification of biased items, for which classical test theory and associated procedures have been unable to provide satisfactory solutions (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991). Item response theory (IRT) purports to overcome these and other limitations of traditional models.

Item Response Theory (IRT): An Historical Overview

IRT has also been termed latent trait theory (LTT) and item characteristic curve theory (Anastasi, 1988; Anastasi & Urbina, 1997; Hambleton & Swaminathan, 1985; Weiss, 1983; Weiss & Yoes, 1991), but the term "item response theory (IRT)" will be utilized for the purposes of this document.

In terms of a definition, IRT is a statistical theory, comprised of a family of mathematical models, that enables the expression of the probability of a particular response to an item as a function of the ability of the test taker, and of certain characteristics of the item (Barnard, n.d.; Hambleton, 1989, 1995; Hambleton & Slater, 1997; Hashway, 1998; Huysamen, 1979; Lord, 1980; Van der Linden & Hambleton, 1997; Wainer & Mislevy, 1990; Weiss, 1995).

The historical roots of item response theory lie in the work of Binet and Simon, who were the first to plot performance levels against an independent variable, and use these in test development (Baker, 1992; Hambleton, 1986; Hambleton & Swaminathan, 1985; Van der Linden & Hambleton, 1997; Weiss & Yoes, 1991). Richardson, Mosier, Lawley, Tucker, Guttman and Lazarfield contributed greatly to the start of the development of IRT in the 1930's and 1940's (Baker, 1992; Hambleton & Swaminathan, 1985; Hulin, Drasgow & Parsons, 1983; Trabin & Weiss, 1983; Van der Linden & Hambleton, 1997; Weiss, 1983; Weiss & Yoes, 1991).

Rasch, Lord, and Birnbaum are also well known names in the history of IRT. They made significant contributions in terms of developing the three basic models for addressing dichotomously scored items of unidimensional achievement and aptitude tests in the 1950's and 1960's (Hambleton & Swaminathan, 1985; Hulin, Drasgow & Parsons, 1983; Van der Linden & Hambleton, 1997; Weiss, 1983; Weiss & Yoes, 1991).

Wright, Samejima, Bock, Fischer, Baker, Weiss, and their colleagues are among the names associated with IRT during the 1970's and 1980's (Hambleton & Swaminathan, 1985).

Perusal of the literature on IRT reveals that another prominent name, especially during the 1980's and 1990's, is Hambleton, who has written many general articles and focussed on particular aspects of the subject, alone and in conjunction with various associates.

Features and Assumptions of IRT

There are three principal postulates of item response theory, and these relate to dimensionality, local independence, and the mathematical forms of item characteristic curves (ICC's) (Hashway, 1998; Hambleton, 1995; Weiss & Yoes, 1991)

Dimensionality

Traditionally, the assumption relating to latent space in IRT models is that only one ability or trait accounts for performance on a test. This is known as unidimensionality. Here all items in a test are considered homogeneous (i.e., they measure a single attribute), and examinee performance on a test reflects the individual's position on one underlying trait or ability (Anastasi, 1988; Anastasi & Urbina, 1997; Barnard, n.d.; Hambleton, 1989, 1990, 1995, 1996; Hambleton & Slater, 1997; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Hashway, 1998; Hulin, Drasgow & Parsons, 1983; Huysamen, 1979; Kolen & Brennan, 1995; Osterlind, 1983; Wainer & Mislevy, 1990; Weiss & Yoes, 1991). Deviation from this traditional assumption, where more than one ability is considered necessary to account adequately for test performance, is known as an assumption of multidimensionality. Certain IRT models that expanded from the original ones to be discussed later in this document are based on the assumption of multidimensionality (Hambleton, 1989; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Van der Linden & Hambleton, 1997; Weiss & Yoes, 1991).

Local Independence

Unidimensionality implies the assumption of local independence (Hambleton, 1989; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Huysamen, 1979). It refers to the idea that items in a test are statistically independent (i.e., uncorrelated) for individuals with the same ability or trait level. This means that the probability of a correct response of an examinee to a particular item is not affected by the required responses to other items in a test (Barnard, n.d.; Hambleton, 1989; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Kline, 1993; Kolen & Brennan, 1995; Lord, 1980; Osterlind, 1983; Weiss, 1995;

Weiss & Yoes, 1991). This assumption does not, however, imply that test items are not correlated over the total group of examinees (Barnard, n.d.; Hambleton, 1989; Hambleton & Swaminathan, 1985; Huysamen, 1979; Kline, 1993; Lord, 1980). Factor analytic procedures can be utilised to investigate the assumptions of unidimensionality (Hambleton, 1989; Hambleton & Swaminathan, 1985; Hashway, 1998; Lord, 1980; Weiss & Yoes, 1991) and local independence (Hambleton, 1989; Hambleton & Swaminathan, 1985; Huysamen, 1979; Kline, 1993; Lord, 1980; Weiss & Yoes, 1991).

Mathematical Forms of Item Characteristic Curve's (ICC's)

This assumption posits that the relationship between an individual's performance on each item and the trait measured by a test can be described by means of a monotonically increasing function. It provides the probability that examinees with different ability levels will answer a particular item correctly. Those higher on the ability continuum will have higher probabilities of answering items correctly than those with lower ability levels. The graphical depiction of this relationship between performance on an item and ability level is known as an item characteristic curve (ICC) (Baker, 1985, 1992; Hambleton, 1989, 1990, 1995, 1996; Hambleton & Slater, 1997; Hambleton, Zaal & Pieters, 1991; Hashway, 1998; Huysamen; 1979; Kolen & Brennan, 1995; Osterlind, 1983; Weiss, 1983; Weiss & Yoes, 1991).

The ICC is non-linear (Hambleton, 1989; Hambleton & Swaminathan, 1985; Huysamen, 1979), and typically S-shaped (Hambleton, 1996; Hambleton & Slater, 1997; Hulin, Drasgow & Parsons, 1983; Huysamen, 1979; Osterlind, 1983). ICC's have also been termed "trace lines" (Lord, 1980; Hulin, Drasgow & Parsons, 1983; Wainer & Mislevy, 1990), "item characteristic functions" (Hambleton, 1995; Hambleton & Swaminathan, 1985; Hashway, 1998; Weiss, 1995; Weiss & Yoes, 1991), and "item response functions" (Barnard, n.d.; Hambleton, 1995; Lord, 1980; Wainer & Mislevy, 1990). Hambleton and Swaminathan (1985) note that the term "ICC" is associated with unidimensional IRT models, whereas the term "item characteristic function" is associated with multidimensional IRT models. However, the term "item characteristic curve (ICC)" will be used throughout this document. ICC's are usually described by one-, two-, or three-parameters (Hambleton, 1989, 1990, 1995, 1996; Hambleton, Swaminathan & Rogers, 1991; Hambleton & Slater, 1997; Osterlind, 1983).

Advantages Stemming From the Features of IRT

When the assumptions of IRT can be met, and there is goodness of fit between the IRT model and the test data, certain advantages are obtained.

Invariance. This is a property manifested in two aspects:

1. Sample-free parameter estimates where item statistics are independent of the particular sample of examinees drawn from a population of interest and used in model parameter estimation (Anastasi, 1988; Anastasi & Urbina, 1997; Barnard, n.d.; Hambleton, 1986, 1989, 1990, 1995, 1996; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991). This concept is also known as "person-free parameter estimates" (Hambleton 1989, 1995).
2. Item-free ability estimates where ability estimates for examinees are defined relative to the pool of items from which the test items are drawn rather than the particular sample of items included in a test. Each examinee has the same ability across the various samples of test items, despite differences in estimates that result from measurement errors and the selection of more or less suitable items. Examinees can therefore be compared, although there might have been differences in sets of test items (Barnard, n.d.; Hambleton, 1986, 1989, 1990, 1995, 1996; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991).

Precision. A statistic is provided that indicates the precision of the estimate of every examinee's ability. This statistic can vary for different examinees (Ham bleton, 1989; Hambleton & Slater, 1997; Hambleton & Swaminathan, 1985) and is based on examinee ability and the number and statistical properties of test items (Hambleton, 1989). This is a direct way of estimating measurement error for every ability level, and is superior to reporting one estimate of error (SEM in CTT) and applying it to all examinees without taking into consideration ability levels (Hambleton, 1996).

There are additional advantages that are more generally associated with IRT.

No reliance on parallel-forms reliability. The concept of parallel-forms reliability associated with CTT is replaced by the concept of statistical estimation and standard errors associated with this (Hambleton, 1989; Hambleton & Swaminathan, 1985).

A common scale. Items and examinees are reported using one scale (Hambleton, 1989; Hambleton & Slater, 1997; Hambleton & Swaminathan, 1985),

which provides great possibilities in terms of facilitating test development, and reporting and interpretation of test scores (Hambleton & Swaminathan, 1985).

Item information functions. These functions indicate the contribution of items to measurement precision along the ability continuum. These can be considered the "building blocks" of test development (Hambleton & Slater, 1997).

Such advantages contribute to the accuracy and efficiency of decision-making, especially when one considers the need for cost-effectiveness in terms of finances and time where higher education admissions decisions are concerned.

Models of IRT

The original IRT models that were developed in the early years of IRT presented interesting quantifications of the relationships between response characteristics and the latent trait investigated. Over time, others replaced these models, as their applicability was limited (Hulin, Drasgow & Parsons, 1983). Three models of IRT were developed and were widely researched and extensively used to solve numerous practical measurement problems (Hambleton & Swaminathan, 1985; Lord, 1980; Van der Linden & Hambleton, 1997). The specification of the mathematical form of the ICC, and the corresponding number of parameters required to describe them determines the particular IRT model (Hambleton, Zaal & Pieters, 1991). These various models addressed dichotomously scored data (i.e., data scored as either correct or incorrect) (Van der Linden & Hambleton, 1997), and are reviewed at a basic level in this chapter.

Prior to commencing with these outlines and explanations, however, it is necessary to explain certain concepts and define certain symbols typically encountered in IRT, and especially in discussions on ICC's. The Greek letter theta (θ) is used to refer to a generalized value along the latent trait continuum (i.e., the ability score) (Baker, 1985; Hulin, Drasgow & Parsons, 1983). At each ability level, there is a particular probability that a person with that ability will respond correctly to an item, and this is represented by $P(\theta)$ (Baker, 1985). In models addressing dichotomously scored items, a score of 1 is usually assigned to a correct response, and a score of 0 to an incorrect response (Hulin, Drasgow & Parsons, 1983).

There are three important technical properties utilized in the description of ICCs, depending on the particular model being used. The first is the difficulty of the item, represented by "b". In IRT, difficulty tends to be a location index because it describes where an item functions along the ability scale, (i.e., it indicates the position of the

curve in relation to the ability scale). The second is discrimination, which is represented by "a". This property reflects the steepness of the curve in its middle section (i.e., at the point of inflection). The steeper the curve, the better the item discriminates between individuals with abilities lower than the item location, and those with abilities above the item location. The final property is called the pseudo-guessing parameter, and it represents the probability of a correct response for individuals with low ability levels. It is represented by "c" (Weiss & Yoes, 1991).

These descriptors not only describe the form of the ICC, but are also used to discuss the technical properties of an item (Baker, 1985). It should be noted that the characters utilized to symbolize these properties are not standard statistical notation. In certain articles and books, difficulty is symbolized as β, discrimination is symbolized as α, whereas the symbol for the guessing parameter remains the same (Baker, 1992). For the purposes of this document, standard statistical notation will not be used because the literature on IRT written for the less statistically inclined does not use such notation.

Earliest IRT Models

One of the original models developed in the 1940's was Guttman's perfect scale, the curve of which is actually a step function. This type of scale is considered deterministic (i.e., error-free) (Weiss & Yoes, 1991) because it has very stringent requirements to which data will seldom fit (Hambleton, 1989; Hambleton & Swaminathan, 1985; Hulin, Drasgow & Parsons, 1983). Probabilities of correct responses are either 0 or 1, and the critical ability level is the point at which these probabilities change from 0 to 1 (Hambleton, 1989).

Two stochastic (i.e., probabilistic) models, developed in the 1950's, are the latent-distance and the latent-linear models. The latent-distance model retains the step-function of Guttman's perfect scale, but probabilities of correct and incorrect responses generally differ from 0 and 1. This model has been utilized by Lazarfield and Henry (1968, in Hambleton, 1989 and Hulin, Drasgow & Parsons, 1983) in the measurement of attitudes but it is not usually likely for social science data sets to be reasonably represented by this model. People rarely behave as consistently as the model depicts. The discontinuity at the point for the critical ability level, and the flatness of the curve both before and after this point, seem implausible (Hulin, Drasgow & Parsons, 1983).

The latent-linear model avoided the step-function's discontinuity because it assumes that the probability of a correct response to an item is proportional to a person's position on the underlying latent trait. In this model, ICC's vary in their intercepts and their slopes to reflect the fact that items vary in difficulty and discrimination, respectively. Changes in either of these parameters result in changes in the other (Hambleton, 1989). However, according to this model, it is possible for individuals to have negative probabilities of a positive response, and also to have probabilities greater than unity, depending on their level of ability. This creates theoretical problems because the distribution of the ability is such that there is no person for whom the probability of an incorrect response is less than zero or for whom the probability of a correct response is greater than unity (Hambleton, 1989; Hulin, Drasgow & Parsons, 1983).

Normal-ogive Models

The normal-ogive model was the first of the empirical models to be developed (Hulin, Drasgow & Parsons, 1983), and it postulated a normal cumulative distribution function for an item (Baker, 1997; Van der Linden & Hambleton, 1997). The normal-ogive model is vital in statistical theory, thus it is not surprising that the normal ogive has been used as a model (Baker, 1992). Although he was not the first to use this model, Lord was the first to propose coherent item response models in which the ICC took the form of one-, two-, and three-parameter normal ogives (Hambleton & Swaminathan, 1985). The following two equations are for the two- and three-parameter normal ogives, respectively:

$$P_i(q) = \int_{-\infty}^{a_i(q-b_i)} \frac{1}{\sqrt{2p}} e^{-z_2/2} dz.$$

$$P \equiv P_i(q) = c + (1-c) \int_{-\infty}^{a_i(q-b_i)} \frac{1}{\sqrt{2p}} e^{-z_2/2} dz.$$

In these equations, $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ answers item i correctly. Difficulty and discrimination parameters characterizing this item are b_i and a_i, respectively. The pseudo-guessing parameter in the second equation is represented by c. The *z* is a normal deviate from a distribution with a mean of b_i and standard deviation of $1/a_i$. This results in a monotonically increasing function of ability. The difficulty index (b_i) represents the point on the ability scale where the individual has a 50 percent probability of answering item i correctly, and the discrimination index (a_i) is proportional to the slope of $P_i(\theta)$ at the point $\theta = b_i$.

Transformation of ability scores and item parameter estimates to more convenient scales is common in order to avoid decimals and negatives. Transforming ability scores so that they have a mean of zero and a standard deviation of one results in values of b that typically vary from about -2 to +2. Values that are closer to the negative limit correspond to items that are very easy, and values that are closer to the positive limit correspond to items that are very difficult. Values of a theoretically range from $-\infty$ to $+\infty$, but in practice negatively discriminating items tend to be discarded from tests. It is also unusual to obtain values of *a* larger than two, thus the usual range for this parameter tends to be zero and two. High values tend to result in very steep curves and low values tend to result in flatter curves (Hambleton & Swaminathan, 1985; Hulin, Drasgow & Parsons, 1983).

In a three-parameter normal ogive, c is the pseudo-guessing parameter, b is the location parameter, indicating difficulty level of the item, and the point of inflection is a function of a, which is proportional to the slope of the curve at the point of inflection, and represents the discriminating power of the item, or degree to which response varies with ability level (Lord, 1980).

One favourable aspect of the normal ogive models is that they allow for the interpolation between pairs of empirical ICC points, and for extrapolation beyond the range of empirical ICC points (Hulin, Drasgow & Parsons, 1983). These models are not without criticism, however, and the main problem associated with them relate to the assumption that the characteristic function corresponding to each item is actually the same for all items. The implication is thus that an individual with a particular level of ability has the same probability of responding correctly to all items (Hashway, 1998). In reality, it is conceivable that for all people, there will be items that are more or less difficult than others, and therefore, the probability of responding correctly to a group of items is a function of both the individual's ability level, and the item under consideration (Hashway, 1998). Also, the equations for the normal-ogive model involve integration, which led to the development and increased favour of logistic models in IRT (Hambleton, Swaminathan & Rogers, 1991; Hulin, Drasgow & Parsons, 1983), due to the explicitness of their relation to item and ability, and their important statistical properties (Hambleton, Swaminathan & Rogers, 1991).

Rasch One Parameter Logistic Model

The one parameter logistic model is one of the most widely used in IRT (Baker, 1992; Hambleton, Swaminathan & Rogers, 1991). It was developed independently in

the 1960's by a Danish mathematician named Georg Rasch, and is thus also commonly known as the Rasch model (Baker, 1985, 1992; Hambleton, 1989; Hambleton & Swaminathan, 1985; Wainer & Mislevy, 1990). The equation for this model is as follows:

$$P_i(q) = \frac{e^{(q-b_i)}}{1+e^{(q-b_i)}}, i = 1,2,3,...,n$$

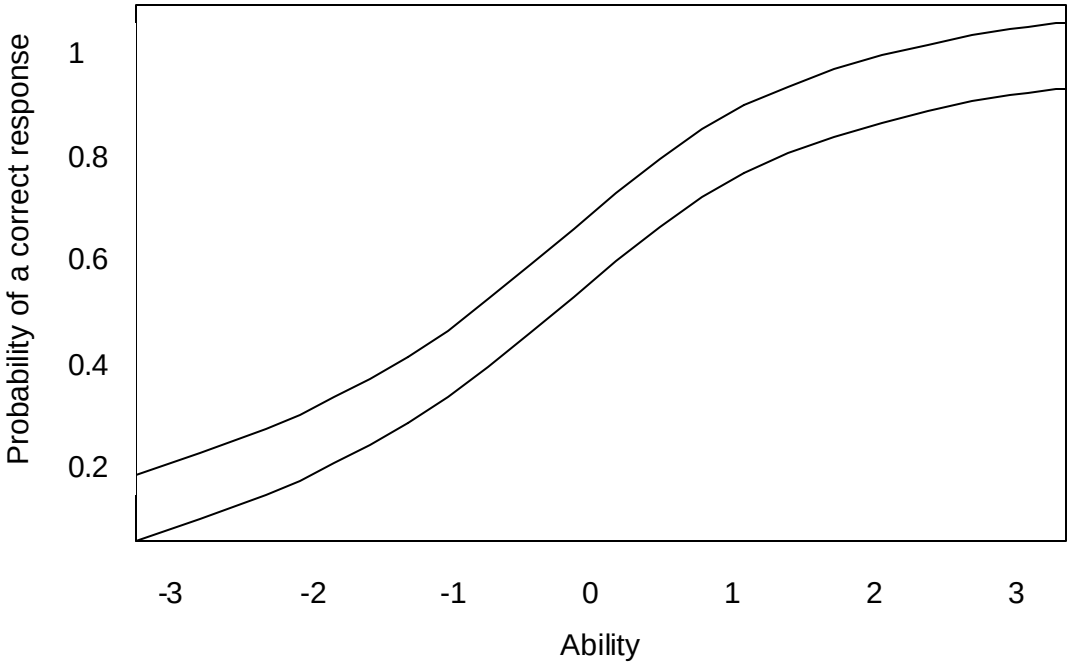
In this equation, $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ answers item i correctly, b_i is the difficulty parameter for item i, n is the number of items in the test, and e is a transcendental number, which is a constant value of 2.718 (correct to three decimals). The difficulty parameter is the point on the ability scale where the probability of a correct response to the item is 0.5. The greater the value of this parameter, the greater the ability level required for an examinee to have a 50 percent chance of answering the item correctly, and thus, the harder the item. Difficult items are positioned to the right (i.e., the higher end) of the ability scale, whereas easy items are positioned to the left (i.e., the lower end) of the ability scale.

The theoretical range of values of b_i is $-\infty \leq b_i \leq +\infty$, but in practice, Baker (1985; 1992) states that transformation of the ability values of a group so that the mean is 0 and the standard deviation is 1, results in values of b_i that typically vary between -3.0 and +3.0. Other sources state that the typical range of values for this parameter is between -2.0 and +2.0 (Hambleton, 1989; Hambleton, Swaminathan & Rogers, 1991). According to this latter view, values that are near -2.0 correspond to very easy items, and those values near +2.0 correspond to very difficult items relative to the group of examinees.

On a graphical depiction of the transformed scale, such as in figure 1, the slope remains the same for different items, but the location of the item varies (Baker, 1985; Hambleton, Swaminathan & Rogers, 1991). In this model, the only characteristic assumed to influence examinee performance is item difficulty, thus, this model is called the one-parameter logistic model. There is no parameter in the equation that reflects discrimination, which is, in effect, equivalent to the assumption that all items discriminate equally among examinees. The lower asymptote of the ICC is zero, which indicates that lower ability examinees have no probability of answering the item correctly, thus no allowance is made for the possibility of guessing (Hambleton, Swaminathan & Rogers, 1991). In other words, in this model, only the difficulty

parameter can adopt different values, discrimination is fixed at one for all items, and guessing is set at zero (Baker, 1985; Hambleton, 1989; Hambleton, Swaminathan & Rogers, 1991; Hulin, Drasgow & Parsons, 1983).

Figure 1: One-parameter logistic ICCs



This model has certain properties that make it an attractive option for test analyses, namely, (a) it is easy to use because it involves only one item parameter, (b) there are fewer parameter estimation problems than with more general models, and (c) the property of specific objectivity is obtained, which allows complete separation of item and ability estimation. In terms of this last point, the result is that ability parameters can be estimated without bias and independently of the items selected from those that fit the model. Also, item parameters can be estimated without bias and independently of the distribution of abilities in the sample of individuals drawn from the population for whom the model fits (Hambleton, 1989).

Despite the simplicity of this model, the assumptions are limited and their appropriateness depends on the nature of the data and the importance of the intended application. It can be applied to data that has been carefully pretested and selected (e.g., very easy tests constructed from a homogeneous bank of items) (Hambleton, Swaminathan & Rogers, 1991; Hulin, Drasgow & Parsons, 1983). It tends to be difficult to find items that fit this model and if the model does not fit the desirable properties mentioned are not obtained (Hambleton, 1989).

Two Parameter Logistic Model

Lord developed a two-parameter item response model based on the cumulative normal distribution in the early 1950's, after which, in the late 1960's, Birnbaum changed the form of the curve by substituting the two-parameter logistic function for the two-parameter normal ogive function (Hambleton, Swaminathan & Rogers, 1991). The equation for this model is as follows:

$$P_i(q) = \frac{e^{Da_i(q-b_i)}}{1+e^{Da(q-b_i)}}, i = 1,2,3,...,n.$$

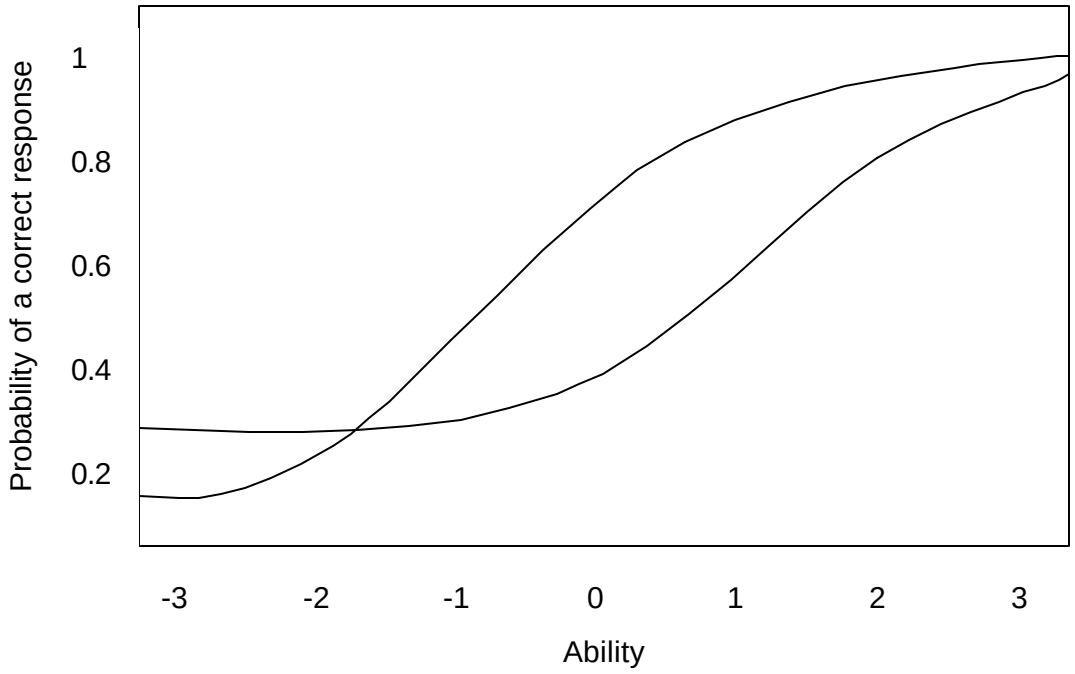
In this equation, $P_i(\theta)$ and b_i hold the same definitions as for the one-parameter model equation. The difference between the one-parameter model equation and this equation is the presence of two additional elements. This includes D, which is a scaling factor that was introduced to keep the logistic function as close as possible to the normal ogive function. It has been demonstrated that when D = 1.702, values of $P_i(\theta)$ for the two-parameter normal ogive and logistic models differ in absolute value by less tan 0.01 for all values of θ.

Secondly, a discrimination parameter has been added and is represented in the equation as a_i. The slope of the curve changes as a function of ability level and a maximum value is attained when the ability level is equal to the difficulty of the item. Thus, the discrimination parameter does not represent the general slope of the ICC, but is rather proportional to the slope of the ICC at $\theta = b_i$. In actuality, the slope at $\theta = b_i$ is a/4, but considering a_i to be the slope at b_i is an acceptable approximation that allows for easier practical interpretation. Theoretically, the range for this parameter is $-\infty \leq a \leq +\infty$, but the normal range in practice tends to be $-2.80 \leq a \leq +2.80$ (Baker, 1985). Items with steeper slopes are more useful for separating examinees into different ability levels than items with flatter slopes. The utility of an item for discriminating among examinees near an ability level θ (i.e., those with abilities $\leq \theta$ from those with abilities $> \theta$) is proportional to the slope of the ICC at θ (Hambleton, Swaminathan & Rogers, 1991).

Despite Baker's (1985) statement about the practical range of values for this parameter, other sources note that negatively discriminating items tend to be discarded from tests. It has also been mentioned that it is unusual for there to be values for a_i that are larger than 2, thus the usual range for this parameter tends to be between 0 and 2. As illustrated in figure 2, higher values of a_i result in ICC's that

are very steep whereas lower values result in flatter curves (Hambleton, 1989; Hambleton, Swaminathan & Rogers, 1991).

Figure 2: Two-parameter logistic ICCs



This model does not make allowance for guessing among examinees (Hambleton, 1989; Hambleton, Swaminathan & Rogers, 1991). The assumption of no guessing seems plausible when one considers that for all items for which there is a positive relationship between performance on the item and the ability measured by the test, the probability of responding correctly to an item decreases to zero as ability decreases (Hambleton, 1989).

Three Parameter Model

It is widely accepted that in testing examinees will get items correct by guessing. Neither of the two models already described took this phenomenon into consideration, but in the late 1960's, Birnbaum modified the two-parameter logistic model to include a parameter representing the contribution of guessing to the probability of a correct response. However, it has been noted that certain mathematical properties of the logistic function were lost in the process, and it is thus technically no longer a logistic model (Baker, 1985, 1992).

The equation for this model is as follows:

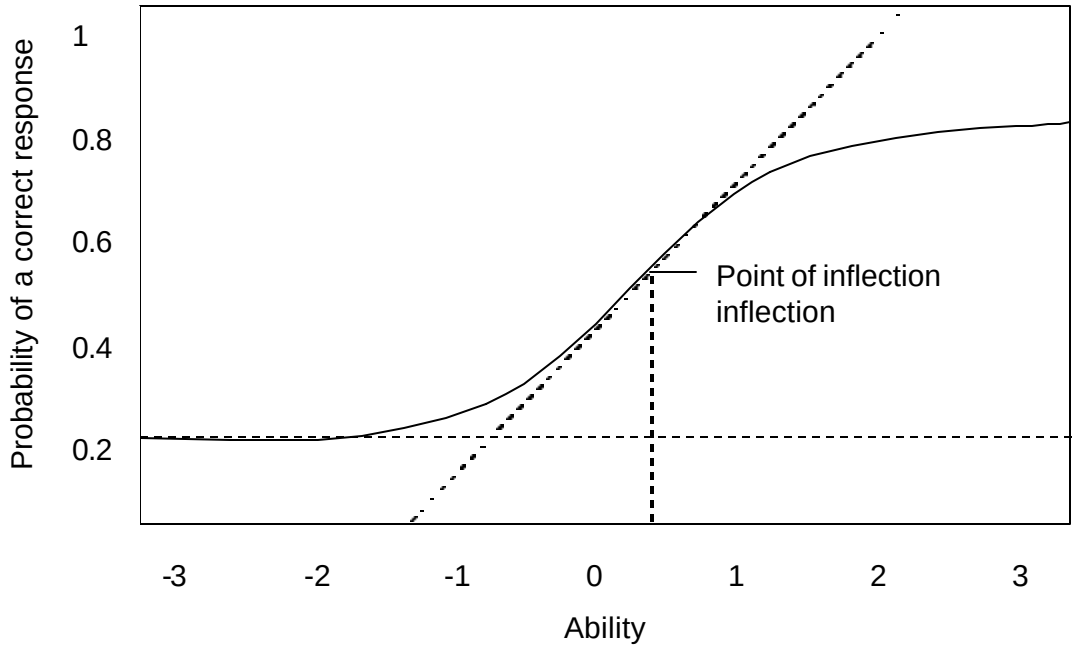
$$P_i(q) = c_i + (1 - c_i)\frac{e^{Dai(q-bi)}}{1+e^{Da(q-bi)}}, i = 1,2,3,..., n.$$

In this equation, $P_i(\theta)$, a_i, b_i, and D hold the same definitions as for the two-parameter model equation. It should be noted that c_i is representative of the probability of responding correctly through guessing alone (Baker, 1985; Kolen & Brennan, 1995). This parameter value does not vary as a function of ability. It represents the probability that low ability individuals will answer an item correctly. Individuals at the lowest and highest ability levels have the same probability of answering the item correctly by guessing (Baker, 1985; Kolen & Brennan, 1995). It is common to refer to this parameter as the pseudo-chance level or pseudo-guessing parameter because where good distractors have been written for multiple-choice items, low ability individuals would actually score higher by randomly guessing the correct answer rather than selecting a plausible distractor (Hambleton, 1989).

Typically, the value of c_i tends to be smaller than the value that would result if low ability individuals were to guess an answer randomly. This value is decided in consideration of the magnitude of the chance that guessing will take place. Without this parameter, individuals with low ability levels would tend to exceed predicted item performance from the best fitting one- and two-parameter models already discussed (Hambleton, 1989). Theoretically, the range of values for this parameter is $0 \leq c_i \leq 1.0$, but in practice, values greater than .35 are considered unacceptable, thus the range $\theta \leq c_i \leq .35$ tends to be utilized (Baker, 1985). See figure 3 for an illustration of three-parameter ICCs.

One impact of the additional parameter is that the definition of the difficulty parameter is changed (Baker 1985; Hambleton, 1989). Rather than b_i being .50 on the ability scale, the probability of a correct response is $(1+c_i)/2$. The probability is halfway between the value of c_i and 1 because c_i provides a floor to the lowest value of the probability of a correct response. Thus b_i defines the point on the ability scale where the probability of a correct response is halfway between this floor and 1. When c_i =zero, then b_i =.50 (Baker, 1985; Hambleton, 1989, Kolen & Brennan, 1995). The discrimination parameter can still be interpreted as being proportional to the slope of the ICC at $\theta = b_i$, but in this model, the value for this parameter is actually $a_i (1 - c_i)/4$ (Baker, 1985). Although these definitional changes for the parameters of difficulty and discrimination are slight, they are important for interpreting test analyses results (Baker, 1985).

Figure 3: Three-parameter logistic ICCs



Interpretation Guidelines for Item Parameter Values

Much of the application of models is computer generated, but it is necessary to have some idea of what the values of parameters mean in order to guide interpretation of information yielded by tests that are based on IRT models.

Item discrimination. This parameter can be described verbally by means of labels that are ascribed to ranges of values of the parameter. Table 1 contains the labels and value ranges for the discrimination parameter applicable only for values derived from a logistic model of IRT. The values must be divided by 1.7 in order to interpret this parameter using a normal ogive model (Baker, 1985)

Table 1: Labels for Value Ranges for Item Discrimination

Label	Value Range
None	0
Very low	.01-.34
Low	.35-.64
Moderate	.65-1.34
High	1.35-1.69
Very High	>1.7
Perfect	Infinity

Note. Adapted from "The Basics of Item Response Theory, " by F. B. Baker, 1985, p. 24. Copyright 1985 by Heinemann Educational Books, New Hampshire.

Item difficulty. In one- and two-parameter IRT models, item difficulty is a point on the ability scale where the probability of a correct response is .50, and in a three-parameter model, it is $(1+c_i)/2$. Interpretation of a numerical value of this parameter is in terms of the place at which the item functions on the ability scale, and the discrimination parameter can be utilized to provide additional meaning to this interpretation. The slope of the ICC is steepest at an ability level corresponding to item difficulty, thus b_i indicates the position on the ability scale where the item functions optimally (i.e., the item distinguishes best between individuals with this level of ability) (Baker, 1985).

Pseudo-guessing. In the three-parameter model, the numerical value of c_i is interpreted directly, because it is a probability (e.g., $c_i = .20$ means that for all ability levels, the probability of answering an item correctly by guessing alone is .20) (Baker, 1985).

Additional points are worth mentioning as a means of summarizing the information about these three models. The slope for the one-parameter model is always the same, and only the location (difficulty level) of the item varies. When utilizing the two- and three-parameter models, the value of the discrimination parameter must be fairly large ((i.e., greater than 1.7) for the curve to be considerably steep. When utilizing one and two-parameter models, a large positive value for the difficulty parameter results in a lower tail of the ICC that approaches zero, whereas when utilizing the three-parameter model, the lower tail approaches the value of the pseudo-guessing parameter. This value is not apparent when the difficulty level is less than zero, and discrimination is less than one, but utilization of a wider range of ability values would cause the lower tail of the ICC to approach the value of the pseudo-guessing parameter (Baker, 1985).

The slope of the ICC is steepest at the level of ability that corresponds to item difficulty, therefore the difficulty parameter indicates the point on the ability scale where the item functions optimally. When utilizing one- and two-parameter models, item difficulty defines the point on the ability scale where the probability of a correct response for individuals with that ability level is .5. However, when utilizing a three-parameter model, this parameter defines the point on the ability scale where the probability of a correct response is halfway between the value of the discrimination

parameter and one. It is only when the pseudo-guessing parameter is zero that these two definitions are equivalent (Baker, 1985).

It has been noted that the one-parameter model is actually a special case of the two-parameter model where the value of the discrimination parameter is set at one (Baker, 1985, 1992; Hambleton, 1995, 1996; Kolen & Brennan, 1995). In the same vein, the one-parameter model can be considered a special case of the three-parameter model where all items are considered to be equally discriminating (i.e., set at one), and the value of the pseudo-guessing parameter is set at zero (Hambleton, 1989, 1995, 1996; Hambleton, Swaminathan & Rogers, 1991; Hulin, Drasgow & Parsons, 1983; Kolen & Brennan, 1995).

Four Parameter Model

Although not commonly used, but still deserving of mention, the four parameter model was developed for situations in which high ability individuals incorrectly answer test items that might even be easy. Reasons for this could be carelessness or a possession of information beyond that assumed by the test item writer.
The equation for this model is as follows:

$$P_i(q) = c_i + (y_i - c_i)\frac{e^{Da_i(q-b_i)}}{1+e^{Da(q-b_i)}}, i = 1,2,3,...,n.$$

In this equation, $P_i(\theta)$, a_i, b_i, c_i, and D hold the same definitions as for the three-parameter model equation. There is an additional parameter, namely y_i, that may assume a value slightly below one, which means that the ICC may have an upper limit less than one. However, this model is most likely of theoretical interest only, as research has been unable to find any practical gains from its utilization (Hambleton, 1989).

Parameter Estimation

In test analysis, an IRT model must be selected for a particular data set, and then it becomes necessary to estimate ability and item parameters. In the models considered thus far, one ability parameter (θ) is estimated for each individual. The parameters that need to be estimated for each item depends upon the IRT model selected (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Hulin, Drasgow & Parsons, 1983).

Successful application of IRT is dependent upon the availability of adequate procedures for estimating model parameters (Hambleton, Swaminathan & Rogers,

1991). Ability and item parameters are usually unknown at some stage of model specification (Hambleton & Swaminathan, 1985). In analyzing test data, a random sample is selected from a target population, and the responses to a set of items are obtained. On the basis of this pattern of responses, parameter estimation is conducted (Hambleton, & Swaminathan, 1985; Kolen & Brennan, 1995; Weiss, 1995).

In practice, there are two principal estimation situations that occur, firstly, estimation of ability with item parameters known, and secondly, estimation of both item and ability parameters (Hambleton, 1989). In some situations, item parameters are assumed to be known. This occurs when items that have been previously calibrated are included in a test. Item parameter estimates derived from earlier analyses are thus treated as the true values (Hambleton, 1989; Hambleton & Swaminathan, 1985).

A number of computer programmes are available for parameter estimation, and different ones utilize different procedures for estimation for one or more of the IRT models discussed (Hambleton, Swaminathan & Rogers, 1991).

Goodness of Fit

IRT has the potential for solving numerous problems in measurement, but its advantages can only be obtained when there is a satisfactory fit between the model and the test data under consideration. Many reported IRT applications have inadequately incorporated the model-data fit aspect, and it's consequences, thus knowledge about the appropriateness of certain IRT models for different applications is not as certain as it appears to be. The utilization of what is now known to be inadequate statistics to conduct goodness-of-fit studies may have resulted in incorrect decisions about the appropriateness of an IRT model applied (Hambleton, 1989; Hambleton, Swaminathan & Rogers, 1991).

The reliance on statistical tests of model fit has been problematic because their sensitivity to sample size is a serious flaw. Nearly any empirical departure from the model in the data being considered will result in a rejection of the null hypothesis of model-data fit if the sample size is large enough. On the other hand, a small sample size is problematic because statistical power is low, thus even large discrepancies in fit might not be detected. Also, estimation errors for parameters tend to be large when sample sizes are small. Sampling distributions for some goodness-of-fit statistics in IRT are not what they have been claimed to be, therefore errors are

possible when such statistics are interpreted according to tabled values of known statistics (Hambleton, 1989; Hambleton, Swaminathan & Rogers, 1991).

Rather than emphasizing the results of significance tests when selecting IRT models, it has been recommended that certain judgements about model-data fit be based on three types of evidence (Hambleton & Swaminathan, 1985):

1. Appropriateness of the assumptions of the model for the test data.
2. Degree to which the expected model properties (i.e., item and ability parameter invariance) are obtained.
3. Predictive accuracy of the model as yielded by real and, if applicable, simulated test data.

There are certain assumptions that must be checked in order to assist model selection. Unidimensionality of the data and non-speededness of the test administration are common to all the models. In addition, the two-parameter logistic model assumes that guessing is minimal, and the one-parameter model assumes that all item discrimination indices are equal. The methods for checking these are not covered here, but are discussed in Hambleton (1989), Hambleton and Swaminathan (1985), and Hambleton, Swaminathan and Rogers (1991).

Ability parameter estimate invariance can be investigated by comparing different samples of test items. Invariance is established when estimates are not excessively different from the measurement errors associated with them. Invariance of item parameter estimates can be investigated by comparing model item parameter estimates obtained in two or more subgroups of the population for whom the test is intended. The resulting plot should be linear and scatter should reflect errors attributable to sample size only. Randomly equivalent samples allows for baseline plots to be obtained (Hambleton, Swaminathan & Rogers, 1991).

There are also numerous methods for checking model predictions that are discussed in Hambleton (1989), Hambleton and Swaminathan (1985), and Hambleton, Swaminathan and Rogers (1991). Much of this process takes place during test development and problems are ironed out prior to the releasing of the test for use in decision-making arenas.

Item and Test Information Functions

A very effective way of describing items and tests, selecting items for a test, and comparing tests is provided by the item information function, denoted $I_i(\theta)$. There are

separate formulae for this function as applicable for one-and two-parameter logistic models, and the three-parameter model.

The role of the difficulty parameter is that more information is obtained when the difficulty value is closer to rather than further from θ. Discrimination must be high for more information to be obtained, and also as the pseudo-guessing parameter value approaches zero (Hambleton, Swaminathan & Rogers, 1991).

The item information function plays an important role in test development and item assessment because they indicate the contribution of items to ability estimation at different points along the ability continuum. However, this is dependent on the item's power to discriminate (i.e., the higher it is, the steeper the slope) and the position at which this contribution is realized to be dependent on the difficulty of the item (Hambleton, Swaminathan & Rogers, 1991).

The utility of this function also depends on the fit of the ICCs to the test data. A poor fit results in misleading statistics and item information functions, and even with a good fit, the value of the item can be limited in all tests if the discrimination value is low and the pseudo-guessing parameter is high. Also, items can provide great information at one end of the ability continuum but be useless elsewhere on the scale (Hambleton, Swaminathan & Rogers, 1991).

The test information function is provided by the sum of the item information functions at θ. Items contribute independently to the test information function, and individual contributions can thus be determined without knowledge of the other items in a test. There is an inverse relationship between the information provided by a test at θ and the precision with which ability is estimated at that point (Barnard, n.d.; Hambleton, 1995; Hambleton, Swaminathan & Rogers, 1991).

An important concept within this is the standard error of estimation, which serves the same role as the standard error of measurement in CTT. It is derived by taking the square root of the item information function (Kline, 1993; Weiss, 1995). The utility of this is that a confidence interval can be established for interpreting the ability estimate. However, the value of the standard error of estimation varies with ability level, and it is the standard deviation of the asymptotically normal distribution of the maximum likelihood estimate of ability for a given true value of θ. The magnitude of this is dependent on the number of items in a test, the quality of items in a test (i.e., in terms of discrimination power), and the match between difficulty of the item and ability of the individual (Hambleton, Swaminathan & Rogers, 1991).

One more point worth mentioning is that the assumption of local independence means that functions for the depiction of ICCs, item information functions and standard errors of estimation are additive. Since a test is a collection of test items, the addition of functions depicting items results in a test response function that can be depicted as a test characteristic curve (TCC). This provides the sum of the probabilities of a correct response to a group of test items, as a function of ability. Dividing this sum of probabilities by the number of items results in the average or expected proportion of correct responses (Barnard, n.d.; Hambleton, 1995, 1996; Weiss, 1995).

Changed Rules of Measurement

The basics of IRT have been briefly discussed, and it seems appropriate to refer back to CTT and draw comparisons between the two approaches. Although IRT is important in measurement, many psychologists and educators are still familiar with CTT rather than with IRT. Prior to the use of tests based on IRT, it is important for individuals in these disciplines to become familiar with the similarities and differences between the approaches, as IRT promises to be the theory on which test development will be based in the immediate future.

Weiss (1983) mentioned that there are similarities between the two approaches and that the concept of IRT has been implicit in CTT for quite some time. He states that the observed test score is not accepted as an exact measurement on an individual, bur is rather assumed to include error, therefore, it functions to some extent as an estimate of an unobservable true score. The consequence is that CTT generally concerns itself more with reliability and, specifically with standard error of measurement that reflects the extent of the error associated with an observed score as an estimator of the true score. His argument is thus that the true score can be considered to be the same as the trait levels used in IRT, as neither the true score in CTT nor the ability in IRT is observable.

In addition, he mentions that CTT assumes a functional mathematical relationship between that which is observed (i.e., observed score) and that which is unobservable (i.e., true score), and that this is assumed to be linear. His argument proceeds with the claim that CTT incorporates a simple linear mathematical model enabling the estimation of a latent trait (i.e., true score) from an observable variable (i.e., observed score), and is therefore not only similar to IRT, but a very simple latent trait model.

On the other hand, Embretson (1997) states that there are fundamental differences between the two approaches, especially in terms of statistical complexity and qualitative concepts. Although many principles of CTT may be derived from IRT, the opposite is not possible. She advocates that the rules from CTT be "revised, generalized or even abandoned" in the application of IRT (1996, p. 341; 1997, p. 21). See table 2 for the new and old rules of measurement.

Table 2: The New and Old Rules of Measurement

Old Rules	New Rules
The standard error of measurement is applicable across all scores in a specific population	The standard error of measurement differs across scores (or response patterns) but generalizes across populations
Longer tests are more reliable than shorter tests	Shorter tests can be more reliable than longer ones
Comparisons of test scores across multiple forms is dependent upon tests parallelism or sufficient equating	Comparisons of test scores across multiple forms is optimal when difficulty levels of tests vary across individuals
Unbiased assessment of item characteristics is dependent on samples that are representative of the population	Unbiased estimates of item characteristics may be obtained from samples not representative of the population
Meaningful scale scores are obtained by comparing positions in a score distribution	Meaningful scale scores are obtained by comparing distances from various items
Interval scale characteristics are achieved by choosing items yielding normal raw score distributions	Interval scale characteristics are achieved by measurement models that are justifiable rather than score distributions

Note. From "The New Rules of Measurement", by S.E. Embretson, 1996, Psychological Assessment, 8 (4), p. 342. Copyright 1996 by the American Psychological Association. Adapted.

The old rules either follow directly from CTT principles, or are implicit in its application, whereas the new rules reflect IRT principles.

Other Models and Future Directions for IRT Research

Research on IRT is ongoing and entire issues of a number of journals have been devoted to developments in this area (Hambleton, 1995). Especially important has been the development of polytomous unidimensional response models and dichotomous and polytomous multidimensional response models (Hambleton, 1995; Hambleton & Slater, 1997; Van der Linden & Hambleton, 1997).

Samejima pioneered the development of polytomous unidimensional response models in the late 1960's when she introduced the graded response model and in the early 1970's when she introduced a model to handle continuous response data. In addition, her work initiated the extension of unidimensonal models to multidimensonal ones (Hambleton, 1995; Hambleton & Swaminathan, 1985; Van der Linden & Hambleton, 1997; Weiss & Yoes, 1991). Also in the early 1970's, Bock introduced a model to deal with multicategory scoring known as the nominal response model, and in the early 1980's, Master introduced the partial credit model to deal with the same type of scoring system (Hambleton & Swaminathan, 1985; Weiss & Yoes, 1991). These models are those for which items have response formats that are discrete and polytomous, either ranked or unranked (Van der Linden & Hambleton, 1997).

Further research has been conducted, and continues, on models for response time or multiple attempts on items in which tests have a time limit and in which response time is recorded and where tests record the numbers of successes on replicated trials. Also, there are models for multiple abilities or cognitive components (Hambleton, 1995; Van der Linden & Hambleton, 1997). Non-parametric models, which involves the relaxation of stringent parametric assumptions for response functions have led to important discoveries and insights (Sijtsma, 1998; Van der Linden & Hambleton, 1997). In addition, there are models for non-monotone items, which have mainly been utilized in attitude research where response functions do not increase monotonically in the underlying variable (Van der Linden & Hambleton, 1997).

Finally, there are models requiring special assumptions about response processes, such as where there are mixtures of response processes or ability distributions, or conditional dependence exists between responses, or response formats allow for partial knowledge of test items. Other research has considered the extension of IRT to multiple groups (Van der Linden & Hambleton, 1997).

Much research is also being conducted in terms of the wider applications of IRT that are not limited to model development.

Although most of the research is being conducted internationally, South African educators and psychologists are becoming more aware of IRT and what the production, acquisition and application of this knowledge can offer in terms of improving testing procedures, and thus the efficiency and effectiveness of decisions based on tests.

Applications of IRT

Test Development

IRT has made important contributions in test construction and development, and this has occurred in three areas, namely item analysis, item selection and item banking.

Item analysis. IRT is employed to facilitate the determination of sample-invariant item parameters through the use of fairly complex techniques and large sample sizes. Although a representative sample is not needed, it is important that the sample be heterogeneous and sufficiently large to ensure adequate item parameter estimates (Hambleton, 1989, 1996).

Another area within item analysis is the utilization of goodness-of-fit criteria to determine which items do not fit the specified response model. Adequate model-data fit is essential for successful item analysis because items may seem to be poor only because the model-data fit is poor. Identification of poor items is usually achieved by analysis of discrimination of discrimination and difficulty indices. Often, CTT item analysis is conducted as well to supplement information for a more accurate decision (Hambleton, 1989, 1996).

Item selection. The purpose of the test is important in determining item selection. Final selection of items depends on how much the respective items contribute to the overall information supplied by the test. Item information functions are most useful in this instance as they allow one to determine the contribution of each item or task to the test information function, independently of other test items (Hambleton, 1989, 1996).

IRT can thus be used to design a test with particular specifications. One procedure for doing this is to decide on a target test information function (TIF) by describing the shape of the test characteristic curve over the range of abilities desired. Items or tasks with known information functions that will bring the test

information function close to the target. The property of additivity allows one to see the effects of adding or omitting a particular item, and the test information function must be calculated after each selection. Items must be added until the target information function is satisfactorily approximated (Hambleton, 1986, 1989, 1996; Hambleton & Slater, 1997).

In practice, statistical and content specifications are balanced to ensure that the test that results has both content validity and the desired statistical properties. This makes possible the construction of a test that discriminates will at any particular area on the ability scale (Hambleton, 1996). Provided the test developer has an idea of the ability of a group of test-takers, tasks or items can be chosen to maximize the test information at a passing score or yield high information in the area of ability covered by the individuals being tested (Hambleton, 1996; Hambleton & Slater, 1997). This in turn contributes to the precision with which the ability parameters are estimated (Hambleton, 1996).

Performance tests often yield lower levels of performance on pretests than on posttests. Thus, one could select easier items for a pretest and more difficult ones for a posttest. This will increase measurement precision for both administrations, in the region of ability where the test-takers are located. In addition, growth can be measured by subtracting the pretest ability estimate from the posttest ability estimate. This is possible because items on both tests measure the same trait and ability estimates are independent of the specific test items included in a test (Hambleton, 1996).

Computer software is available for the performance of what is alternatively called "optimal test design" or "computerized test assembly" (Hambleton & Slater, 1997).

Item banking. Item banks or item pools facilitate test construction and development. These can be defined as a collection of pre-calibrated assessment material that is assembled and stored in one location. Access of an item bank for item selection allows a test to be born without writing and research on psychometric properties having to be conducted for each item. Item information, based on content, difficulty and discrimination, allows for discernment about item selection, and when there are content-related, technically sound items, assessment quality is maximized (Hambleton, 1986, 1996; Kline, 1993).

IRT facilitated the yielding of maximum benefits of item banks, as item parameter values are independent of the sample from which they were obtained. This invariance of item parameters allows for items that have been pre-tested at various times and with different samples to be included in the same item bank (Hambleton, 1986, 1996; Hambleton & Slater, 1997).

Test Equating

Test equating is a statistical process that is used to adjust scores on separate tests so that the scores can be used interchangeably. These techniques adjust for differences in difficulty among forms that are intended to be similar in either difficulty or content, or both. Other methods that are similar to equating are known as scaling for comparability and linking (Kolen & Brennan, 1995).

The equating process is utilized in situations where alternate forms of a test exist and scores obtained on the different forms must be compared. A good example would be test adaptation, when international comparisons must be made and tests have to be equivalent linguistically and culturally. In terms of IRT, the property of parameter invariance makes it possible to separate the difficulty of a test from the ability of the sample from which the scores were obtained, thus it seems that equating is unnecessary for tests designed with IRT. However, there is always a difference between theory and practice. In practice, different tests will be on different scales, and it is therefore necessary to place them on the same scale in order that they may be compared (Hambleton & Slater, 1997).

There are two types of equating, namely, vertical and horizontal. In the former, tests of varying difficulty must be placed on the same scale, and in the latter, tests that are about the same in terms of difficulty must be placed on a common scale (Hambleton & Slater, 1997; Kolen & Brennan, 1995).

Item difficulty parameters are linearly related in IRT, which makes scaling fairly easy, especially when items have been calibrated using IRT. In terms of horizontal equating, CTT and IRT equating procedures are equivalent when items are not already calibrated with IRT. Vertical scaling, however, reveals IRT equating procedures to be superior to those of CTT methods when items have not been calibrated using IRT (Hambleton & Slater, 1997).

IRT equating has other advantages in that multiple tests can be equated easily. Changes in a test require re-equating, which simply involves removal of the items to be eliminated, and revision of the test characteristic curve (i.e., the sum of ICC's in

the test) without re-calibrating the test. In addition, pre-equating is possible when item parameters are known, which means tests can be equated before they are administered (Hambleton & Slater, 1997).

Score Reporting

Large-scale state, national and international assessments are of interest to the public, and IRT models are being used in score reporting (Hambleton, 1995). Two features that make reporting of scores easier are that (a) a more accurate standard error of measurement is calculated for each ability score, and (b) it is possible to predict individuals' performances on items that have not been administered but calibrated on the same scale as administered items (Hambleton & Slater, 1997).

The ICC plays a role as well because of the assumption of invariance (Hambleton, 1995). In this regard, once ability has been estimated, this estimate can be graphically depicted by means of ICC's of items not administered, allowing for greater description of an individual's ability level. The validity for these inferences depends on model-data fit, however, but the advantages of this aspect are obvious for decision-making (Hambleton & Slater, 1997).

Performance Assessment

There are a number of complications in performance assessments, including how they should be scored, as responses tend not to be dichotomous multiple-choice types of format. Polytomous IRT models that are useful in such circumstances include Bock's nominal-response model, Samejima's graded-response model, and Master's partial-credit model. However, the utility of these models is only evident after raters have done the scoring. IRT does not help in the judgemental process of scoring such assessments (Hambleton & Slater, 1997).

In terms of setting standards for performance assessments, IRT can be used to depict scores or score profiles across assessment exercises, which can facilitate the classification of individuals into groups such as basic, proficient or advanced (Hambleton, 1996).

In addition to assisting with setting performance standards, IRT models can be utilized for the identification of problematic response patterns for individuals and groups of individuals on particular and groups of items. This can facilitate successful diagnosis of problem areas for individuals and groups (Hambleton, 1995). Such diagnoses can prove most useful in the South African context where it is necessary

to know the proficiencies of learners in order to tailor programme delivery to their needs. This is a concept inherent in the NQF (Foxcroft, 1999).

Differential Item Functioning: Test/Item Bias

Differential item functioning (DIF) is said to exist when individuals with equal ability, but from different subgroups, do not have the same probability of responding correctly to an item. Basically, DIF means that an item exhibits bias for different groups (Hambleton, 1989, 1996; Hambleton & Slater, 1997; Hambleton, Swaminathan & Rogers, 1991).

IRT can assist in detecting DIF through the comparison of ICC's for different groups that are of interest, because if there is a discrepancy between the ICC's for two groups, then DIF is said to be present for that item (Hambleton, 1989, 1996; Hambleton & Slater, 1997; Hambleton, Swaminathan & Rogers, 1991). There are a number of IRT methods involving ICC comparisons for detecting DIF, and although they are not proven to be better than CTT methods, they are being used with success (Hambleton & Slater, 1997). DIF is discussed more comprehensively in the next chapter.

Computerized Adaptive Testing

IRT, along with advances in technology, has made computerized adaptive testing (CAT) feasible (Green, 1983; Hambleton & Slater, 1997). Computerized adaptive testing incorporates two aspects, namely, computerized administration and that its difficulty is tailored to the ability level of the examinee on the basis of the individual's responses to items (Hambleton, 1989; Hambleton & Slater, 1997; Weiss, 1995), in effect, creating an individualized test (Hambleton & Slater, 1997).

IRT has played a major role in the implementation of CAT to the extent that most applications of CAT have benefited from and been dependent upon its application (Bunderson, Inouye & Olsen, 1989; Dodd, De Ayala & Koch, 1995; Green, 1983; Hambleton & Slater, 1997; Hambleton, Zaal & Pieters, 1991; Kingsbury & Houser, 1993; McBride, 1997b; Stocking, 1987; Van der Linden, 1995; Wainer, 1990; Weiss, 1983, 1985a, 1995; Weiss & Vale, 1987). IRT has been especially useful in constructing the item pool for CAT, developing strategies for item selection during the administration process, scoring, and providing alternative methods of terminating the test (Kingsbury & Houser, 1993; McBride, 1997b).

Computerized adaptive testing is discussed in more detail in the following chapter.

Practical Considerations

Embretson (1996, 1997) names a few reasons for the limited use of IRT in measurement. First, IRT is considerably more sophisticated statistically than CTT. Second, the researchers whose careers are currently at a peak in measurement did not receive academic training in IRT when they were completing their post-graduate programmes. Finally, IRT is difficult to learn and master outside of a course context and keeping abreast of literature on new developments in the field is unlikely to enable complete practical understanding of the principles.

As an extension of Embretson's (1996, 1997) first point, Hambleton and Swaminathan (1985) mention that IRT requires advanced knowledge or understanding of mathematics. It is a theory based on strong assumptions and it's utility is dependent on the availability of computers. Hambleton (1996) noted that the models are complex and in practice, parameter estimations can arise, especially when researchers work with small sample sizes and short tests. Model-fit is also problematic, especially when the assumption of unidimensionality is violated. More research is thus needed in these areas for model development.

A Brief Critique

Although IRT constitutes a widely accepted improvement over CTT, some researchers have their doubts about the way IRT has been embraced without much consideration of the underlying philosophy. It has been mentioned, for example, that the theory that has been operationalised in IRT models is inadequate. In reality, what has been termed item response theory is actually item response modelling. It has also been pointed out that there remain a number of unsettled controversies in IRT the resolution of which has been the focus of much research (Hutchinson, 1991).

Although there are documented differences between CTT and IRT, it is probably better to view IRT as an extension of CTT (Barnard, n.d.; Weiss, 1983). The ideas of IRT are implicit in CTT. For example, the true score in CTT is analogous to the ability estimate of the trait level in IRT. In CTT there is an assumed (albeit linear) relationship between the visible observed score and the invisible true score, thus enabling the estimation of the true score, analogous to the estimation of a latent trait from that which is observed (Weiss, 1983).

In addition, IRT models are characterized by supposition. Abilities and item traits are inferred from examinee performance on items in that high scorers are presumed to have high ability and low scorers are presumed to have low ability.

Difficulty level is dependent on responses of examinees across the ability spectrum of the sample utilized. This implies that the measurement and perhaps the operational definition of person-item characteristics is actually test dependent (Helms, 1997).

Despite varied application, it is worth noting that even proponents of IRT admit that the assumption of unidimensionality is usually violated. Apart from there being a number of dimensions inherent within one test, which is what multidimensional models of IRT address, performance is probably influenced by aspects such as motivation, anxiety, speed, guessing tendencies when in doubt about answers, and other cognitive skills (Hambleton, Swaminathan & Rogers, 1991). Helms (1997) points out that this means the issue of cultural equivalence and related aspects of bias and fairness remain unresolved, because IRT models do not include traits external to the test, such as group membership (i.e., race, culture and SES).

This is almost addressed, as was mentioned earlier in this chapter, by some more recent research that has considered the extension of IRT to multiple groups (Van der Linden & Hambleton, 1997). However, what is important is that there is awareness that IRT does not solve all problems, but it is progressive and has made enormous contributions to many different areas of assessment.

Advances in theory and technology are dynamic and new discoveries and their applications are inherent in progressive societies, and they are important in transforming societies such as South Africa. The utilization of new knowledge based on sound research can contribute toward the increasing competitiveness of South African society in the international arena even as it aids and facilitates social reconstruction at the level of education.

Although not yet widely used, it seems to be the theory for the immediate future. It is even possible that at some point in the distant future psychometrics and edumetrics will progress beyond item response theory and computerized adaptive testing, which is discussed in the next chapter.

CHAPTER FOUR: ADVANCES IN PSYCHOMETRICS: COMPUTERIZED ADAPTIVE TESTING

IRT, along with advances in technology, has made computerized adaptive testing (CAT) feasible (Green, 1983; Hambleton & Slater, 1997). Generally, computers have greatly facilitated the processes of test construction, administration and scoring (Anastasi & Urbina, 1997; Bunderson, Inouye & Olsen, 1989; Hambleton, Zaal & Pieters, 1991).

Advances in computer technology have improved the efficiency of the testing process, specifically in terms of administering, scoring and even interpreting test information for large numbers of individuals in a shorter period of time than has been possible with the traditional paper-and-pencil methods. This has positive implications for university admissions procedures that move toward including large-scale testing as part of their entrance requirements.

This chapter contains a description and explanation of CAT. It begins with a brief description of the history of adaptive testing, the move to computerized adaptive testing, the issues involved in the change to CAT, the advantages of this testing procedure, and the focus of research developments within CAT. Also covered is how paper-and-pencil testing and CAT are different from each other and certain practical aspects that require consideration prior to implementing computerized assessment measures.

Adaptive Testing: An Historical Overview

Every test is administered according to some or other testing algorithm, which is a set of rules governing the items presented to an individual taking the test, and the order in which these are presented. Testing algorithms are comprised of information concerning how to commence the test, continue it, and end the test. Traditional paper-and-pencil tests in which the score is the number of items correct are typically commenced by starting with the first item, continuing with the next sequentially numbered items, usually in order, and stopping once the last question has been attempted (Thissen & Mislevy, 1990). Such tests are comprised of a set of items that have been pre-selected to constitute a measuring instrument to measure a particular trait. All the questions are administered to each person who takes the test (Weiss, 1995).

This is the conventional approach, with the exception of type being the intelligence test introduced by Alfred Binet in the early 1900's (Green, 1983; Hambleton, Zaal & Pieters, 1991; Thissen & Mislevy, 1990; Schoonman, 1989; Weiss, 1983, 1995; Weiss, & Vale, 1987). Binet's test employed an adaptive testing algorithm, the administration rules of which were more complex (Thissen & Mislevy, 1990).

Adaptive testing, also known as tailored testing, can be defined as a paradigm of testing in which tests are individually constructed during administration through the selection of items that are appropriate in difficulty level for the individual, on the basis of the individual's responses to items, until a satisfactory estimate can be obtained of the individual's ability (Chang & Ying, 1996; Green, 1983; Hambleton, 1989; Hambleton & Slater, 1997; Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; Kline, 1993; Lord, 1980; McBride, Wetzel & Hetter, 1997; Schoonman, 1989; Stocking, 1987; Thissen & Mislevy, 1990; Wainer, 1990; Weiss, 1983, 1985a, 1995; Weiss, & Vale, 1987).

Early Adaptive Tests

Binet's intelligence test contained items classified according to levels of development, known as mental age levels, which corresponded to increasingly difficult items. Testing started with items identified upon consideration of the individual's chronological age, and each was scored as they were administered. Administration continued until two mental age levels were identified, namely, basal and ceiling levels, in which a pre-specified number of items were answered correctly and incorrectly, respectively (Hulin, Drasgow & Parsons, 1983; Thissen & Mislevy, 1990; Weiss, 1985a, 1995; Weiss & Vale, 1987). The effective range of measurement for the individual fell between these two levels (Weiss, 1995; Weiss & Vale, 1987), which is where items were neither too easy nor too difficult (Lord, 1980; Schoonman, 1989).

Two common shortcomings of individually administered clinical instruments include the possibility of bias being introduced by the examiner, where race, gender and ethnicity could impact negatively upon interactions between the examiner and the person being tested. The second problem is the cost involved in individual administration, both in terms of time and finances (Green, 1983; Hulin, Drasgow & Parsons, 1983).

The concept of adaptive testing remained dormant for decades following Binet's introduction of it, and paper-and-pencil adaptive tests were briefly examined again during the 1950's, but abandoned because of the complexity of administration rules (Weiss, 1983, 1995). Simple adaptive testing procedures that were based on classical test theory (Stocking, 1987; Weiss, 1983) and contained mechanical branching rules (McBride, 1997a; Weiss, 1983, 1995), were suggested and examined, and dominated the research in the 1970's and 1980's (Thissen & Mislevy, 1990).

Strategies for adaptive testing can be divided into two broad types, namely two-stage and multistage (Hambleton, 1989; Hambleton, Zaal & Pieters, 1991).

Two-stage Testing

Two-stage testing typically requires an individual to take two tests. Every individual is presented with the first test, known as the routing test, and on the basis of their score on this test, another test that is either easier or more difficult, known as the optimum test, is administered (Anastasi & Urbina, 1997; Hambleton, 1989; Hulin, Drasgow & Parsons, 1983; McBride, 1997a; Schoonman, 1989; Thissen & Mislevy, 1990; Van der Linden, 1995; Weiss, 1985a). Thus, there is only one level of adaptation, and that is between the first and second test (Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983). An ability estimate is then derived from a combination of scores for the two tests (Hambleton, 1989). Issues that emerged in the research conducted around these tests involved (a) test length of the second test, which was often longer, (b) which score should be considered appropriate for deciding whether an easier or more difficult test should be administered in the second session, and (c) interpretation of the aggregated results from different tests administered in the second session (Hulin, Drasgow & Parsons, 1983; Thissen & Mislevy, 1990; Van der Linden, 1995).

Multistage Testing

Multistage strategies involve branching decisions following responses to individual items. These can be either fixed or variable. In the former, the same item structure is utilized for all individuals, but each person can progress through the structure in a unique way. In the latter, there is an item bank or pool from which items are selected that will reduce uncertainty about the ability estimate for the individual taking the test (Hambleton, 1989; Hambleton, Zaal & Pieters, 1991).

One example of a fixed type is the flexi-level test, introduced by Lord in the 1970's. It is conventional in that it is a paper-and-pencil test and it is self-scored. The test typically has an odd number of items that are arranged in order of difficulty, with only one item per difficulty level. Individuals begin by answering an item of moderate difficulty (Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; Weiss, 1985a, 1995). Answers are scratched onto answer sheets. If the answer is incorrect, individuals are directed to an item that is a little easier, whereas if the answer is correct, they are directed to an item that is a little more difficult (Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; Lord, 1980; McBride, 1997a; Thissen & Mislevy, 1990; Van der Linden, 1995; Weiss, 1985a, 1995).

However, the disadvantage of this type of test is that complex instructions are required and it is the responsibility of the examinee to follow them. A major problem arises when instructions are not followed correctly because computing scores for such individuals is then difficult (Hulin, Drasgow & Parsons, 1983; Lord, 1980; Thissen & Mislevy, 1990). Issues for such tests, as reflected in research, concern (a) the impact of complex instructions on test performance (Hulin, Drasgow & Parsons, 1983; Thissen & Mislevy, 1990), (b) the distribution of item difficulties in the test form, (c) the selection of an appropriate discontinuation criterion (Van der Linden, 1995), and (d) only item difficulty tends to be utilized and other item characteristics that influence test performance are ignored (Hambleton, Zaal & Pieters, 1991; Weiss, 1985a).

Another fixed-type is pyramidal testing, also known as the staircase method, which contains items arranged into a lattice-like structure, based on difficulty. There are thus a number of items per difficulty level. Each test begins with the same item, and every individual takes the same number of items. Items are selected by branching through gaps in the lattice that are alongside each other in order to converge on those items that are similar in difficulty to the ability level of the individual taking the test (Anastasi & Urbina, 1997; McBride, 1997a; Weiss, 1985a).

Also fixed-type in nature are stradaptive (stratified adaptive) tests, which contain items that are arranged into mutually exclusive sets or levels of difficulty, within which items are arranged according to discrimination ability. Branching occurs between levels. These tests incorporate variable entry and variable termination. At each stratum, the first unused item is administered, and branching occurs until a ceiling is

reached, which is a level where none of the items is answered correctly (McBride, 1997a; Weiss, 1985a).

Such branching tests constituted a major area of research into adaptive testing. Individuals begin with items of moderate difficulty and proceed to a more difficult item when they respond correctly, and to an easier item when they respond incorrectly. Prior to the testing, all possible branching pathways are already established (Hulin, Drasgow & Parsons, 1983; McBride, 1997a; Weiss, 1985a). Branching variations include (a) increases and decreases in item difficulty by a constant amount, (b) increases and decreases in difficulty that are a function of constants, dependent on correct and incorrect responses, and (c) step size decreases between successive item difficulties where difficulty is a consistent estimate of the ability being tested (Hulin, Drasgow & Parsons, 1983).

Testlets

The concept of testlets was introduced in the 1980's. A testlet is a group of items from one content area that is developed as a whole and inherently holds a set number of predetermined paths that an individual may take. They are therefore small enough to manipulate and large enough to contain their own context. Basically, the procedure for administration of testlets involves first making an estimate of an individual's ability to determine the initial testlet to be administered. Thereafter, ability is estimated after each testlet until the discontinuation criterion has been reached.

The purposes of testlets can be divided into two broad streams. First, was to maintain some level of control over the structure of the completed test. Second, was the issue of fairness because examinees of similar ability levels could be compared on scores derived from tests of similar content (Wainer et al., 1990).

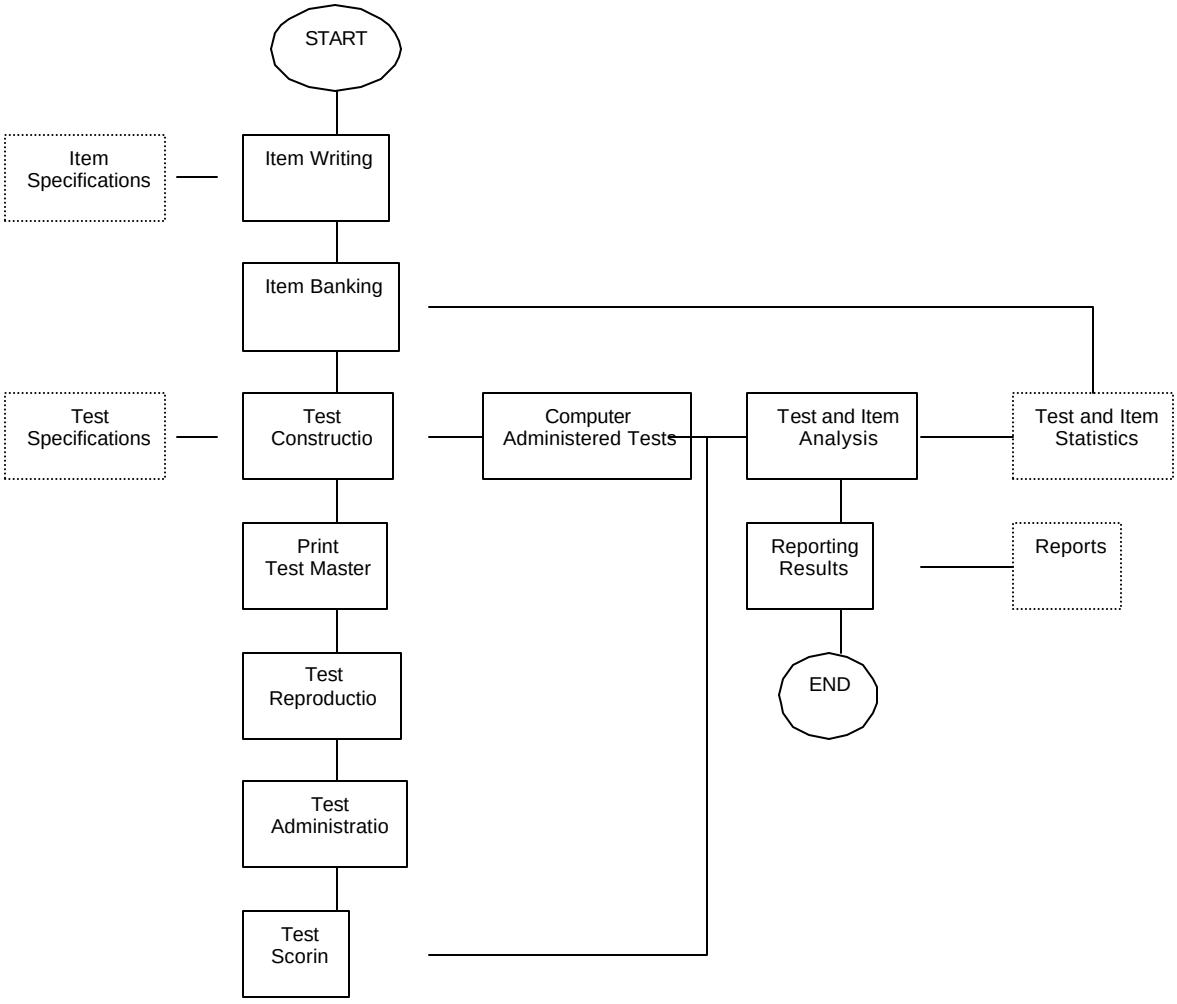
Three uses of testlets represent different kinds of testlet construction. One involves a group of content balanced testlets, equivalent in content and difficulty, which are randomly selected for presentation to examinees. These are useful when it needs to be determined whether or not examinees fall above or below a particular level of ability. In such instances, the distance between their level of ability and the desired level of ability is irrelevant. Another involves linear testlets that are linearly administered because item sequence is based on a single problem, passage or diagram, and all examinees respond to items in all the testlets available so that content is balance between testlets. The third involves hierarchical testlets that are administered linearly, where the individual is routed through items appropriate to their

performance. Correct answers lead to items addressing a more difficult concept and incorrect answers lead to items addressing easier concepts. At the end of the sequence of testlets, individuals are grouped into a number of theoretically ranked levels, based on their patterns of responses to the testlets (Wainer et al., 1990).

The Impact of Computers on Adaptive Testing

Computers have facilitated the testing process in the areas of developing tests, administrating them, scoring them, and even reporting test results (Baker, 1989; Bugbee, 1996; Bunderson, Inouye & Olsen, 1989; Stocking, 1987; Van der Linden, 1995). The functional flow of the testing process is illustrated in figure 4. This process assumes the definition of the construct, and item and test specifications.

Figure 4: Functional flowchart of the testing process



Adapted from "Computer Technology in Test Construction and Processing," by F. B. Baker, 1989, in R.L. Linn (Ed.), Educational Measurement (3rd edition), p. 411. Copyright 1989 by Macmillan.

One specific area in which computers have facilitated progress is the administration is the adaptive test (Green, 1983; Lord, 1980; McBride, 1997a; Sands & Waters, 1997; Thissen & Mislevy, 1990; Van der Linden, 1995). Although original applications of computers involved administration of conventional paper-and pencil test versions whose items had been transferred onto computer for presentation purposes, and were then scored on computer, it was recognized that computers could do more than merely speed up these processes (Bunderson, Inouye & Olsen, 1989; Linn, 1989).

The opposite poles of conventional test construction are referred to as peaked and rectangular. In the former, a set of items is selected for inclusion in a test with difficulties concentrated around one level of difficulty. In the latter, a group of items are selected for inclusion in a test so that there is an equal number of items per the number of difficulty levels spanning a desired and useful range of difficulty. Both types of conventional test, and their variations along the continuum, contain a bandwidth-fidelity dilemma. The peaked test allows a precision of measurement at the point of peak, but has little capacity to differentiate individuals at other levels along the scale. The rectangular test allows differentiation among ability levels along the scale, but it's measurement precision is fairly low (Weiss, 1985a).

It was generally accepted that conventionally administered tests, whether paper-and-pencil or computerized, inherently contained much time-wastage in that correct responses to easy items by high ability individuals and incorrect responses to difficult items by low ability individuals provided relatively little information about their respective ability levels (Sands & Waters, 1997; Wainer, 1990; Weiss, 1985a). In addition, there was the possibility of boredom on the part of high ability individuals and frustration on the part of low ability individuals, which could result in careless and random responses, respectively, introducing greater measurement error into the testing process. It was also recognized that adaptive testing tailors the test to the ability level of the person taking the test, and that computers could collect and evaluate information during the administration of the test (Sands & Waters, 1997; Wainer, 1990).

CAT has become a practical alternative to the traditional paper-and-pencil test.

In many areas where tests are utilized to enhance decision-making, CAT has been implemented either as an adjunct to or a replacement of paper-and-pencil tests (Kingsbury & Houser, 1993; McBride, 1997a; Stocking, 1987). It seems logical that the transformation in higher education should not only acknowledge but also incorporate this medium of assessment in admissions procedures in order to improve decision-making in selection and placement.

It has already been mentioned that IRT played a major role in the implementation of CAT to the extent that most CAT applications have benefited from and been dependent upon the use of IRT (Bunderson, Inouye & Olsen, 1989; Dodd, De Ayala & Koch, 1995; Green, 1983; Hambleton & Slater, 1997; Hambleton, Zaal & Pieters, 1991; Kingsbury & Houser, 1993; McBride, 1997b; Stocking, 1987; Van der Linden, 1995; Wainer, 1990; Weiss, 1983, 1985a, 1995; Weiss & Vale, 1987). IRT has been especially useful in constructing the item pool for CAT, developing strategies for item selection during the administration process, scoring, and providing alternative methods of ending the test (McBride, 1997b).

Strategies for Initiating, Continuing and Terminating CAT

Every adaptive test has certain characteristics (Hulin, Drasgow & Parsons, 1983; Lord, 1980; McBride, 1997b; Thissen & Mislevy, 1990; Van der Linden, 1995; Wainer, 1990; Weiss, 1995; Weiss & Vale, 1987):

1. A pool of items that has been previously calibrated so that such information as difficulty and discrimination is known for each item.
2. A procedure for selecting the first item, which is either determined on the basis of the item parameters of items in the item pool (i.e., one of moderate difficulty) if no prior information is available on the individual's, or on the basis of an ability estimate derived from previous information about the individual taking the test.
3. A method for selecting items, which is usually based on responses to items that have already been administered to the individual.
4. A scoring procedure, which involves scoring items as they are administered or determining scores for groups of items at a number of points during the administration process, and at the end of administration.
5. A method for termination, which may be dependent on the individual's performance, or be fixed to a certain number of items.

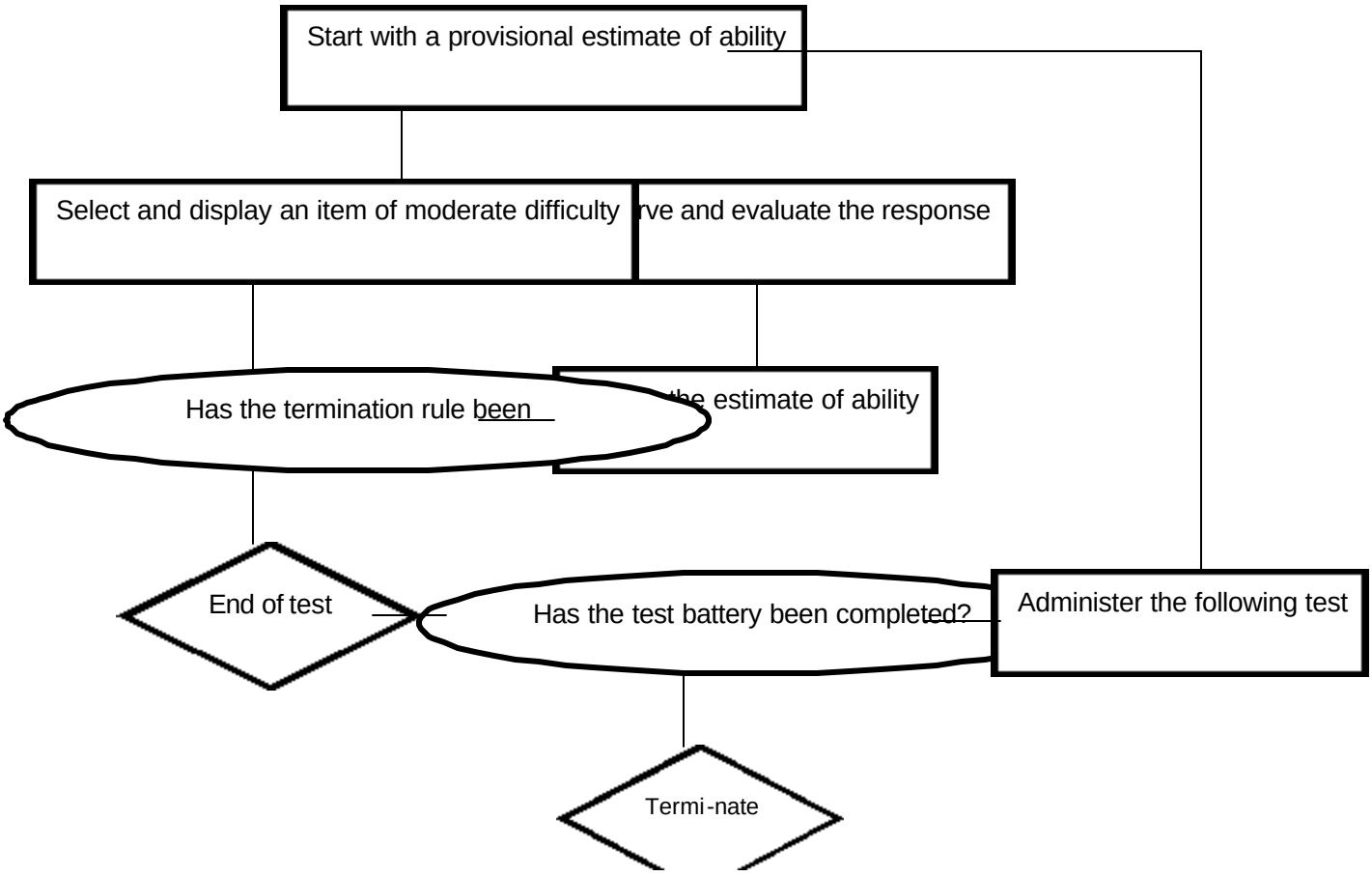
Apart from the first point, these characteristics require the implementation of strategies for the solution of issues involved in CAT applications (Weiss & Vale, 1987). Typically, these are grouped into initial item selection, continued item selection and termination issues (Hulin, Drasgow & Parsons, 1983; McBride, 1997b; Thissen & Mislevy, 1990; Van der Linden, 1995; Wainer, 1990; Weiss & Vale, 1987).

The steps involved in administering adaptive tests are:

1. Make a preliminary specification for an estimate of the individual's ability.
2. Select and administer an item that will yield the most information at that estimated level of ability. Typically, if the item is answered correctly, a more difficult item is presented, and if answered incorrectly, an easier item is presented.
3. Update the ability estimate for the individual after each item administered.
4. Continue administering test items until a designated test termination criterion has been satisfied (Bunderson, Inouye & Olsen, 1989; Lord, 1980; Mills & Stocking, 1995; Stocking, 1987; Stocking & Swanson, 1993; Thissen & Mislevy, 1990).

The adaptive testing process can be depicted in the form of a flowchart, as in figure 5.

Figure 5: Flowchart of the adaptive testing process



Adapted from "Testing Algorithms" by D. Thissen and R.J. Mislevy, 1990, in H. Wainer (Ed.), Computerized Adaptive Testing: A Primer, p. 108. Copyright 1990 by Lawrence Erlbaum Associates.

The design of adaptive tests is such that the test administrator can control the precision of measurement and maximize the efficiency of the testing process because test items are selected for the individual being tested during the administration process (Weiss, 1985a; 1995).

The strategies for initiating, continuing, scoring and terminating CAT are part of the test programme and the decisions pertaining to these strategies are made during the process of test construction.

Item pool calibration

This aspect was covered in the previous chapter on IRT, however, it is important to note that adaptive testing places great demands on test items (McBride, 1997b; Wainer, 1990). The efficiency of any adaptive test is largely dependent on the available number of items (with calibrated, varying difficulties). The larger the number of items and the larger the sample on which the items are calibrated, the better is the performance of the testing system (Schoonman, 1989). The pool available for administration must be much larger than the number of items administered to any particular individual (Lord, 1980). Although item selection is contingent on responses, each individual is presented with a subset of items from a relatively large bank of test items (McBride, 1997b).

As a guideline, the minimum number of items in a pool tends to be 100 (Bunderson, Inouye & Olsen, 1989; Weiss, 1985a). It is, however, preferable to have as many items as possible (for security reasons) across all possible ability levels (for measurement reasons) that are in accordance with all aspects of the purpose of the test. Although theoretically the goal, the number of items written, pretested and accepted into the pool tends to be limited by economics (Mills & Stocking, 1995; Weiss, 1985a). CAT administration allows for the inclusion of new items for pretesting, and requires the removal of certain items, temporarily or permanently, either for security reasons (Mills & Stocking, 1995), or because ongoing research has revealed them to be obsolete (Way, Steffen & Anderson, 1998). Issues surrounding test and item security, and measures for addressing these concerns, are discussed further in this chapter.

Initiating the CAT

There are a few possibilities for beginning a CAT, and these are based on an estimate of the individual's ability level (Hulin, Drasgow & Parsons, 1983; Weiss, 1995). If nothing is known about an individual, the option is to present a first item from the pool that is of moderate difficulty (Folk & Smith, 1998; Green, 1983; Hambleton, Zaal & Pieters, 1991; Lord, 1980; Mills & Stocking, 1995; Weiss & Vale, 1987). This estimate is derived from the mean ability level of the population (Thissen & Mislevy, 1990; Van der Linden, 1995; Wainer, 1990; Weiss & Vale, 1987). This is known as a "fixed entry level" (Weiss, 1985a) or "constant entry level" adaptive test (McBride, 1997a).

If some relevant information is available about the individual, such as educational level or previous test results, an item can be presented that is better matched to their ability level (Lord, 1980; Hambleton, Zaal & Pieters, 1991; McBride, 1997a; Thissen & Mislevy, 1990; Van der Linden, 1995). This is known as a "variable entry level" adaptive test (McBride, 1997a; Weiss, 1983, 1985a). The concept of fairness emerges with the second option in that it is possible that the final test result is biased by prior information, if the information utilized is related to group membership (Thissen & Mislevy, 1990; Van der Linden, 1995). However, it has been demonstrated that a poor choice of the first item will have a minor effect on the final result, unless the test is very short (Hambleton, Zaal & Pieters, 1991; Lord, 1980; Thissen & Mislevy, 1990).

Another consideration that could influence the selection of the first item is that of providing the individual with a so-called "success experience". In such an instance, the first item could be one that is slightly easier than average. However, this has potentially negative implications for the item pool in that the number of easier items available might only be beneficial in terms of the first item administered (Mills & Stocking, 1995).

Sequential item selection

The basic principle in CAT is to administer a slightly more difficult item to the individual who answers correctly and a slightly easier item to the individual who answers incorrectly (Bunderson, Inouye & Olsen, 1989; Hambleton, Zaal & Pieters, 1991; Lord, 1980; Weiss, 1985a). Thus responses are utilized on the previous item(s) to adapt the next item to the current ability estimate of the individual (Thissen & Mislevy, 1990; Van der Linden, 1995; Weiss, 1985a, 1995; Weiss & Vale, 1987).

Usually, item selection is conducted so that there is an expected probability of 50 percent on a correct answer per individual (Schoonman, 1989), which is the most discriminating item for that individual (Hambleton & Slater, 1997) and the item yields maximum information about the individual (Lord, 1980).

Two IRT techniques are mainly used for continued item selection and these are maximum information and Bayesian methods (Chang & Ying, 1996; Folk and Smith, 1998; Hambleton, 1989; Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; Kingsbury & Houser, 1993; Thissen & Mislevy, 1990; Van der Linden, 1995; Weiss, 1983, 1985a, 1985b, 1995; Weiss & Vale, 1987). The concept of item information is important in item selection (Weiss, 1985a, 1995). This was described in the previous section on IRT.

In maximum information item selection, the likelihood of a particular pattern of item responses can be calculated for any point on the ability scale, and the point at which the likelihood is highest is the ability estimate (McBride, Wetzel & Hetter, 1997; Wang & Vispoel, 1998). Thereafter, the most informative item (in terms of difficulty and discrimination) not yet administered, is selected for presentation (Folk & Smith, 1998; Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; Lord, 1980; Thissen & Mislevy, 1990; Van der Linden, 1995; Weiss, 1983, 1985a, 1995; Weiss & Vale, 1987). Thus, the error of measurement is reduced at each step in the process of administration (Van der Linden, 1995; Weiss, 1985a, 1995). The main problem with this method is the tendency for psychometrically desirable items to be overutilised, posing problems for test and item security (Folk & Smith, 1998; Hambleton, Zaal & Pieters, 1991).

There are three kinds of Bayesian item selection, namely, Owen's method, expected a posteriori (EAP) and maximum a posteriori (MAP). Bayesian techniques assume an initial ability distribution, called the prior distribution (Hulin, Drasgow & Parsons, 1983; Thissen & Mislevy, 1990; Wang & Vispoel, 1998). After each response, the likelihood associated with that response is combined with the information about the prior ability distribution to create an adjusted ability distribution, known as the posterior distribution (Wang & Vispoel, 1998). The process is sequential in that each posterior distribution created becomes the prior distribution to be combined with the likelihood of each response to update the ability estimate (McBride, Wetzel & Hetter, 1997; Wang & Vispoel, 1998). Bayesian methods select for administration the item that will maximize posterior precision (Thissen & Mislevy,

1990), or maximally reduce the posterior variance of the ability estimate (Folk and Smith, 1998; Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; McBride, 1997a; Thissen & Mislevy, 1990; Van der Linden, 1995; Weiss, 1983; Weiss & Vale, 1987). The main problem with these methods is the tendency toward bias, especially at ability extremes (Wang & Vispoel, 1998), possibly because of the utilization of inappropriate prior distribution estimates for the ability level of the individual (Hambleton, Zaal & Pieters, 1991).

It has been demonstrated that Bayesian procedures yield less random error, greater administrative efficiency, and approximate actual standard errors than do maximum likelihood procedures (Wang & Vispoel, 1998). Owen's method is the Bayesian procedure that is most commonly utilized (Folk & Smith, 1998; Hambleton, Zaal & Pieters, 1991; McBride, 1997a; Thissen & Mislevy, 1990; Van der Linden, 1995), although it has been demonstrated that EAP provides the best results of the three Bayesian procedures (Wang & Vispoel, 1998).

Scoring

The two IRT methods that are used during test administration to obtain an ability score for the adaptive test are maximum-likelihood and Bayesian estimation (Green, 1983; Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; McBride, Wetzel & Hetter, 1997; Thissen & Mislevy, 1990; Van der Linden, 1995; Weiss, 1995). These are methods for estimating the ability level of the individual during the testing process (Hambleton, 1989; Thissen & Mislevy, 1990; Weiss, 1983, 1995). Generally, a correct answer to an item raises the ability estimate, whereas an incorrect answer lowers the ability estimate (Weiss, 1985a). The ability estimate is consistently altered after every item administered, as this is the basis for item selection (Mills & Stocking, 1995; Weiss, 1995; Weiss & Vale, 1987).

Although maximum likelihood scoring methods are most often used with maximum information item selection, and the Bayesian scoring approach is used with Bayesian item selection, the two scoring methods and item selection strategies can be utilized in a reverse combination (Weiss, 1983).

Terminating the CAT

There are a few typical criteria for terminating a CAT, and these can be applied in purity or in combination (Hambleton, Zaal & Pieters, 1991; Thissen & Mislevy, 1990). There are, however, two types of CAT, namely fixed-length and variable-length tests (McBride, Wetzel & Hetter, 1997).

In fixed-length CATs, the criterion is to stop the test once a fixed number of items have been administered (Bunderson, Inouye & Olsen, 1989; Folk & Smith, 1998; Green, 1983; Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; McBride, 1997a; McBride, Wetzel & Hetter, 1997; Mills & Stocking, 1995; Thissen & Mislevy, 1990; Weiss & Vale, 1987). This option is easy to implement and it is possible to predict item usage rates more precisely (Thissen & Mislevy, 1990). Also, it allows for more direct control of testing time limits and simplifies test schedules (Folk & Smith, 1998). A disadvantage is that individuals are measured with different degrees of precision, and the abilities of those at the extremes will be measured less accurately (Thissen & Mislevy, 1990).

Variable-length CATs utilize other stopping criteria (Folk & Smith, 1998; Hambleton, Zaal & Pieters, 1991; McBride, 1997a; McBride, Wetzel & Hetter, 1997; Thissen & Mislevy, 1990; Weiss, 1995; Weiss & Vale, 1987)). One criterion is to stop the test when the standard error reaches or is less than a specified value (Bunderson, Inouye & Olsen, 1989; Green, 1983; Hambleton, Zaal & Pieters, 1991; Hulin, Drasgow & Parsons, 1983; McBride, 1997a; Van der Linden, 1995; Weiss, 1985a, 1995; Weiss & Vale, 1987). This results in a uniform standard error of measurement across all individuals (Green, 1983; Van der Linden, 1995; Weiss, 1995). It is possible that the use of this termination rule introduces bias for certain ability estimates (Stocking, 1987).

Another possibility is to utilize the individual's response pattern as a basis for stopping the test, which means that the test will be terminated at the point where the most information has been derived (Bunderson, Inouye & Olsen, 1989; Kingsbury & Houser, 1993; Van der Linden, 1995).

It is also possible to terminate solely on the basis of elapsed time, though this is more appropriate for speed than for power tests (Thissen & Mislevy, 1990). Usually, however, a time limit is applied that is considered reasonable (i.e., approximately 95 percent of individuals are able to complete the test) (Folk & Smith, 1998).

Generally, the purpose of the test will determine which termination criterion is most appropriate (Kingsbury & Houser, 1993; Weiss & Vale, 1987). Tests that are used to facilitate diagnostic decisions, such as in clinical evaluations, career guidance and placement situations, require an accurate estimate of the individual's ability. Tests utilized for classification, such as in employment selection or admissions situations, require that individuals be separated into two or more groups on the basis

of test scores. Thus, a precise knowledge of the person's score is unnecessary; it is only necessary to know whether the person's score falls above or below a particular cutpoint (Weiss & Vale, 1987).

One problem with variable length CAT's relates to an issue of fairness. Low ability individuals who have taken shorter tests might argue that their treatment was unfair in comparison with other low ability individuals who had taken longer tests because shorter test lengths are associated with an underestimation of test score. On the other hand, high ability individuals with longer tests might argue that their treatment was unfair in comparison with other high ability individuals who had taken shorter tests because longer test lengths are associated with an underestimation of test score (Stocking, 1997).

Most programmes of adaptive testing have implemented fixed-length CATs and have applied what is considered to be a reasonable time limit (Folk & Smith, 1998). Fixed length CAT's tend to be preferred over variable length CAT's because (a) empirical studies have shown that this kind of termination rule results in acceptable reliabilities and validities, (b) administering the same number of items avoids certain public relations issues associated with variable-length testing, (c) they are easier to administer, and (d) more research is required to assess the relative precision of variable-length testing and it's operational implications (Moreno, Segall & Hetter, 1997).

Advantages of CAT

Adaptive testing involves the adjustment of a set of items in a test in accordance with an individual's ability. This intimates that the goal of this type of test is to determine and present only those items that will yield maximally useful information pertaining to the ability level of the individual being tested (Laatsch & Choca, 1994). The achievement of this goal enhances the efficiency of measurement in all stages of the testing process and also offers a number of additional advantages in other areas relating to testing.

These advantages cannot be emphasized sufficiently for higher education admissions and placement procedures that strive to meet the goals of equality and equity when university entrance decisions are being made.

Administration

Reduction in test length. CATs tend to be considerably shorter than their paper-and-pencil counterparts (Dodd, De Ayala & Koch, 1995; Mills & Stocking, 1995;

Moreno, Segall & Hetter, 1997; Way, Steffen & Anderson, 1998). Research has demonstrated that test length can be reduced by up to 50 percent, without compromising the quality of measurement (Bunderson, Inouye & Olsen, 1989; Maurelli & Weiss, 1981; Vispoel, 1993; Weiss, 1985b, 1995; Weiss & Vale, 1987). This results in savings on testing time as well (Bugbee, 1996; Linn, 1989; Schoonman, 1989; Smittle, 1991). Thus, a CAT allows for shorter tests with greater precision of information relating to the individual's ability (Anastasi & Urbina, 1997; Bugbee, 1996; Chang & Ying, 1996; Legg & Buhr, 1992; Sands & Waters, 1997; Smittle, 1991).

Flexibility within test sessions. CATs are self-paced (Green, 1983; Smittle, 1991), which yields an ideal power test (Green, 1983). Individuals have the freedom to start when they are ready, continue and finish the tests in the battery at a pace with which they are comfortable (within reasonable time constraints) (Sands & Waters, 1997; Wainer, 1990).

Minimization of negative testing experiences. Individuals taking a test that is tailored to their ability level will not become frustrated with questions that are too difficult for them or bored with questions that are too easy for them (Mills & Stocking, 1995; Smittle, 1991). Instead, individuals are challenged without being discouraged (Green, 1983; Smittle, 1991; Wainer, 1990). Research has also indicated that the majority of individuals prefer, and are thus more motivated, to take computerized tests rather than paper-and-pencil versions (Bugbee, 1996; Sands & Waters, 1997; Vicino & Moreno, 1997; Vispoel, 1993).

Standardization. The computer controls the process of administration in terms of instructions that are presented and the mode and medium of presentation (Kline, 1993; Moreno, Segall & Hetter, 1997; Sands & Waters, 1997; Schoonman, 1989).

Simplification of test revision. Pretesting of new items can take place during the testing without influencing the measurement process too dramatically (Green, 1983; Thissen & Mislevy, 1990). Thus, psychometric information can be obtained for experimental items in a way that is more cost-effective than the conventional piloting of test items (Green, 1983; Sands & Waters, 1997; Schoonman, 1989).

Continuous testing. Tests can be scheduled more regularly to approach the convenience of the examinee (Bugbee, 1996; Mills & Stocking, 1995; Way, Steffen & Anderson, 1998).

Scoring

Error reduction. Computers almost eliminate the errors that can occur during conventional scoring procedures, namely clerical errors associated with hand-scoring and reliability problems associated with scanning equipment (Kline, 1993; Moreno, Segall & Hetter, 1997; Sands & Waters, 1997).

Immediate results. Computer scoring is much quicker and results are available almost immediately following the testing session (Bugbee, 1996; Davey & Nering, 1998; Kline, 1993; Sands & Waters, 1997; Smittle, 1991; Wainer, 1990).

Measurement Precision

CATs yield more accurate estimates of ability than their paper-and-pencil counterparts (Dodd, De Ayala & Koch, 1995; Linn, 1989; Mills & Stocking, 1995; Smittle, 1991; Vispoel, 1993; Way, Steffen & Anderson, 1998) and this precision is especially greater for those individuals at the extreme ends of the ability continuum (Anastasi & Urbina, 1997; Bunderson, Inouye & Olsen, 1989; Sands & Waters, 1997).

Test Security

Test material. The stationery associated with paper-and-pencil testing is eliminated in CATs, thus removing the possibility of test material being stolen prior to testing sessions, and answer booklets being marked by examinees (Bunderson, Inouye & Olsen, 1989; Green, 1983; Sands & Waters, 1997; Smittle, 1991; Wainer, 1990). Test information is stored in the computer or system and programmes can be encrypted to enhance test security (Bunderson, Inouye & Olsen; 1989).

Cheating reduction. Each examinee takes a test that is individualized on the basis of their ability, thus the likelihood of cheating is reduced because the items presented to individuals sitting alongside each other will tend to differ, and copying becomes impossible (Anastasi & Urbina, 1997; Bunderson, Inouye & Olsen, 1989; Davey & Nering, 1998).

Research Issues in CAT

CAT has been implemented in a variety of environments for the efficient and effective measurement of a variety of psychological variables (Stocking & Swanson, 1993; Weiss & Vale, 1987). One of the first operational CATs to be implemented was the Graduate Record Examination (Bugbee, 1996; Davey & Nering, 1998; Mills & Stocking, 1995; Schaeffer, Steffen, Golub-Smith, Mills & Durso, 1995). Also during the 1990's, the Defense Department of the USA officially introduced a computerized

adaptive version of the Armed Services Vocational Aptitude Battery (ASVAB) (Sands, Waters & McBride, 1997). The application of CAT for large-scale high-stakes testing has resulted in the emergence of a number of practical issues that have been the focus of much research since their inception during this past decade (Mills & Stocking, 1995). It seems logical that any institution considering the implementation of CAT for this kind of testing, such as universities that need to make admissions and placement decisions, should familiarize themselves with these issues and what research has revealed on them.

Test Security and Item Exposure

Security of tests and items is important both in conventional paper-and-pencil testing and computerized adaptive testing (Davey & Nering, 1998). Tests that are high-stakes are those on which important decisions are based, at least in part, on test results (Stocking & Lewis, 1995b). The main area of concern for test and item security pertains to the exposure of items to potential examinees (Davey & Nering, 1998; Stocking & Lewis, 1995b), thus organizations spend a great deal of time and effort to ensure the security of tests and their items (Stocking & Lewis, 1995b).

There are three goals inherent in large-scale, high-stakes adaptive testing, namely: (a) the maximization of test efficiency by selecting only those items appropriate to the ability level of the examinee; (b) the assurance that tests measure the same composite of multiple traits for each examinee by controlling the nonstatistical nature of items presented; and (c) the protection of item pool by controlling the rates at which items are administered (Davey & Parshall, 1995). There are different approaches to the attainment of these goals, and each yields different adaptive testing algorithms. Any such algorithm contains the ordering of items in the pool according to their desirability for presentation as the next item. Order differences typically reflect specific definitions of item optimality and particular methods of ability estimation. Attempts to control item exposure can thus be seen as modifications imposed on this ordering (Stocking & Lewis, 1995a, 1995b).

A number of methods of controlling item exposure have been applied in the past (and present) that involve increasing conditionality for items. These include:

1. Simple randomization in which a group of items is identified as fairly equal in optimality and the next item is selected randomly from this group (Mills & Stocking, 1995; Stocking & Lewis, 1995b).

2. Randomesque control entailed identifying the best item for administration at a particular point. The second, third and fourth best questions are identified as well, thus the best question is administered only 40 percent of the time, and the second, third and fourth best questions are administered 30 percent, 20 percent and 10 percent of the time respectively (Davey & Nering, 1998; Mills & Stocking, 1995; Thissen & Mislevy, 1990). Although easy to implement and understand, the method does not differentiate between popular questions that are often selected and less popular ones whose numbers are truly random, therefore, it may not prevent overuse of some items (Davey & Nering, 1998; Stocking and Lewis, 1995b)
3. The "INFO4" procedure in which every item in the pool is ranked in descending order according to their information function at the present level of estimated ability. These values are then raised to the fourth power, a maximum is placed on them, and they are then normed to sum to one so that a cumulative function is formed. A random number is generated and the position of the corresponding item is found for the value of the random number, interpreted as a cumulative probability. Thus, the item becomes the next one for administration. Unfortunately, this procedure could be dependent of the particular item pool for which it was developed and not applicable to other pools (Stocking & Lewis, 1995b).
4. The exposure control parameter method considers an examinee as being randomly sampled from a typical group of examinees, and differentiates between the probability that an item is selected by some CAT algorithm and the probability that an item is administered, given it's selection. This method seeks to control the general probability that an item is administered and to ensure that the maximum value over all probabilities of administration is less than a particular value that is the desired maximum rate of item usage. Exposure control parameters are determined for each item through a series of simulations, following each of which the proportion of times each item is selected as the best one and the proportion of times each item is administered, are tallied separately. If the proportion of times each item is selected as the best one is less than or equal to the desired maximum rate of item usage, then the exposure control parameters are set to one and simulations continue. This occurs until they have stabilized and the maximum

observed proportion of administration for all items is approximately equal to the desired value of item usage. Test length has an influence on the number of items in the pool that must have an exposure control parameter of one (Davey & Nering, 1998; Folk & Smith, 1998; Stocking & Lewis, 1995a, 1995b; Thissen & Mislevy, 1990). However, this procedure will administer very easy or very difficult items every time they are selected because these questions are unlikely to exceed the specified maximum exposure rate, even if presented every time they are selected (Davey & Nering, 1998).

5. The multinomial method conditions exposure control parameters on items and the ability levels of the examinee being tested. It contains two phases. The adjustment phase utilizes the exposure control parameter procedure to develop a series of exposure control parameters that differ across ability levels for each item, and the following selection phase utilizes the exposure control parameters to override the optimal selection of the next item. In effect, a list of exposure parameters is assigned to each question, the number of parameters being equivalent to the number of ability levels on which tallies were based. During administration, the individual's present ability estimate is used to determine which of the exposure parameters associated with a selected question is relevant (Davey & Nering, 1998; Folk & Smith, 1998; Stocking & Lewis, 1995a).
6. Conditional exposure control assigns exposure parameters to items in the same way as is done in the exposure control parameter method, but two lists of exposure parameters are assigned to each item. The first list contains values that limit the frequency with which items can be administered to examinees at each ability level. The second list contains values that functions of the number of items in the pool so that every item is assigned a parameter in conjunction with each other item in the pool to limit the frequency with which items are allowed to occur concurrently. This method allows the combating of item clusters (i.e., sets of items that appear together frequently. Simulations allow a tally of frequencies related to these two lists and these are compared with set maximum limits and adjusted if necessary. The process continues until there is stabilization (Davey & Nering, 1998).
7. Weighted deviations model (WDM) ordering takes into account nonstatistical item properties or features in conjunction with statistical properties of items.

Desired balances between psychometric and content concerns are characterized as a set of constraints that are weighted by the test constructor. This method also permits specification of overlapping items that may not be administered in the same individual's test. Also, item selection can be restricted to blocks of items if they have something in common with a feature deemed important by the test constructor. At each item selection, the pool or an appropriate subset of the pool is ordered from most desirable to least desirable. The most suitable item selected for administration is one that minimizes the weighted sum of positive deviations from the specified target constraints (Folk & Smith, 1998; Stocking, 1997; Stocking & Lewis, 1995a, 1995b; Stocking & Swanson, 1996).

8. Optimal constrained adaptive testing (OCAT) seeks to maximize information for the present examinee ability estimate subject to a set of specified constraints. Rather than selecting one item per point in the test, all remaining items are selected for the present ability estimate, and at each subsequent point, remaining items are re-selected, in order to ensure that a complete test can be constructed to meet the set of constraints (Folk & Smith, 1998).

Other methods used to date are more theoretical than empirically based, impact upon the item pool itself rather than formulae for item exposure. These include:

1. A stratified question pool according to item discrimination so that at each point of administration, items are selected from only one of the strata. The stratum utilized changes as the test proceeds, beginning from those strata containing the least discriminating items, and moving to those containing the most discriminating items. Within each stratum, items are selected on the basis of appropriate difficulty levels for the examinee (Davey & Nering, 1998).
2. Thoroughbreds and plugs constitute another approach where the pool is divided into a group of highly discriminating, popular questions (thoroughbreds), and a group of less popular, less frequently administered items (the field). For each item to be selected, it must be decided whether to draw the item from the thoroughbreds or from the field. Thoroughbred items are selected using any of the standard CAT algorithms, thus controlling their use, whereas those items selected from the field, called plugs, are drawn either randomly or by matching difficulty and ability, thus ensuring a more balanced use of these items (Davey & Nering, 1998).

Item pool management, in conjunction with exposure control methods, is an important aspect of test security. There are a number of strategies that have been researched:

1. Very large pools are an obvious choice, but they are costly to create and maintain because some items will become obsolete and new ones will need to be added (Davey & Nering, 1998).
2. Rotation of pools is another possibility. There would have to be a number of small item pools that could rotate into and out of action unpredictably. This rotation could take place over time and geography (Davey & Nering, 1998).
3. The vat approach combines the elements of a large item pool and item pool rotation. The starting point is a large pool (the vat) containing up to 30 conventional forms. Smaller pools are created by drawing items from the vat, either systematically or randomly. Systematic drawing of items produces pools with specific content traits. These smaller pools are rotated in and out of use, with the addition of new pools periodically being included occasionally (Davey & Nering, 1998; Way, Steffen & Anderson, 1998).
4. Another approach involves having one large main pool containing more popular, overutilised items, and unpopular, underutilized items, and dividing them into parallel pools, each capable of supporting adaptive testing on their own. Prior to administration, one of the pools is randomly selected and testing proceeds (Stocking & Swanson, 1996).
5. Finally, there is an extension of the main pool from which independent pools are derived, and that is to create overlapping pools from the independent pools. Such independent pools are created by firstly determining the number of replications per item (i.e., the number of overlapping pools in which each item should appear), and then to assign each replication of each item to an overlapping pool. Thereafter, the independent and overlapping pools can be administered concurrently or not (Stocking & Swanson, 1996).

One last aspect related to the item pool management is that there should be rules in place that govern item use over time. Such rules would pertain not only to item exposure rates for items but also the extent of overlap for smaller pools drawn from a larger one. A system of rules would contain docking rules that essentially result in the removal of items from the vat for a period of time. A closed system would contain a fixed docking rule for item retirement that is dependent on rate of item

exposure, whereas an open system would allow retired items to be reintroduced at a later stage. However, docking rules should in practice be variable in terms of how strictly they are applied, depending on the items available in the vat for particular content areas (Way, Steffen & Anderson, 1998).

Item pools tend to be in a constant state of flux because not only are items retired for different reasons, new items are pretested and added to the pools regularly (Mills & Stocking, 1995). The pretesting and seeding of items should be done carefully (Parshall, 1998; Thissen & Mislevy, 1990). Item development and prestesting holds it's own problems because the demand for good items and pretest data is greater for computerized adaptive testing than for other delivery models (Parshall, 1998).

Item Ordering, Skipping, Omissions and Review

Conventional paper-and-pencil tests allow for examinee control over the order in which items are attempted, although there is a pre-specified sequence according to which items are presented. All items are available for perusal at any time during administration. Also under examinee control is the omission of items and review of items. This is not the case with computerized adaptive tests because each item presented is dependent on responses to previous items, and only one item is presented at a time (Mills & Stocking, 1995; Stocking, 1997).

The standardization of CAT's has been questioned when considering context effects because different individuals receive tests tailored to their ability rather than being exposed to the same items in the same order, thus context effects could be different (Bunderson, Inouye & Olsen, 1989).

Skipping, omission and review of items is currently impossible with a CAT (Folk & Smith, 1998; Green, Bock, Humphreys, Linn & Reckase, 1984; Kingsbury & Houser, 1993; Mills & Stocking, 1995; Stocking, 1997; Wainer et al., 1990). Permitting these options would reduce measurement precision and decrease the efficiency of the adaptive design (Mills & Stocking, 1995; Stocking, 1997; Vispoel, 1993).

A fairness issue emerges because allowing examinees control of the order of item administration can unintentionally result in giving examinees control over the actual items administered. This would yield a worthless measuring instrument if all examinees took advantage of the option, or an unfair instrument if only a few examinees capitalized on the possibility (Stocking, 1997). Allowing skipping, omission

and revision of items could result in the examinee constructing for themselves an easy test on which they would score very well. This would also result in larger errors in ability estimates, a situation from which low to average ability individuals would most likely benefit, the possibility of bias excluded (Mills & Stocking, 1995; Stocking, 1997).

One of the reasons examinees prefer to be able to skip, omit or review items is that items in a test might overlap to some extent, resulting in context effects or cross-information. This involves one item cueing the correct response for another item (Bunderson, Inouye & Olsen, 1989; Mills & Stocking, 1995; Stocking & Swanson, 1993; Thissen & Mislevy, 1990). Item pools need to be carefully scrutinized for such items, and these can be controlled by placing constraints on the exposure of such items in the same test (Bunderson, Inouye & Olsen, 1989; Mills & Stocking, 1995; Stocking & Swanson, 1993).

Although examinees have expressed concern over their being prevented from skipping, omitting and reviewing items, research has revealed that not allowing these options eliminates irrelevant variance (Mills & Stocking, 1995). In addition, prohibiting these options contributes to the protection of the item pool (Mills & Stocking, 1995; Stocking, 1997). However, the possibility of incorporating these options to a limited degree into adaptive testing procedures continues to be investigated (Folk & Smith, 1998; Mills & Stocking, 1995; Stocking, 1997).

Content balancing

Content balancing relates to the administration of test items that require different knowledge or skills related to one trait (Bunderson, Inouye & Olsen, 1989). The issue arises when the item pool is apparently unidimensional but has been designed to include a number of specific goals and sub goals (Kingsbury & Houser, 1993). A combination of these attributes is reflected in the examinee's score (Thissen & Mislevy, 1990). It is necessary to avoid administering too many items associated with one particular subgoal of a test. Content balancing is usually accomplished by incorporating content specifications as constraints in item exposure control methods (Bunderson, Inouye & Olsen, 1989; Kingsbury & Houser, 1993; Thissen & Mislevy, 1990).

Scoring Procedures

In adaptive testing, the conventional method of scoring by number-correct is not appropriate since different examinees take different sets of items (McBride, Wetzel & Hetter, 1997; Stocking, 1987, 1994). Adaptive test scores tend to be expressed in terms of IRT scales (McBride, Wetzel & Hetter, 1997; Mills & Stocking, 1995). Specifically the ability estimate, a transformation of the standard error of measurement, or the test information function based on the pattern of examinee responses can be utilized in deriving a test score (Bunderson, Inouye & Olsen, 1989).

The main problem that arises in large-scale high-stakes testing is that examinees do not understand the foundations of the different scoring approaches available in adaptive testing. Alternative methods of scoring tests based on IRT, whether paper-and-pencil or computerized, have been and continue to be researched, including a few number-correct procedures that seem to hold promise (Dodd & Fitzpatrick, 1998; Stocking, 1994).

Another aspect impacting on scoring is incomplete tests. Such instances are not problematic in conventional paper-and-pencil testing based on classical test theory, but it poses some difficulty for tests scored on the basis of IRT. A decision must then be made as to whether or not incomplete tests will be scored, and thereafter, how much of the test must be completed before a score is generated (Folk & Smith, 1998; Mills & Stocking, 1995; Moreno, Segall & Hetter, 1997).

It is not feasible to base scores on a small number of responses, as the examinee could then manipulate the test by answering correctly as few questions as possible (Folk & Smith, 1998; Mills & Stocking, 1995; Moreno, Segall & Hetter, 1997). This would introduce a certain amount of bias in favour of the examinee, into the final score. To counteract this, a penalty impacting upon the score can be imposed for incomplete tests (Moreno, Segall & Hetter, 1997).

Examinee Attitudes

A number of questions relating to fairness can arise from the implementation of computerized adaptive testing, and one is that different people take different tests. However, when one considers that individualized tests can be regarded as loosely parallel tests, from this perspective, all tests are equivalent (Green, 1983).

Another issue is that different examinees have different levels of exposure to computers prior to being administered a CAT, and as a result it has been asked whether familiarity and greater experience with a computer does not perhaps

advantage an examinee in some way. Green (1983) states that this is a computer generation and experience has revealed that this is not such a problem. Studies have demonstrated that performance is not affected by previous experience with computers (e.g., Legg & Buhr, 1992; Weiss, 1985b).

Although it is reasonable to accept that attitudes range from great enthusiasm to active, long-lasting dislike of computers (Bugbee, 1996), it has been documented that examinees from all groups are in generally favour of computerized adaptive testing, regardless of the extent of their prior exposure to computers (Bugbee, 1996; Legg & Buhr, 1992; Schaeffer, Steffen, Golub-Smith, Mills & Durso, 1995; Schoonman, 1989; Vicino & Moreno, 1997; Vispoel, 1993), and even express preference for computerized administration over paper-and-pencil administration (Bugbee, 1996; Sands & Waters, 1997; Vicino & Moreno, 1997; Vispoel, 1993). The only aspect of CAT that is perceived as frustrating by examinees is the inability to skip, omit or revise items during administration (Legg & Buhr, 1992; Schaeffer, Steffen, Golub-Smith, Mills & Durso, 1995; Vicino & Moreno, 1997; Vispoel, 1993).

Generally, computerized testing has been accepted, indicating that perceived benefits outweigh those of paper-and-pencil tests (Bugbee, 1996). Previous exposure (or lack thereof) to computers can be regarded an important consideration in the South African situation, but there is no reason to suspect that attitudes toward CAT should be any different in our society from what research has revealed exist in other countries.

Equivalence and Differences Between Paper-and-Pencil Tests and CATs

There is no doubt that the use of computers affects testing, and much of their influence has been positive, but general acceptance of computerized assessment does not necessarily validate computer-administered tests (Bugbee, 1996).

Prior to the American Psychological Association introducing guidelines for the use and interpretation of computer-based tests in 1996, there was no precedent (McBride, 1997c). Since their publication, much research has been conducted and studies comparing paper-and-pencil tests with computerized and computer-adaptive counterparts in order to investigate their equivalence have yielded contradictory results (Bugbee, 1996).

Bugbee (1996) mentions certain aspects that relate to the establishment of equivalence:

1. It is the responsibility of the test developer to demonstrate the equivalence of paper-and-pencil tests and computer-based tests.
2. Equivalence is established by meeting either of two criteria as evidenced in comparisons of actual or rescaled scores, namely (a) that alternate test forms have equal means and distributions, and (b) that interchangeable test scores have equal means and distributions, reliabilities and criterion related validity.

Aside from the required demonstration of equivalence, there are a number of differences between conventional paper-and-pencil tests and computerized adaptive tests. These have been mentioned throughout this chapter, and are summarized in Table 3.

Table 3: Paper-and-Pencil Versus Computerized Adaptive Tests

Paper-and-Pencil Testing	Computerized Adaptive Testing
Periodic scheduling	Continuous scheduling
Administration materials include test booklets and answer sheets	Administration material include computer hardware and software
Linear	Adaptive
Test items are pre-specified for examinees and each examinee receives the same set of items	Examinees receive different sets of items (i.e., individualized tests)
Order of items under examinee control	Item ordering is computer generated
Skipping, omission and review of questions possible and allowed	Skipping, omission or review of items not permitted
Cheating is a real possibility	Cheating is considerably reduced
Tests tend to be fairly long	Tests tend to be fairly short
Time rigid within test sessions in that all examinees begin simultaneously, proceed at similar rates and end at the same time	Flexibility within test sessions and self-paced
Inherently greater standard error of measurement	Greater measurement precision
Scoring could require stencils or other materials, thus taking time and possibly including errors	Scores are computer generated, thus available almost immediately and accurate

Scoring is based on the number of items correct derived from CTT	Scoring is based on ability estimates or SEM or information functions derived from IRT

Practical Considerations

There are a number of aspects that need to be considered prior to implementing CAT, and these are pertinent to the South African context, especially as financial implications play a role in the decisions made in this regard.

Systems

Early projects on CAT were abandoned mainly on the basis of cost-benefit analyses, which revealed that computer administered tests were far more expensive than their paper-and-pencil counterparts (Hambleton, Zaal & Pieters, 1991). However, advances in computer technology made computerized testing more feasible as personal computers replaced mainframes and hardware decreased in cost. The main problem now is not so much one of cost but rather one of choice, as there are a great variety of systems and peripherals available for CAT implementation (Hambleton, Zaal & Pieters, 1991; Weiss, 1985a).

The most feasible way of implementing CAT for large-scale testing is to utilize a system with a main processor and a number of independent units. The main processor would be used for development, data communication and storage, monitoring of the progress of individuals through the test(s), and analyses of results (Green, 1990; Hambleton, Zaal & Pieters, 1991).

Certain aspects influence the decision on the nature and capacity of the independent terminals. One relates to the delay between a response entry and the presentation of the next item. This delay should be brief (i.e., one second). Secondly, there must be a decision as to whether response time should be measured and, if so, the accuracy requirement of this measurement (Green, 1990). Thirdly, the need to use graphic displays for certain items impacts the system choice (Green, 1990; Hambleton, Zaal & Pieters, 1991).

Storage capacity for each terminal is another consideration, and this depends on the number and types of items for administration. Also important is portability, which relates to how often, if ever, the testing equipment must be moved, because relatively permanent installations are easier to equip than are ones that must be moved frequently (Bunderson, Inouye & Olsen, 1989; Green, 1990).

The components of an operational CAT system are numerous and involve hardware and software elements that are fairly complex (Bunderson, Inouye & Olsen, 1989).

Hardware

It is conceivable that hardware influences the examinee experience of taking a computerized test (Segall, 1997). Standard hardware components tend to be the better option, the basics of which include the monitor, CPU, keyboard and/or mouse (Bunderson, Inouye & Olsen, 1989).

Monitor screens must be legible and visible, especially if there is a great deal of reading required in terms of comprehension items or instructions. Single-colour displays are adequate, and variations in brightness and contrast are not recommended (Bunderson, Inouye & Olsen, 1989; Green, 1990).

The keyboard must be available for responses to be entered. A simplified keyboard is usually sufficient for the examinee, as it is unnecessary for all the keys to be utilized during testing. One option is to construct a special unit or to provide an overlay for the standard keyboard so that only relevant keys are accessible. The keyboard should also be movable for comfort. Other response devices can be used instead of a keyboard, such as a touch-sensitive screen or a light pointer, but these may need additional supervision (Green, 1990). A mouse can be utilized, but the keyboard has been demonstrated to be better than the mouse when inexperienced computer users take computerized tests (Schoonman, 1989).

Software

Most software in existence is research-oriented or is dependent on the item pool of a larger test developer (Van der Linden, 1995). Software for CATs that allow developers to implement adaptive versions of their own tests is not readily available. The one exception to this is MicroCAT, which incorporates the option of loading an item pool and defining testing procedures (Hambleton, Zaal & Pieters, 1991; Van der Linden, 1995).

The main issue where software requirements are concerned revolves around exchangeability. The system implemented must be efficient for the current CAT application, but also be designed in such a way as to incorporate the possibility of expansion or new developments. All components in the process from planning to the reporting of test results should be included (Hambleton, Zaal & Pieters, 1991).

One important aspect during administration is that the programme must allow for entry and resetting in different ways, which includes the ease with which the test can be restarted at the current item, an alternative current item or the beginning of the test itself or another test. Software errors are inevitable, although attempts are made to avoid them, thus the best way to deal with them is to try to minimize their effects when they do occur (Green, 1990).

Interface Conventions

An important consideration is what can be termed interface conventions. Screen formats should be standard for all item types and a clear set of rules should be applied for paging or scrolling. Such rules should be quick and easy to learn (Bunderson, Inouye & Olsen, 1989), but it is necessary to allow time for individuals to become familiar with these conventions by incorporating sufficient practice examples, and making provision for skipping of these examples, as there will be individuals who feel comfortable with the processes more quickly than others (Bunderson, Inouye & Olsen, 1989; Legg & Buhr, 1992).

Guidelines for evaluating CATs were prepared by Green, Bock, Humphreys, Linn and Reckase (1984) and covered six areas for examination, namely, dimensionality, reliability, validity, item-parameter estimation and item selection procedures, item pool characteristics, and human factors related to system design. These guidelines are summarized in Appendix 1.

Recent Research on Future Possibilities

Computerized testing and adaptive testing by computer has progressed dramatically since the idea took hold in the 1970's. There are many and varied possibilities that have been and continue to be investigated where the application of computers in assessment is concerned, too many to have been covered in this chapter. Additional areas of recent and continued research, perhaps not mentioned directly in this chapter, include the following:

1. Underlying IRT methods as applied to CAT (e.g., Chang & Ying, 1996; Dodd, De Ayala & Koch, 1995; Wang & Vispoel, 1998);
2. Test models, including design of complex computerized tasks (e.g., Luecht & Clauser, 1998; Van der Linden, 1998);
3. Item development and pretesting, and maintenance and protection of item pools (e.g., Davey & Nering, 1998; Parshall, 1998; Way, Steffen & Anderson, 1998).

4. Delivery models such as computerized adaptive testing on the item and testlet level, linear-on-the-fly testing and computerized mastery testing (e.g., Folk & Smith, 1998);
5. Use of multimedia in large-scale computerized testing (e.g., Bennet, Goodman, Hessinger, Ligget, Marshall, Kahn & Jack, 1997); and
6. Alternatives for scoring CAT's (e.g., Dodd & Fitzpatrick, 1998; Plake, 1998).

Development in technology promises to be dynamic, and this will continue to impact upon assessment in psychometrics and edumetrics and other related fields of application. These developments should be incorporated in the institutions of any society that is in transformation, as transformation implies improvement and advancement. Research-based and theoretical innovations in information technology will continue to be geared toward the facilitation and enhancement of accuracy and efficiency in all spheres of society, and it seems logical to make use of this in the higher educational arena since it is available.

Certain people- and group-related issues remain, however, despite advances and improvements, and these pertain to concerns about bias and fairness. These issues are discussed in the next chapter, with special emphasis on their pertinence in admission and placement.

CHAPTER FIVE: ASSESSMENT ISSUES IN ADMISSION AND PLACEMENT: BIAS AND FAIRNESS

This chapter is concerned with the issues surrounding selection and placement, specifically within the field of education. The uses of tests are mentioned, with special emphasis on selection and placement, followed by a differentiation of the concepts of bias and fairness. Thereafter, bias is discussed in more detail and the methods used to detect it and rectify it, following which two notions of fairness are considered and also decision models that have been proposed to enhance fairness for selection decisions. Finally, the causes of bias and fairness issues are examined, with special reference to multicultural settings, and particular mention of the South African situation.

Test Uses: Defining Selection and Placement

Tests have a variety of uses in education, including the following:

1. Selection or admission in which individuals are either accepted or rejected. Sequential selection includes screening, which involves the rapid, rough designation of individuals into possible acceptance and rejection groups (Anastasi, 1988; Brown, 1983; Cronbach, 1990).
2. Placement, which is developmentally focused in that individuals are assigned to appropriate streams for optimal effects of outcomes on the basis of a single criterion (Anastasi, 1988; Brown, 1983; Cronbach, 1990).
3. Classification in which individuals are assigned to appropriate streams for optimal effects of outcomes on the basis of two or more criteria (Anastasi, 1988; Brown, 1983; Cronbach, 1990).
4. Diagnosis which involves the conceptualization of a problem or situation being experienced to explain it in terms of relative strengths and weaknesses to facilitate a decision on the method to remedy it (Brown, 1983; Cronbach, 1990).
5. Licensing, which is a mandatory procedure similar to selection (Cronbach, 1990).
6. Certification, which is a voluntary procedure to specialize in an area (Cronbach, 1990).

Selection or admission and placement reduces wastage in terms of time and money and increases the possibility that people are suited for a programme, which

raises productivity and individual satisfaction. Allowing every individual an opportunity to try every alternative available is unrealistic (Schoonman, 1989). This view is also propagated in the White Paper on Higher Education (1997) and the National Plan for Higher Education (2001).

Differentiating Bias and Fairness

The terms "bias" and "fairness" are often used interchangeably in colloquial language, and where testing is concerned, it seems obvious that a biased test is unfair and that an unbiased test is fair. These two concepts are quite different within the context of measurement.

Test bias exists when the test makes systematic errors in measurement of a specific attribute, or prediction of a criterion or outcome. Test fairness refers to value judgements with regard to decisions and/or actions made on the basis of test scores. Bias relates to the statistical, empirical attributes within the scientific context of test construction, development, administration, scoring and interpretation (Murphy & Davidshofer, 1991; Osterlind, 1983; Owen, 1992; Reynolds & Brown, 1984). Thus, it exists if the testing procedure is unfair pertaining to an identifiable group of individuals (i.e., based on race, culture, language, gender, age, SES and even sexual orientation) (Rust & Golombok, 1989). Fairness is inseparable from the relative psychological, socio-political and philosophical context in which decisions are made (Murphy & Davidshofer, 1991). Table 4 defines the characteristics of bias and fairness.

Table 4: Defining Characteristics of Bias and Fairness

Bias	Fairness
Refers to test scores or to predictions based on test scores	Refers to actions taken or decisions made on the basis of test scores
Is based on statistical characteristics of scores or predictions	Is a value judgement regarding outcomes
Is defined empirically	Is defined in philosophical or political terms
Can be scientifically determined	Cannot be scientifically determined

Note. From Psychological Testing: Principles and Applications (2nd ed.) by K.R. Murphy and C.O. Davidshofer, 1991, Englewood Cliffs, NJ: Prentice Hall. Copyright 1991 by Prentice Hall. Reprinted.

Test Bias: Identification and Correction

Test bias has traditionally been inferred when mean differences are identified between the scores of two groups. However, such mean differences could be, and probably are, real differences that the test has identified (Cole & Moss, 1989; Reynolds & Brown, 1984). Certain fallacies exist concerning the definition of test bias (Jensen, 1984; Kline, 1993; Taylor, 1987):

1. The egalitarian fallacy states that any mean difference between the scores of two groups on a test necessarily indicates bias.
2. The culture bound fallacy states that group differences are a result of the culture bound nature of items (i.e., tests designed for one group are biased against any other group).
3. The standardization fallacy states that a test standardized for one group is inherently biased against another group, if used on another group.

Bias relates to intrinsic aspects of test themselves, and although reliability plays a role (Rust & Golombok, 1989), the main source of bias seems to lie in validity issues (Anastasi, 1988; Cole & Moss, 1989; Fox & Zirkin, 1986; Goldstein, 1996; Murphy & Davidshofer, 1991; Reynolds & Brown, 1984; Rust & Golombok, 1989; Vane & Motta, 1986). Specifically, tests may be biased in content, construct and criterion-related predictive validity.

Reliability

A test may be differentially reliable for different groups. One way of establishing reliability is by making use of the test-retest method preferably using parallel forms (Kline, 1993; Owen, 1992) for the different groups and differences in the reliability coefficients could indicate the presence of bias. Internal consistency reliability should also be the same for the two groups, after allowing for item difficulty (Cole & Moss, 1989; Kline, 1993; Owen, 1992; Reynolds & Brown, 1984). However, internal consistency approaches tend to identify differences in relationships among items or scores across groups, items or scores that are abnormal in one group in relation to other items or scores. These only detect how item and score relationships differ across groups and do not directly imply bias (Cole & Moss, 1989).

Content and Construct Validity

Content validity refers to the adequacy of the test content as a sample of the defined domain from which inferences are made (Cole & Moss, 1989). Bias can occur when items and answers or item format or presentation in a test may be more

familiar for one group than for another (Cole & Moss, 1989; Hambleton, Clauser, Mazor & Jones, 1993; Kok, 1992; Owen, 1992; Reynolds & Brown, 1984; Walsh & Betz, 1985). Item offensiveness is also a consideration in that such items may subtly elicit emotional or attitudinal responses that affect performance (Anastasi, 1988; Cole, 1981; Hambleton, Clauser, Mazor & Jones, 1993). For example, a test may include words implying or pictorially representing social stereotypes relating to roles of inferiority and superiority in society based on race, culture/ethnicity, SES, and/or gender (Anastasi, 1988; Cole & Moss, 1989; Hambleton, Clauser, Mazor & Jones, 1993; Mensh & Mensh, 1991; Rust & Golombok, 1989; Walsh & Betz, 1985).

Items may represent an adequate sample of the particular content domain, but the domain must also be justified on the grounds of relevance (Cole, 1981; Cronbach, 1990; Fox & Zirkin, 1986; Goldstein, 1996; Kok, 1992). Construct validity has an impact upon the interpretation and explanation of performance on a test in that test content is administered in order to obtain a score that is used for a decision with an intended outcome. Thus, it is context-based (Cole & Moss, 1989; Fox & Zirkin, 1986).

Items that measure a construct that is unrelated to what the test is designed to measure are irrelevant and, as such, are biased if they result in mean differences in group scores (Cole, 1981; Kok, 1992). Construct validity concerns the meanings attached to words or concepts that are being measured. It is important that these are consistent for all groups being assessed by a measuring instrument (Van der Vijver & Poortinga, 1997; Walsh & Betz, 1985). Bias exists when a test measures different traits for different groups or measures the same trait but with different degrees of accuracy for different groups (Owen, 1992).

Violation of content relevance in any way is referred to construct irrelevant variance, and indicates the assessment is too broad and contains additional reliable variance. However, there is also the possibility of construct underrepresentation, which means the assessment is too narrow and excludes important facets of the construct (Messick, 1994).

Validity issues have been addressed by various judgemental and empirical methods (Cole & Moss, 1989; Hambleton, Clauser, Mazor & Jones, 1993; Owen, 1989). One popular judgemental method involves employing representatives of various groups to participate in panels as reviewers or act as consultants to contribute to or evaluate the items of tests being developed (Anastasi, 1988; Camilli

& Shepard, 1994; Cole & Moss, 1989; Hambleton, Clauser, Mazor & Jones, 1993; Reynolds & Brown, 1984; Walsh & Betz, 1985). Empirical methods are more statistical in nature and tend to focus on the differences in performance on individual items in a test (Cole & Moss, 1989; Hambleton, Clauser, Mazor & Jones, 1993). Research conducted typically utilizes two groups, either referred to as focal and reference groups (Camilli & Shepard, 1994; Dorans & Potenza, 1994; Holland & Wainer, 1993; Potenza & Dorans, 1995), or majority and minority groups (Hambleton, Clauser, Mazor & Jones, 1993).

Differential Item Functioning (DIF)

This concept was mentioned previously in chapter three. Differential item functioning (DIF) refers to a psychometric difference in the manner that an item in a test functions for two comparable groups. The presence of DIF indicates that members of two comparable groups (i.e., groups that are matched according to the construct being measured by the instrument) perform differently on a particular item. The utilization of equivalent groups is important because it is necessary to differentiate between differences in item functioning, indicating item bias, and real differences in ability (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Two definitions exist for consideration of DIF, namely, an unconditional definition and a conditional definition. The former states that an item is potentially biased if individuals who are equivalent in all respects excepting for group membership, perform differently on an item to a degree that is more or less than what would be anticipated from the groups performance on other test items. The latter states that an item is potentially biased if individuals with the same level of ability, but representing different groups, do not have the same probability of answering the item correctly (Hambleton, Clauser, Mazor & Jones, 1993; Holburn, 1992; Taylor, 1987; Van der Vijver & Poortinga, 1991). An important difference in the utilization of the term DIF as opposed to the term bias is that the focus is on the results of the statistical analyses rather than on inferences of the effect (Hambleton, Clauser, Mazor & Jones, 1993). In addition, DIF is a relative term in that it compares group performances taking into consideration each group's overall performance on a test (Camilli & Shepard, 1994; Holland & Wainer, 1993).

Early methods of identifying DIF concentrated on statistical significance tests to identify items on which groups performed differently (Hambleton, Clauser, Mazor & Jones, 1993), but IRT approaches of comparing ICCs are currently the popular

methods of identifying DIF (Hambleton, Clauser, Mazor & Jones, 1993; Kline, 1993, Osterlind, 1983; Rust & Golombok, 1989). These are estimated separately for each group. These are the same for items that demonstrate no DIF, but when they are different by more than would be expected from an examination of sampling error, DIF is suspected (Hulin, Drasgow & Parsons, 1983; Lautenschlager, Flaherty & Park, 1994). In addition, discrepancies in item performance across groups can be equal across the range of ability, and in this instance, DIF is said to be uniform. When discrepancies in item performance are inconsistent across the range of ability, DIF is said to be non-uniform (Camilli & Shepard, 1994; Hambleton, Clauser, Mazor & Jones, 1993; Van der Vijver & Poortinga, 1991; Van der Vijver & Poortinga, 1997). Figure 6 illustrates the ICCs demonstrating uniform DIF for two groups and Figure 7 illustrates the ICCs demonstrating non-uniform DIF for two groups.

Figure 6: Uniform DIF as identified by a comparison of ICCs

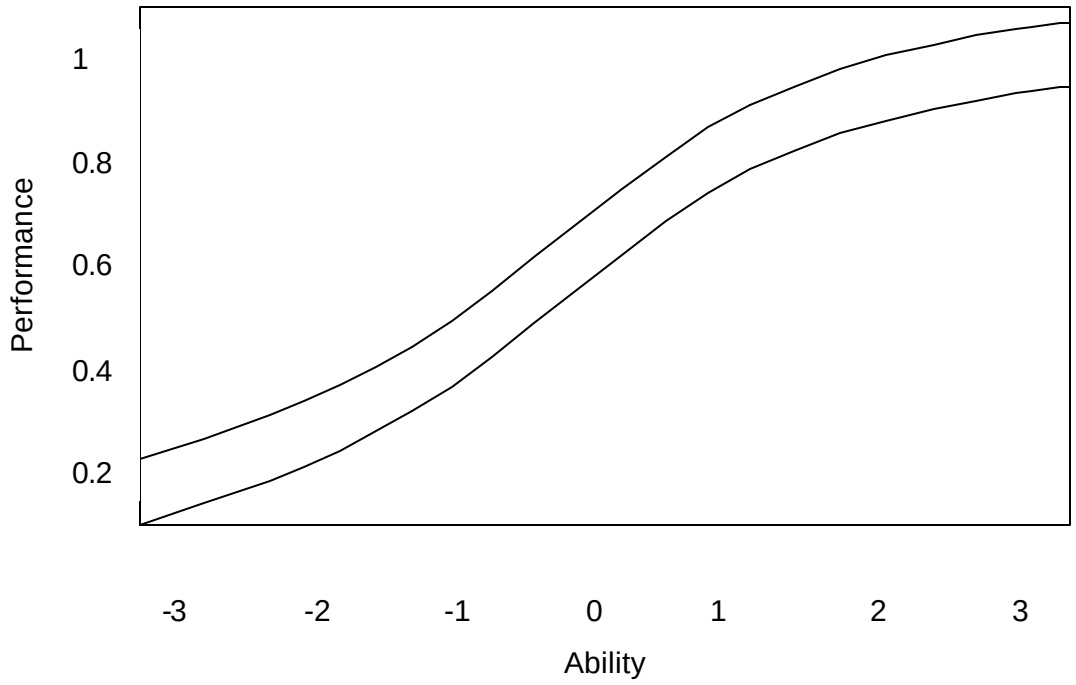
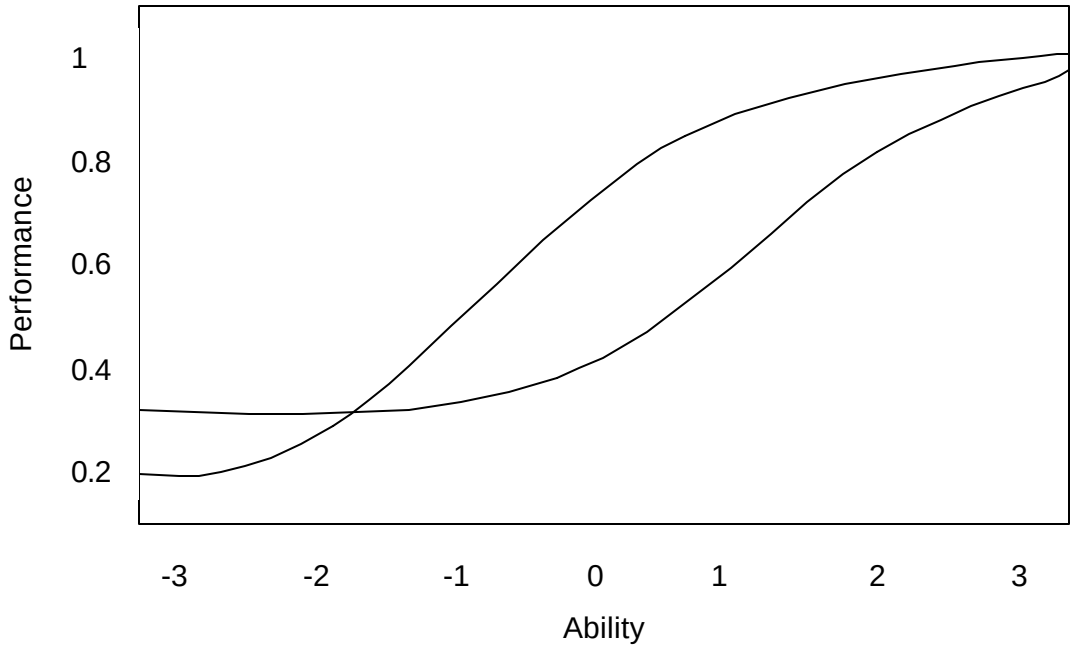


Figure 7: Non-uniform DIF as identified by a comparison of ICCs



There are various methods that have been proposed for examining DIF (Lautenschlager, Flaherty & Park, 1994). Early methods of detecting DIF assumed dichotomous scoring, regardless of the item format, but later procedures developed for assessing DIF addressed polytomously scored items. Advances in theory and technology play an integral part in the improvement of investigating issues pertaining to bias and fairness.

Currently, DIF procedures can be classified into two approaches, namely (a) CTT methods that utilize the observed score as the matching variable, and (b) IRT approaches that utilize an estimate of latent ability, which is a function of the observed data (Dorans & Potenza, 1994; Holland & Wainer, 1993; Potenza & Dorans, 1995). These classifications can be cross-classified with two other procedures, namely (a) parametric procedures that employ a functional form for the relationship between item score and the matching variable, and (b) non-parametric procedures that do not make use of assumptions about the structure of the data (Dorans & Potenza, 1994; Potenza & Dorans, 1995). See Table 5 for a cross-classification of DIF procedures. A brief explanation of each is provided in the numerical order given in table 5.

Table 5: Cross-classification of DIF Procedures

	Parametric	Non-Parametric
Dichotomous DIF		
CTT	Two-way ANOVA (1)	Chi-squares (6)
	Transformed item difficulty (2)	Standardization (7)
	Point-biserial correlations (3)	Mantel-Haenszel (8)
	Factor analysis (4)	
	Logistic regression (5)	
IRT	General IRT-Likelihood ratio (9)	Simultaneous Item Bias (SIBTEST) (14)
	Limited information IRT – Likelihood ratio (10)	
	Log linear IRT – Likelihood ratio (11)	
	IRT-D^2 (12)	
	Lord's Chi-square (13)	
Polytomous DIF		
CTT	Polytomous logistic regression (15)	Mantel (16)
		Polytomous standardization (17)
		H1 and H3 (18)
		Generalized Mantel-Haenszel (19)
IRT	General IRT – Likelihood ratio (20)	Polytomous SIBTEST (23)
	Partial credit (21)	
	Generalized partial credit (22)	

Note. From "DIF Assessment for Polytomously Scored Items: A Framework for Classification and Evaluation' by N.J. Dorans and M.T. Potenza, 1995, Applied Psychological Measurement, 19 (1), p. 25. Copyright 1995 by the American Psychological Association. Adapted.

Two-way ANOVAs. These methods of detecting bias were very common in early investigation of item bias. They typically yield an indication of item difficulty for different groups through the presence of a group by item interaction (Camilli & Shepard, 1994; Cole & Moss, 1989; Jensen, 1984; Kline, 1993; Murphy & Davishofer, 1991; Osterlind, 1983; Owen, 1992; Taylor, 1987; Rust & Golombok,

1989). Significant group by item interactions usually require the application of a secondary procedure to identify particular items that are biased (Camilli & Shepard, 1994; Cole & Moss, 1989; Osterlind, 1983; Taylor, 1987).

Transformed item difficulty. This method is useful as difficulty levels of the items are transformed to standardized scores. Separate transformations are performed for each group and the pattern of item difficulties is then compared for the two groups. Significant differences indicate the presence of bias (Angoff, 1993; Camilli & Shepard, 1994; Cole & Moss, 1989; Hulin, Drasgow & Parsons, 1983; Osterlind, 1983; Owen, 1992; Taylor, 1987). However, certain of these procedures omit consideration of item discrimination indices and can thus give an inaccurate idea of the existence of item bias. In addition, some tend to over identify bias and others tend to miss the presence of bias completely (Camilli & Shepard, 1994; Cole & Moss, 1989).

Point biserial correlations. The coefficients yielded for particular items and total scores for two groups are compared for significant differences. This is similar to the way internal test structure is investigated for consistency (Cole & Moss, 1989; Hulin, Drasgow & Parsons, 1983; Kline, 1993; Owen, 1992; Rust & Golombok, 1989).

Factor analyses. Factor loadings for two groups are examined within the limits of standard errors, and bias is detected through non-equivalence (Cole & Moss, 1989; Hulin, Drasgow & Parsons, 1983; Jensen, 1984; Kline, 1993; Osterlind, 1983; Owen, 1992; Rust & Golombok, 1989). It is uncertain that these procedures are sensitive to the degree of bias that might be expected in reality (Cole & Moss, 1989).

Logistic regression. Item-score regressions with intersection points for two groups are compared for significant differences (Camilli & Shepard, 1994; Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Chi-squares. This approach compares frequencies of correct and incorrect responses to items at different levels of test scores (Angoff, 1993; Cole & Moss, 1989; Hulin, Drasgow & Parsons, 1983; Kline, 1993; Osterlind, 1983; Owen, 1992; Rust & Golombok, 1989).

Standardization. This method yields an average overall index of DIF by averaging differences in expected item scores across total score levels, weighting each difference by focal group relative frequencies (Angoff, 1993; Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Mantel-Haenszel. This method cross-tabulates frequencies of correct and

incorrect responses for the groups concerned and then measures the degree of DIF under the limitation that the odds-ratio is constant across all score levels (Angoff, 1993; Camilli & Shepard, 1994; Dorans & Potenza, 1994; Potenza & Dorans, 1995).

General IRT-Likelihood ratio. This approach uses the Bock-Aitken marginal maximum likelihood estimation algorithm for parameter estimation for a number of models. Comparison is made between a compact model (postulating identical ICCs) and an augmented model (postulating different ICCs). This is most useful as it accommodates a variety of models (Dorans & Potenza, 1994; Potenza & Dorans, 1995; Thissen, Steinberg & Wainer, 1993).

Limited information IRT – Likelihood ratio. This method utilizes the normal cumulative distribution IRT models with generalized least squares parameter estimation techniques and likelihood ratio tests to evaluate significance of observed differences. Comparison is made between a compact model and an augmented model. Typically, large sample sizes are required (Dorans & Potenza, 1994; Potenza & Dorans, 1995; Thissen, Steinberg & Wainer, 1993).

Log linear IRT – Likelihood ratio. This approach utilizes cross-classification of data to assess goodness of fit between a compact and augmented model, the significance of differences being determined through the application of maximum likelihood estimation procedures (Dancer, Anderson & Derlin, 1994; Dorans & Potenza, 1994; Potenza & Dorans, 1995; Thissen, Steinberg & Wainer, 1993). However, it mostly relies on the application of one-parameter logistic models, which do not incorporate discrimination (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

IRT-D^2. This approach analyzes all items at the same time. It applies marginal maximum likelihood estimation, followed by iterations of another algorithm to estimate item parameters for the two groups concerned. The iterations yield standard errors for estimates of item parameters and the ratios of parameter differences to their standard errors are then used to assess the significance of differences noted. However, the three-parameter model is applied, allowing only the difficulty parameter to differ (Dorans & Potenza, 1994; Potenza & Dorans, 1995; Thissen, Steinberg & Wainer, 1993).

Lord's Chi-square. This procedure assumes that the three-parameter model fits the data for the two groups concerned. Discrimination and difficulty are allowed to differ and a chi-square is used to test for significant differences in these parameters

across the groups (Camilli & Shepard, 1994; Dorans & Potenza, 1994; Potenza & Dorans, 1995).

SIBTEST. This procedure postulates a DIF-free multidimensional model as underlying test performance. It actually assesses differential test functioning and incorporates the idea of construct-irrelevant variance. It employs an index that parallels the standardization procedure already mentioned (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Polytomous logistic regression. These approaches involve a different set of pairwise comparisons between score categories or combinations of these. Included here are comparisons of item performance in adjacent categories across groups, continuation-ratio logits, and the proportional odds model. However, they incorporate no descriptive measure of DIF, complicating interpretation, and results may be inconsistent across models because each estimates different sets of odds ratios (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Mantel. This is merely a generalization of the Mantel-Haenszel procedure to the polytomous situation. The comparison is between expected mean item scores for the groups concerned (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Polytomous standardization. This is merely a generalization of the standardization procedure to the polytomous situation concerned (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

H1 and H3. These methods fall within the general standardization framework in which differences in item scores are weighted on the basis of statistical considerations across levels of the observed total score to obtain a summary measure of DIF. Both indexes yield normally distributed statistics with a mean of zero and standard deviation of one. Unfortunately, the test statistics are sample size dependent and do not measure the degree of DIF (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Generalized Mantel-Haenszel. This procedure compares entire item response distributions conditional on the observed scores. The statistic yielded is sensitive to differences in conditional response patterns between groups concerned (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

General IRT – Likelihood ratio. A number of models can be utilized and they can be classified into difference models and divide-by-total models. Difference models speculate the probability of selecting a response category as the difference

between two adjacent cumulative probabilities (e.g., the graded response model). Divide-by-total models express the probability of selecting a response category as an exponential divided by a sum of exponentials (e.g., nominal response model, partial-credit model, and the rating-scale model). The general nominal response model employs a series of likelihood ratio tests to evaluate the significance of DIF by comparing a compact model with a variety of augmented models in which item category response functions differ (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Partial credit. The probability of selecting a particular category is a function of points of intersections for adjacent categorical response curves, or step parameters. All items have the same discrimination parameter (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Generalized partial credit. This model postulates that slope parameters are the same for different groups and tests for differences in step parameters (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

Polytomous SIBTEST. This constitutes an extension of SIBTEST to the polytomous situation. Item performance is regressed onto an estimate of ability. Differences in empirical item/true score regressions are averaged across score levels with a focal group weighting function (Dorans & Potenza, 1994; Potenza & Dorans, 1995).

A multi-method multi-sample approach for detecting bias is preferable as methods based on different theoretical approaches are likely to identify different sets of items as biased, although there should be a great deal of overlap. The multi-method multi-sample approach involves the application of more than one method of detecting item bias, based on each of the different definitions of item bias. These should be conducted using different samples. The items thus identified depend on the frequency with which the methods identify items from the different samples (Taylor, 1987; Holburn, 1992). This inherently entails the assessment of validity through the utilization of convergent and discriminant evidence (Cole & Moss, 1989; Messick, 1994).

There is no item bias detection method in existence that can identify pervasive item bias (i.e., bias that affects all the items equally). If analyses do not reveal any bias, it is not conclusive that none exists. Bias is a matter of degree in that references range from relatively minimal bias to substantial bias. Critical values based on

statistical tests used to determine cut points for demarcating an item as biased or bias-free, and these are subjectively decided (Holburn, 1992).

Criterion Related Predictive Validity

A test provides information about present performance (Anastasi, 1988; Cronbach, 1990; Kline, 1993; Murphy & Davidshofer, 1991). Such information would be irrelevant if there was no way of predicting future performance from this (Cronbach, 1990). Criterion related validity, also known as predictive validity, is determined by the ability of test scores to predict performance in another area external to the test itself (i.e., the criterion) (Anastasi, 1988; Cole & Moss, 1989; Cronbach, 1990; Fox & Zirkin, 1986; Owen, 1992; Vane & Motta, 1986). Bias exists when a test consistently yields differences in the predictor-criterion relationship for different groups (i.e., consistently over- or under-predicts the performance of certain groups on a criterion) (Cole & Moss, 1989; Walsh & Betz, 1985).

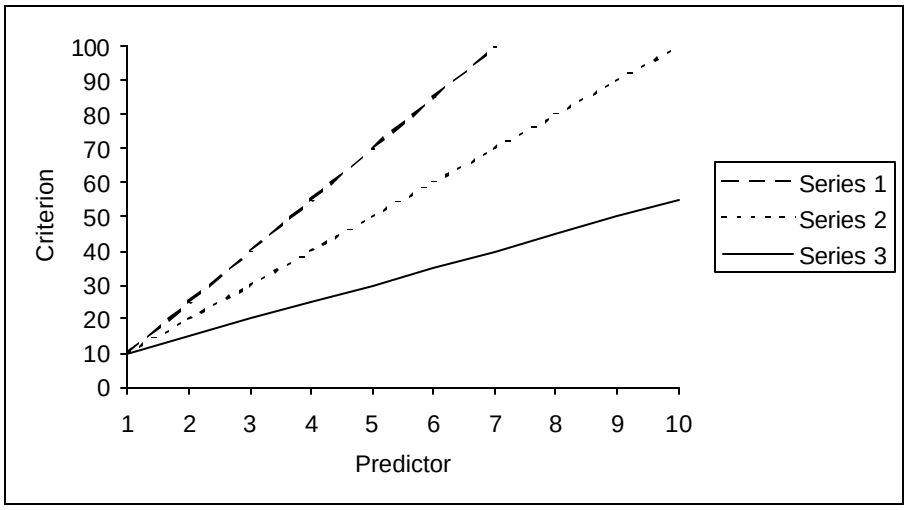
Tests that have established high predictive validity against some particular criterion contain criterion relevant content. Investigation is conducted to determine whether it is effective across groups for its intended purpose. The predictive characteristics of test scores are less likely to vary among groups when a test is intrinsically relevant to criterion performance (Anastasi, 1988). However, prediction systems can differ in standard errors of estimate, regression line slopes and regression line intercepts, despite identical predictor-criterion correlation coefficients (Anastasi, 1988; Cronbach, 1990; Cole & Moss, 1989; Holburn, 1992; Hulin, Drasgow & Parsons, 1983; Murphy & Davidshofer, 1991; Owen, 1992; Rust & Golombok, 1989; Taylor, 1987; Walsh & Betz, 1985).

The main way of examining for bias is to compare regression equations and regression lines for different groups. If significant differences emerge across groups, some form of bias may be established (Anastasi, 1988; Cronbach, 1990; Holburn, 1992; Hulin, Drasgow & Parsons, 1983; Jensen, 1984; Murphy & Davidshofer, 1991; Rust & Golombok, 1989; Schmitt, Hattrup & Landis, 1993; Taylor, 1987; Walsh & Betz, 1985).

The slope of a regression line is equal to the correlation between predictor and criterion when both scores are expressed in standard score units. If the correlation values differ significantly for the groups, resulting in different slopes, then the test exhibits differential predictive validity in terms of the criterion (Brown, 1983; Jensen, 1984; Owen, 1992; Reynolds & Brown, 1984; Walsh & Betz, 1985).

Figure 8 depicts slope bias where several sample regression lines help illustrate how this phenomenon operates. Scores on a hypothetical test are placed on the X-axis and are used to predict performance in some area on the Y-axis. The top line represents Group one (series three), the bottom line represents Group two (series two) and the middle line represents the regression line for both groups together (series one). Assume that selection is dependent on a minimum criterion score of 50; in each case, the same test score would predict a different criterion score. Examination of figure 8 reveals how criterions can be either overpredicted or underpredicted for different groups with different regression lines. Slope bias discriminates against the group with the steeper regression line and in favor of the group with the flatter regression line (Hunter, Schmidt & Rauschenberger, 1984; Hulin, Drasgow & Parsons, 1983; Jensen, 1984; Walsh & Betz, 1985).

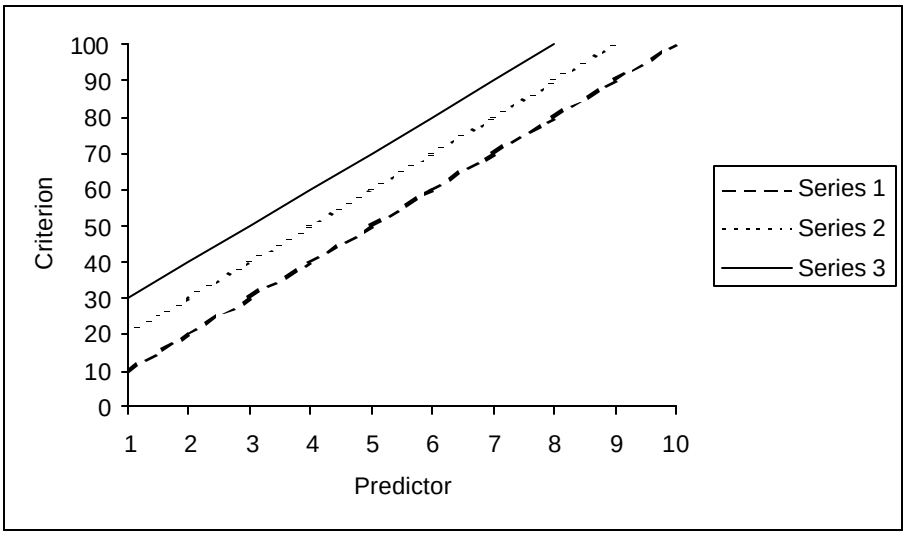
Figure 8: Illustration of slope bias



In intercept bias, regression line slopes are equal but the differences emerge in the points at which the lines intercept the Y-axis or criterion, indicating differences in prediction of the criterion (Brown, 1983; Jensen, 1984; Reynolds & Brown, 1984). Figure 9 depicts intercept bias where several sample regression lines help illustrate the working of this phenomenon. The top line represents Group one (series three), the bottom line represents Group two (series one) and the middle line represents the regression line for both groups together (series two), and 50 is the minimum criterion score for prediction. Overprediction and underprediction can be seen here too for the different groups. Intercept bias discriminates against the group with the higher or

larger intercept, and in favour of the group with the lower or smaller intercept (Camilli & Shepard, 1994; Hunter, Schmidt & Rauschenberger, 1984; Hulin, Drasgow & Parsons, 1983; Jensen, 1984; Walsh & Betz, 1985).

Figure 9: Illustration of intercept bias



Psychometric research on regression lines indicates little, if any, evidence of predictive discrimination where minority groups is concerned. In many cases, use of overall regression lines is in fact likely to benefit these groups, as their scores tend to be overpredicted, and be disadvantageous for majority groups, as their scores tend to be underprediced. Such benefits are generally true for all groups who obtain lower mean test scores (Anastasi, 1988; Camilli & Shepard, 1994; Cole, 1981; Hunter, Schmidt & Rauschenberger, 1984; Walsh & Betz, 1985).

Tests can be valid and relatively bias free, yet still produce scores that result in negative outcomes for certain groups. This relates once again to the distinction between bias and fairness, where the former concerns the quality of the test and the latter concerns uses of test scores. In fact, since conventional tests are usually found to be generally unbiased on technical criteria, the unfair use of tests is implicated as being responsible for causing any identified bias (Cole, 1981; Walsh & Betz, 1985). An understanding of this concept is particularly important when considering the previous situation in South African society.

Test Fairness

Decision Models for Fair Test Use

A shift in focus came about from evaluations of test bias to the design of strategies enhancing fairness in test use where minority groups were concerned. The original strategy of fairness in selection was based solely on predicted criterion performance, but a variety of other models were proposed to increase the selection number of individuals from lower scoring groups (Anastasi, 1988).

One decision model is that proposed by Cleary in the 1960's, namely that selection of individuals is based only on predicted criterion scores (Anastasi, 1988; Camilli & Shepard, 1994; Cole, 1981; Cole & Moss, 1989; Hulin, Drasgow & Parsons, 1983; Hunter, Schmidt & Rauschenberger, 1984; Huysamen, 1995; Lourens, 1984; Reynolds & Brown, 1984; Rust & Golombok, 1989; Willingham & Cole, 1997c). Differences in regression lines could result in the utilization of the two separate regression lines or a multiple regression equation with test score and group membership as predictors (Huysamen, 1995; Lourens, 1984; Schmitt, Hattrup & Landis, 1993). This strategy ignores goals outside of criterion performance that may be part of the selection process (Anastasi, 1988; Cole & Moss, 1989; Rust & Golombok, 1989).

In the early 1970's, Thorndike produced the constant ratio model where selection was determined by a cut point demarcated by the equivalence of proportion selected to proportion successful (Camilli & Shepard, 1994; Cole, 1981; Cole & Moss, 1989; Hunter, Schmidt & Rauschenberger, 1984; Huysamen, 1995; Rust & Golombok, 1989; Schmitt, Hattrup & Landis, 1993).

Darlington simultaneously proposed the modified criterion model, which supposedly maximized validity and did not consider the quota system (Cole, 1981; Cole & Moss, 1989; Hunter, Schmidt & Rauschenberger, 1984; Rust & Golombok, 1989; Schmitt, Hattrup & Landis, 1993).

Cole thereafter provided a conditional probability model where all potentially successful candidates should have an equal chance of being accepted (Cole, 1981; Cole & Moss, 1989; Hunter, Schmidt & Rauschenberger, 1984; Huysamen, 1995; Rust & Golombok, 1989; Schmitt, Hattrup & Landis, 1993; Willingham & Cole, 1997c).

In the mid-1970's, Gross and Su suggested a threshold utility model that employed statistical decision theory to incorporate social influences directly into the

decision process (Rust & Golombok, 1989). Peterson and Novick formalized this approach by explicitly determining utilities associated with the outcomes of success or failure for different groups (Cole, 1981; Cole & Moss, 1989).

Upon their introduction, these models appeared to follow procedures quite distinct from the regression model but it later emerged that they could all be expressed as variants of one comprehensive model. The main differences among the proposed models could be explained in terms of their relative implicit value judgements relating to utility of outcome, being either favourable or unfavourable. These value judgements, combined with the probability of each outcome, are applied in computing the total expected utility of the decision strategy. Such analyses of fair test use reveals that the models contain different definitions of fairness in that they implicitly allocate different values of acceptance and rejection of potential successes and failures within different groups (Anastasi, 1988; Brown, 1983; Cole & Moss, 1989).

Individuals have different values, assumptions and standards of evidence when formulating questions about test score differences (Cole & Moss, 1989; Van der Vijver & Poortinga, 1997). As Goldstein (1996) notes, criteria for judging whether group differences on an instrument constitute bias must ultimately derive from judgements conditioned by prevailing cultural and political perspectives.

Psychometric tradition is one that technical criteria alone should determine test content rather than social or political criteria. The established psychometric method of dealing with equity problems is to seek technical solutions. It is, however, legitimate to seek political or social means of resolving these issues by introducing them earlier in the process of test construction. Traditional notions of bias, fairness and equity in assessment actually derive from an assumption that there exists a method whereby certain groups are not disadvantaged by tests. However, there is no external criterion of fairness and decisions inevitably include socio-cultural and political values. Use of the terms bias and fairness rather than acknowledging and referring to group differences implies a belief in an objective judgement criterion that in reality does not exist (Gipps & Murphy, 1994; Goldstein, 1996).

Notions of Fairness

There are basically two notions of fairness that have emerged, each with different underlying philosophical and ethical premises. The distinctions between

them lead to definite contrasts with regard to test use and decision-making (Reschly, 1986).

Equal treatment. The equality principle lies in treating every person alike with regard to opportunities and selection procedures and criteria, regardless of race, culture/ethnicity, language, SES or gender of individuals. A useful and fair test is one that is characterized by equal validity and prediction, and equivalent test scores relating to the same criterion performance level across groups. Mean differences for groups on the predictor are acceptable if approximately similar criterion group differences exist. The main problem lies in the fact that group differences exist at all stages, leading to unequal outcomes. The effects of this principle contributed to the development of another notion of fairness, namely equal outcomes (Reynolds & Brown, 1984; Reschly, 1986). This notion can also be termed the merit assumption, and incorporates (a) a quality conception that states that selection should optimize the work-related quality of individuals, (b) a group-blind conception that stipulates selection should not consider group membership, and (c) individualism, which proposes that selection should be founded on predictors unrelated to group membership (Lourens, 1984).

Equal outcomes. The equality principle lies in having selection, classification and placement percentages in proportion to the group percentages of the general population (usually by the application of the quota system). Discrimination is identified when there are substantial variations from these percentages. Fairness relates to all groups having an equal, proportional share of available rewards. Standardized tests are regarded as generally useless for resulting in equal outcomes, and this gave impetus for the development of pluralistic norms for conventional tests and differential weighting of test scores of minority groups (Reynolds & Brown, 1984; Reschly, 1986). This notion can also be termed the remedial assumption, which incorporates the quota conception, which states that groups should have opportunities that are proportional to their numerical representation in society, and a compensation conception, which stipulates that selection practices should compensate for past discrimination or adverse impact (Lourens, 1984; Reynolds & Brown, 1984).

There are vast differences in the underlying philosophies and ethics of these two positions because one proposes that discrimination emerges through differential treatment and the other proposes that discrimination emerges through non-

differential treatment, at least to the point of establishing proportional group outcomes (Reschly, 1986).

It has been argued that the consequences of unfairness in the past are so pervasive that a period of reverse discrimination is necessary and justified, but this view has been criticized. Opponents of this perspective argue that to make decisions on anything other than merit represents a compromise of standards in various spheres, the consequences of which could be disastrous (Reschly, 1986).

In addition, the general public tends to be against equal outcomes and reverse discrimination, which might in turn impact negatively on the progress of equal opportunity. The idea of an ethic of hard work and earning one's rewards tends to be scorned, resulting in a loss of efficiency and productivity in society. Use of tests is a critical issue in controversies about the best manner for the achievement of equity and fairness (Reschly, 1986). In South Africa, the White Paper on Higher Education (1997) and the National Plan for Higher Education (2001) stipulates that a balance must be obtained as equity and equality are as important as the maintenance of standards to produce quality graduates in higher education.

It is the responsibility of psychometrists and psychologists to produce tests high in quality and low in bias, and present options and estimates of probable consequences of various policies, but it is the responsibility of society to determine and implement the ethics and policies surrounding test use within the broader socio-political process involving their political representatives (Reschly, 1986). Legislation pertaining to uses of tests must be updated in accordance with scientific knowledge and advancement, and if this is not done, appropriate pressure should be exerted by interested parties that this be done (Kline, 1993).

The Causes of Bias-Fairness Issues

The principal area of concern is the misinterpretation of scores. Tests demonstrate what an individual can do at a particular point in time, but they do not show why performance is at an observed level. External environmental and internal person-related factors need to be investigated to explain this (Anastasi, 1988; Kline, 1993).

There will remain a concern about differences in group scores that are cultural in origin as culture affects the development of psychological traits of certain group members. Differences in experience are manifested in test performance. Tests constitute samples of behaviour and to the extent that culture influences behaviour,

its influence will and should be identified by tests. Ruling out cultural differentials from tests could lower validity as a measure of the behaviour domain they were designed to assess, and it would thus fail to provide the kind of information needed to correct the conditions that impair performance. Tests cannot compensate for disadvantages by eliminating their effects from their scores. In fact, they should reveal such effects so that appropriate remedial steps can be taken. Concealing the effects of deprivation or devising tests insensitive to such effects retards progress toward a true solution of social and educational issues (Anastasi, 1988). This concept is also important to bear in mind when interpreting the test results of South African learners applying for tertiary programmes.

Test-Related Factors

Certain factors reduce validity of tests, including motivation, anxiety, previous exposure to testing situations, rapport with the examiner, and other variables influencing performance but irrelevant to the broad domain being investigated (Anastasi, 1988).

Gender Differences

Documented differences that have emerged through research include that females have greater verbal ability than males, and the reverse is true for visual-spatial ability. As a result, females tend to perform better on language assessments and males perform better on mathematics and science-related assessments (Willingham & Cole, 1997a). Testing and research has also revealed that females have stronger academic records than males at every educational level, their academic work habits tend to be stronger, and they demonstrate more positive attitude and effort academically. In addition, males and females exhibit distinct patterns of values and interests in academics, leisure and occupational preferences. However, there is substantial overlap and wide individual differences within the two groups. There is also indication that culture impacts upon these identified differences (Willingham & Cole, 1997b).

Socio-economic Status

SES can be defined as the amount and quality of economic resources available to an individual. Usually, these are inferred from external indicators, such as educational levels, income, and the like. One consequence of SES is a lack of exposure to material and experiences that stimulate learning and development, merely because of socialization practices associated with social class (Helms, 1997).

Eells and colleagues conducted one of the earliest research studies that investigated test bias in the 1950s. The focus was on test bias in intelligence tests among different socio-economic groups. Although it has been documented that mean test scores differ for the different socio-economic groups, most research has concentrated on mean test score differences found among ethnic and language groups (Camilli, 1993; Holburn, 1992; Reynolds & Brown, 1984). In fact, the ultimate goal of the research by Eells and colleagues was to utilize the knowledge gained to address cultural bias issues (Camilli, 1993; Reynolds & Brown, 1984).

Research conducted on the relationship between SES and test performance has revealed that this variable impacts on experience of item difficulty, typical response patterns, familiarity with the material, and motivational issues. There is apparently an interaction between SES and item format (Owen, 1992).

Culture and Language

Problems emerge with the construction of reliable and valid measuring instruments that can be used across cultures. A test that is valid in one culture is not necessarily valid in another culture (Dana, 1996). The problem of cross-cultural testing was recognized as early as 1910. Some of the earliest of these kinds of tests were constructed in order to assess the large numbers of immigrants to the USA. Others were developed as a result of the need to research relatively isolated cultural groups who had had little or no contact with Western technology which provided the framework for the development of psychological tests (Anastasi, 1988).

Every test tends to favour those from the culture of its origin. Cultural differences are a function of the values and demands placed upon their members. When a person must adjust in order to function, compete and be successful within a culture or subculture other than the one in which they were reared, such differences tend to become disadvantages or handicaps (Anastasi, 1988; Owen, 1992).

Traditionally, tests attempted to eliminate certain parameters characterizing cultural variety. Thus, so-called culture-free tests were developed, based on the idea that an individual's behaviour was covered by a cultural layer that could be penetrated by using such tests. This idea is now regarded as a fallacy, as it is now accepted that all behaviour is in some way a reflection of the individual's cultural environment and tests, being samples of behaviour at a given point in time, should reflect cultural influences. Attempting to devise culture-free test is therefore a futile exercise (Anastasi, 1988; Holburn, 1992).

Thereafter, the trend was to construct tests based on assumptions of common experiences for different cultures, and terms such as culture-common, culture-fair, and cross-cultural were considered appropriate. However, all tests cannot be universally applicable or equally fair to all cultures (Anastasi, 1988), so the term culture-reduced was adopted (Murphy & Davidshofer; 1991) because although cultural differentials can be lowered, they can never be eliminated from test performance.

The practical problems of cross-cultural testing are common in pluralistic societies where there are subcultures or minority cultures within the majority culture. The main concern lies in the applicability of available tests to minority groups (Anastasi, 1988; Fouad, 1993). Assessment within a multicultural context requires cultural competence in order to be acceptable, credible, beneficial and ethical. This entails culture-specific service delivery styles, preferably in the first language of the individual, cultural orientation evaluation, appropriate assessment methodology and measuring instruments, including adequate adaptations of standard tests, and feedback to the person or significant others (Dana, 1996).

Evaluation of cultural orientation involves assessing the level of acculturation regarding the dominant culture for a particular individual. Categories include traditional, marginal, bicultural and assimilated. The assessor should be aware of cultural differences that might influence test performance, but be unrelated to what the test is designed to measure. Assessment should be conducted in the first language of the individual (Dana, 1996).

Language differences, especially proficiency in the language in which the test has been constructed, can influence test performance in ways unrelated to criterion performance (Anastasi, 1988; Kok, 1992; Owen, 1992). The APA (1985) has stated that where testing is conducted in English with individuals for whom English is not a first language, and even with speakers of certain dialects of English, that test taps proficiency or literacy in addition to what it has been designed to measure. This presents certain challenges as the results may not accurately reflect a person's competence if performance depends on familiarity with the language in which testing is conducted. Language affects problem solving on certain tasks. Individuals process information more quickly and accurately in the language most familiar to them. When tested in another language, attention must be divided between decoding the linguistic

statement of a problem and applying strategies and memory to the interpretation and solution of the problem (Duran, 1989).

Language must therefore be taken into account when developing, selecting, administering and interpreting test results, as speakers of two or more languages display different levels of proficiency and efficiency when faced with certain tasks presented in a language that is not their mother tongue (APA, 1985). The term language proficiency refers to the degree of control an individual has obtained in terms of having learned and functional ability to use a language system. It is reflected in four skills, namely, comprehending, speaking, reading and writing (Duran, 1989).

In terms of the tests themselves, there are three main criteria that represent adequate cross-cultural comparisons: (a) the test must be validly translated, (b) items must be equivalent, and (c) both test and individual items must be bias free (Fouad, 1993; Van der Vijver & Poortinga, 1997).

Simple translations are rarely sufficient and usually some adaptation and revision is required (Anastasi, 1988). Translation can be demonstrated through established procedures (Dana, 1996). Valid translation is a process involving three or four steps. Literal translation occurs by translating the test from English into another language, often done independently by more than one person. Back translation is also done independently and involves translating the test back into English and then comparing the two English forms. If too many items are discrepant upon comparison of the original and back translated forms, a committee of consensus might have to reconcile the differences. Finally, a bilingual field test is conducted in which bilingual individuals are administered the two language forms in random order less than a week apart and a correlation is then determined between the two forms (Fouad, 1993; Hulin, Drasgow & Parsons, 1983). Comparability of the two forms can never be assumed. Independent establishment of reliability, validity and norms for the translated version should take place for any group in which the test will be used (Anastasi, 1988; Van der Vijver & Poortinga, 1997), as it cannot be assumed that psychometric properties of a test are comparable across languages or even dialects (APA, 1985).

Equivalence needs to be established on various levels. Functional equivalence refers to the behaviour being measured (Fouad, 1993; Helms, 1992; Owen, 1992; Taylor, 1987). Conceptual or linguistic equivalence refers to similarities of meaning attached to behaviour constructs or concepts (Dana, 1996; Fouad, 1993; Helms,

1992; Hulin, Drasgow & Parsons, 1983; Taylor, 1987) and this is usually established through translation and back translation procedures (Fouad, 1993). Metric equivalence relates to the psychometric scales, which impacts upon conceptual equivalence in terms of measuring the same constructs and concepts within different cultures (Dana, 1996; Fouad, 1993; Helms, 1992; Owen, 1992; Taylor, 1987). Differences in norms, cultural variables, response sets, item statistics, correlations and factor loadings could influence metric equivalence (Dana, 1996).

Helms (1992) cautions that these relate to the tests themselves but suggests that other forms of equivalence must also be included for there to be true cultural equivalence:

1. Equivalence of testing conditions where testing procedures are familiar and acceptable to all groups;
2. Equivalence of context where evaluations must be conducted in standardized settings; and
3. Equivalence of sampling where subject samples representing various groups should be comparable at the test development, validation and interpretation stages.

Both cognitive and personality tests might use verbal or non-verbal stimuli to tap relevant aspects, but where cultural loading in verbal items relates mainly to the meanings of words, non-verbal items can be culturally loaded in terms of using culture symbols. Although tests can be used in other culture than the dominant one, they inevitably are changed in the process of making them applicable for such use and the common result is then that the test becomes valid within rather than between cultures and cross-cultural comparisons then become difficult to interpret.

The South African Context

The main issue in South Africa is culturally relevant test usage in our pluralistic society. There is an attitude of concern over the fact that certain tests designed in the USA or UK primarily for westernized individuals are inappropriate for use in this country because adequate norms are not available and that they were not originally designed for use in other groups (Foxcroft, 1997; Shuttleworth-Jordan, 1996). The use of cognitive tests is particularly controversial as it has been firmly established that socio-cultural factors influence performance. In particular, it is acknowledged that the contents of such tests represent learning and performance thus reflects the testee's learning opportunities and contextual experiences. Also, culture dictates

relevance for its maintenance, and impacts upon the personality, motivation, and cognition of its members, which again is reflected in test performance (Shuttleworth-Jordan, 1996).

Shuttleworth-Jordan (1996) maintains that the following should be borne in mind in the concern about the applicability of (specifically cognitive) tests developed in other cultures to the South African context:

1. The dynamic nature of socio-cultural differences and the changing positions of people along a continuum of lesser to greater levels of urbanization, westernization and literacy;
2. The common brain-behaviour relationships and corresponding cognitive processes in humanity; and
3. Norm-based paradigms categorize people purely on the basis of test scores.

She refers to points that Cronbach (1990) mentioned regarding cross-cultural test usage in the USA:

1. The greater the level of acculturation and urbanization, the less adjustment of scores is required;
2. Mean differences among groups is largely reduced when participants are matched in terms of parent education, occupation and income;
3. Performance differences resulting from cultural influences are not static; and
4. Criticisms of cultural loading have not been supported, as group differences in performance on such items are small, despite being labelled as unfair.

She states that there are signs that the gap between different groups on cognitive tests might be narrowing and could even disappear as socio-cultural differences are minimized and reduced and that certain principles emerge based on the points previously mentioned:

1. In assessment planning, the complex and evolutionary nature of socio-cultural differences must be taken into account;
2. Issues about testing need to be differentiated relating to levels of orientation regarding literacy, urbanization and westernization; and
3. There is emerging evidence of basic commonalities among groups regarding neuro-behavioural function as expressed in performance on cognitive tests.

Language is a variable that plays a major role in influencing test performance (Shuttleworth-Jordan, 1996) and could represent a barrier, especially for Black people. Despite the existence of eleven official languages in South Africa, English is

the first language of less than ten percent of the total population and yet it is in fact the most dominant language in the country. It is the medium of instruction for most tertiary institutions (Peirce & Stein, 1995) and is also often the most frequently chosen medium of instruction for lower levels of education (Foxcroft, 1997). Therefore, it is desirable to determine the testee's level of proficiency in English before testing them, and even then, bilingual assessment is advisable when performance is dependent on previous learning (Foxcroft, 1997).

The assessment procedure and its related context must be considered, as there are instances where the examinee might be unfamiliar with the testing situation. It should be borne in mind that in such situations there is an inequitable relationship between examiner and examinee despite, presumably, a common purpose and set of expectations. The characteristics of the conventional testing situation give rise to meanings within a particular genre that is constituted within and by a particular social occasion, and these may be ritualized or informal. A question of validity arises with regard to the testing environment when the examinee is unfamiliar with such circumstances (Peirce & Stein, 1995).

In South Africa, there are very few culturally relevant tests available and the empirical certainty that those that have been standardized are bias free for pluralistic populations is not established. There is thus a need to develop and norm culturally relevant tests for this country to enhance the fair and ethical use for tests but test development has slowed down rather than increased. One main practical problem is that it is difficult and expensive to begin a new, separate campaign of test construction (Shuttleworth-Jordan, 1996). Also, since South African society is in transition and members of the various cultural groups are undergoing a process of acculturation, it might be better to make use of international tests, as this would maintain a link with international research. The primary requirement would then be to adapt and modify the test for different groups in South Africa and gradually develop applicable norms for local communities (Foxcroft, 1997; Shuttleworth-Jordan, 1996). However, the new measures will be necessary for those who are less literate, urbanized, acculturated and westernized (Foxcroft, 1997).

The development of culturally relevant tests should be performed by a panel of experts, the members of which should include representatives of the various cultural groups in South Africa. They should collaborate on each part of the process of test construction (Foxcroft, 1997). Adaptation of existing internationally recognized tests

includes the selection of an appropriate measuring instrument, adequate translation of the instrument, selection of an experimental design, administration of the adapted version, pilot testing, and the determination of psychometric equivalence (Van Ede, 1996).

The utilization of existing tests in the South African context that entails benefits from the body of international research is relevant and applicable in educational measurement. This is especially the case in higher education, which is in transition, and requires not only new procedures based on advances in theory and technology for the admission of learners, but also a paradigm shift toward a perception of placement, which better meets the needs of learners and South African society.

CHAPTER SIX: PROBLEM FORMULATION

It has already been mentioned that human judgement is fallible when it comes to making important decisions (Dahlstrom, 1993), and following the abolition of Apartheid in South Africa, there are numerous decisions to be made in order to redress previous discriminatory systems (Foxcroft, 1994). Many of these decisions are complicated by the lack of precedent for them, and admissions and placement decisions in education constitute part of this category (Foxcroft, 1994).

Traditionally, matriculation results (alone or converted) have been principally relied upon for making admission decisions in higher education in South Africa (Foxcroft, 1999; Greyling & Calitz, 1997), as matriculation performance was long considered the best, most accurate and fair predictor of academic success at tertiary level (Mitchell & Fridjhon, 1987). However, with the unequal opportunities within the Apartheid era being obvious and well documented, more recent South African research has indicated that matriculation results are unreliable predictors of academic performance, especially for historically disadvantaged students (e.g., Greyling & Calitz, 1997; Skuy, Zolezzi, Mentis, Fridjhon & Cockroft, 1996; Smit, n.d.). The result has been an advocation that higher education systems develop alternative admission criteria and routes (Foxcroft, 1999; Huysamen, 1996; Koch, 1997; Nunns & Ortlepp, 1994; Skuy, Zolezzi, Mentis, Fridjhon & Cockroft, 1996; Smit, n.d.). As was discussed in chapter two, one alternative admission procedure constitutes the establishment of testing programmes to determine the potential linguistic proficiency and numeracy skills of prospective students that underlie academic success (Foxcroft, 1999).

Interest in the development of procedures and measures to broaden access to higher education institutions has grown within the context of the changes that have taken place within education in South Africa. This has provided a further opportunity to introduce testing programmes where the focus is on using the test information to develop the learner. The National Qualifications Framework (NQF) has been instrumental in facilitating the adoption of a learner-centered outcomes based approach in higher education institutions, the effectiveness of which requires that lecturers know the level at which learners are operating upon entrance to a programme in order for them to foster learning experiences that are tailored to the

needs of learners. A logical way of determining learners' strengths and weaknesses would be to assess them prior to their entering a programme. In addition, establishing profiles for each prospective learner would be useful for the guidance of their choices in terms of the degree programme that would be most appropriate for them (Foxcroft, 1999).

The recently initiated transformation in education means that matriculation exemption will no longer be a statutory requirement for entrance to universities. Higher education institutions will be able to determine appropriate entry-level requirements beyond the statutory minimum but admissions criteria will have to be sensitive to the educational backgrounds of prospective students and will have to incorporate recognition of prior learning, which is an essential aspect of the NQF (White Paper on Higher Education, 1997). Entry-level proficiencies of learners must be delineated and operationalised, and this indicates the need for assessment of applicants to higher education institutions (Foxcroft, 1999).

It has been reasoned that the formulation of a multi-stage admissions policy that uses a regression model approach in which admissions test results, school performance and demographic factors are included, would enhance the matching of learners' entry-level knowledge and skills with degree programme requirements (Foxcroft, 1999; Huysamen, 1996). However, "high stakes" admissions testing holds the possibility of discrimination among applicants due to past educational disadvantages, thus the test battery would have to meet standard psychometric criteria and be stringently researched (Foxcroft, 1999; Nunns & Ortlepp, 1994).

Taking this and the heterogeneity of the South African population into consideration, it has been argued that assessing for potential rather than traditional intellectual assessment is a more valid and fair means of determining which applicants to higher educational institutions will be most successful at tertiary studies (Shochet, 1994; Taylor, 1994). However, considering the cost of developing new instruments to measure cognitive and non-cognitive aspects relevant for academic performance, it seems more feasible to adapt and norm existing internationally researched measures to test for entry-level competencies than to develop new ones (Foxcroft, 1997; Shuttleworth-Jordan, 1996).

Taking cognizance of the specifications pertaining to admissions criteria in the White Paper on Higher Education (1997), it is obvious that there needs to be a move away from admissions and selection testing programmes that operate using a gate-

keeping focus, and a move towards placement assessment that is developmentally-focussed.

Prior to the development of an admissions and placement assessment battery, however, the entry-level proficiencies to be measured must be determined. Certain aspects have been identified as being both relevant and important for coping with academic courses, and these include the following (Foxcroft, 1999):

1. Numerical and mathematical proficiency and problem-solving;
2. Academic literacy and English proficiency, especially relating to receptive language ability in terms of understanding of information, being able to identify key ideas, make inferences, read critically, and evaluate how arguments are developed;
3. Expressive language or writing skills, in terms of addressing a given task, developing a logical argument, and using language effectively;
4. Non-verbal reasoning and problem-solving;
5. Non-cognitive aspects, including academic self-efficacy pertaining to personal planning, self management, self-concept and self-appraisal, persistence and motivation, career goals, leadership positions, and community involvement;
6. Work history; and
7. School performance.

An assessment battery used for admissions and placement purposes should thus aim to assess as many of the above-mentioned aspects as possible.

There have been considerable advances on theoretical and technological fronts that have had a great impact on education and psychometrics in recent years. The utilization of internationally researched instruments that have applied these advances in the form of item response theory and computerized adaptive testing could prove to be valuable and efficient in the South African tertiary educational context.

An American based admissions and placement assessment battery that measures generic entry-level proficiencies has been implemented at a tertiary institution in the Eastern Cape and is in the process of being researched. This study forms part of a larger research project that is ongoing. It builds on preliminary findings of cluster analytic studies using the language and mathematical proficiency subtests in the admissions and placement assessment battery (Foxcroft, 1999; Koch, Foxcroft & Watson, 1999). Other studies within the larger project focus on the research

conducted on the non-cognitive aspect of the placement assessment battery (Watson, Foxcroft & Koch, 1999). The various research studies surrounding this assessment battery ultimately contribute toward demonstrating the applicability of using these measures to facilitate admissions procedures at a South African university.

Determining whether or not there is a relationship between scores on proficiency tests that measure generic entry-level proficiencies and the traditionally used matriculation results was deemed a logical route of investigation in the present study. Matriculation results need to be supplemented in order to improve traditional decision-making procedures. The existence of some relationship might be expected, as matriculation results are useful for determining success at university for certain groups, though the relationship might not be very strong, as matriculation results are variable predictors for different groups, as already mentioned.

Also, it was considered logical to investigate whether there is a relationship between the scores on proficiency tests that measure generic entry-level proficiencies, matriculation results, and academic performance as the purpose of the test battery is to facilitate decision-making related to the admission of learners who have the best chance of experiencing success in tertiary education.

Given that the battery is comprised of a number of tests which are not combined into a total score as this would not yield meaningful information, a multivariate analysis that explores underlying patterns among the test scores and matriculation performance in the data set, could facilitate the developmental interpretation of the test information for the relevant degree programme. In addition, such research could be helpful in identifying high and low risk learner profiles and appropriate cutpoints for the tests in the assessment battery in the future (Sireci & Robin, 1999). Consequently, in the present study, cluster analysis was considered an appropriate means of investigating the underlying patterns of performance on the assessment battery. Previous exploratory cluster analytic research with the battery at the university concerned, revealed three profile groupings, namely, a low risk group, a mixed medium risk group with lower numerical and language proficiency, and a high risk group (Foxcroft, 1999). Thus, it was important to explore this further in the present study to see whether similar or different groupings of the patterns of performance would emerge, and which could thus be used to aid admissions, placement, and development decisions.

However, one factor that influenced the design of the present study and the interpretation of the results was the existence of two categories of degree programmes. The principle distinction between them being that mathematics is a prerequisite for enrollment in the one category of degree programmes but not in the other. Based on this difference, it was deemed appropriate to separate the sample into these two groupings of degree programmes and to investigate their respective patterns of performance on the relevant proficiency tests that measure generic entry-level proficiencies and matriculation results.

Research Objectives

The primary and secondary aims of this research were as follows:

Aim 1. To describe the relationship between matriculation results and performance on the Arithmetic and Reading Comprehension tests for first-year learners in the faculties of Arts, Health Sciences, Education and Law.

Aim 2. To describe the relationship between matriculation results and scores on the Arithmetic, Elementary Algebra and Reading Comprehension tests, and academic performance during the first year of tertiary studies Arts, Health Sciences, Education and Law.

Aim 3. To identify and describe underlying patterns of performance (clusters) that emerge based on performance on the Arithmetic and Reading Comprehension tests and matriculation results for learners in the faculties of Arts, Health Sciences, Education and Law.

Aim 4. To describe the relationship between matriculation results and performance on the Arithmetic, Elementary Algebra and Reading Comprehension tests for first-year learners doing programmes in the Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy.

Aim 5: To describe the relationship between matriculation results and scores on the Arithmetic, Elementary Algebra and Reading Comprehension tests, and academic performance for first-year learners doing programmes in the Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy.

Aim 6. To identify and describe underlying patterns of performance (clusters) that emerge based on performance on the Arithmetic, Elementary Algebra, and Reading Comprehension tests and matriculation results for first-year

learners doing programmes in the Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy.

Contingent on aims three and six, the following secondary aims were:

a) To determine whether the clusters can be validated internally;

b) To describe the clusters comprehensively in terms of demographic aspects such as age, gender, culture and home language.

The following chapter documents the methodology utilized in order to achieve these aims.

CHAPTER SEVEN: METHODOLOGY

This chapter incorporates a documentation of the type of study utilized to achieve the aims outlined in Chapter 6, a description of the way the sample was obtained and characteristics of those who participated in the research, the measures used and the general procedure for data collection. Finally, the statistics used to analyze the data are described and explained.

Research Method

This study was primarily exploratory and descriptive. The type of data used was quantitative in nature as it took the form of numerical scores for each of the psychometric measures included in the assessment battery as well as for matriculation and academic performance.

For the purposes of this research study, a correlational method was employed, as the relationships between variables as well as underlying patterns in the data set were explored. Correlational research is concerned with association and is defined by the examination of the nature of relationships between or among variables (Cozby, 1989; Dooley, 1995; Huysamen, 1994; Locke, Silverman & Spirduso, 1998; Somer & Somer, 1991; Wilkinson & McNeil, 1996).

The two disadvantages of this type of method relate to causality and pose basic threats to internal validity. Firstly, it is not possible to state the direction of cause and effect between or among the variables observed (Cozby, 1989; Dooley, 1995; Somer & Somer, 1991; Wilkinson & McNeil, 1996). Secondly, an observed association between or among variables might stem from another extraneous variable (Cozby, 1989; Dooley, 1995; Wilkinson & McNeil, 1996). In an effort to counteract the latter limitation of the correlational method in the present study, it was necessary to take into account the influence of potentially spurious extraneous variables when the results were interpreted (Cozby, 1989; Dooley, 1995; Wilkinson & McNeil, 1996).

In addition to the correlational method, a between-groups design was used in the present study when the cluster groupings that emerged were internally validated. This contributed to the strengthening of the internal validity of this research study and simultaneously provided some explanation for cluster groupings.

Participants

This research employed a non-probability convenience sample. The likelihood of particular individuals being included to participate in the study was unequal and unknown, as participants were selected on the basis of their availability and accessibility.

Convenience sampling is advantageous to the extent that it is cost-effective with respect to time and finances. A principal disadvantage of convenience samples is that participants do not represent the general population, the implication being that the results of the research have limited generaliseability beyond the participants included in the study (Bailey, 1987; Cozby, 1989; Dane, 1990; Dooley, 1995; Wilkinson & McNeil, 1996).

Participants for this particular study were extracted from an existing database, namely, the admissions and placement assessment records of a tertiary institution in the Eastern Cape. The scores on the admissions and placement assessment battery of 193 learners were utilized for this research. These students were registered for degree programmes and had completed at least the first semester of their first year of tertiary studies.

The demographics of the learner group for this study were characterized by heterogeneity in terms of gender, age, culture, home language, socio-economic status (SES), and other background variables such as previous work experience.

The total sample was divided into two groups, depending on the degree programme for which learners registered. This distinction was made on the basis of the different admissions requirements of various degree programmes relating to matriculation mathematics. Certain programmes require mathematics as a matriculation subject and that a certain standard of performance be attained, while others do not have this requirement. Thus, the sample was divided into two groups, namely, non-mathematics based programmes and mathematics-based programmes. The non-mathematics based programmes was comprised of students registered for Health Sciences (excluding Pharmacy), Arts, Education, and Law programmes, which do not require mathematics as a matriculation subject, and the mathematics based programmes was comprised of learners registered for programmes in the Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy, where a certain standard in mathematics as a matriculation subject was an admissions prerequisite.

The Non-Mathematics Based Degrees Group. This group was comprised of 16 male and 52 female students. Ages ranged between 17 and 35 years, with the mean age being 19 (SD = 3.31). Table 6 provides a breakdown of the participants in terms of the biographical variables of culture and home language

Table 6: Breakdown of Participants Following Non-Mathematics Based Degree Programmes According to Culture and Language (N = 68)

Biographical Variable	n	Percentage
Culture Group		
Black	22	32.35
Coloured	14	20.59
Indian	2	2.94
White	29	42.65
Chinese	1	1.47
Home Language		
English	22	32.35
Afrikaans	17	25
English and Afrikaans	7	10.29
Xhosa	16	23.53
Other African Language[a]	6	8.82

Note. [a]This category incorporates the spectrum of Black languages mainly spoken outside of the borders of the Eastern Cape.

The Mathematics Based Degrees Group. This group was comprised of 56 male and 69 female students. Ages ranged between 16 and 39 years, with the mean age being 18.02 (SD = 2.14). Table 7 provides a breakdown of the participants in terms of the biographical variables of culture and home language.

Table 7: Breakdown of Participants Following Mathematics Based Degree Programmes According to Culture and Language (N = 125)

Biographical Variable	n	Percentage
Culture Group		
Black	48	38.4
Coloured	18	14.4
Indian	7	5.6
White	49	39.2

Chinese	3	2.4
Home Language		
English	44	35.2
Afrikaans	26	20.8
English and Afrikaans	6	4.8
Xhosa	37	29.6
Other African Language[a]	9	7.2
Other[b]	3	2.4

Note. [a]This category incorporates the spectrum of Black languages mainly spoken outside of the borders of the Eastern Cape. [b]This category incorporates any language not commonly spoken in South Africa.

SES was not reported for either of the two groups, not because it is not considered important, but rather because data was not gathered on this variable in the database from which the data for the present study was extracted.

Table 8 shows the number of participants per faculty represented in the two groups.

Table 8: Faculty Representation Within the Sample (N = 193)

Faculty	n	Percentage
Non-Mathematics Based Group	68	35.23
Health sciences (excluding Pharmacy)	37	19.17
Arts	8	4.15
Education	3	1.55
Law	20	10.36
Mathematics Based Group	125	64.77
Pharmacy	12	6.22
Building disciplines	12	6.22
Economic sciences	77	39.89
Natural sciences	24	12.44

The external validity of this study, or the extent to which the results are generaliseable beyond the participants of the sample, is questionable because of the sampling strategy utilized. However, the distribution of biographical characteristics in

the groups included in this study were sufficiently reflective of the first-year intake of the tertiary institution where the study was conducted for some general conclusions to be tentatively drawn from the present findings.

Measures

The tests making up the admissions and placement assessment battery are briefly described in this section in terms of their content, administration and scoring, and psychometric properties. Furthermore, the way in which matriculation and academic performance were operationalised will also be described below.

Accuplacer Computerized Placement Tests. The Accuplacer System was developed in the USA and is comprised of four components, including a battery of computerized adaptive tests, to provide information to assist the placement, advising and guidance of students entering a tertiary educational institution. The principal purpose of the test battery is to determine into which degree programmes learners would be appropriately placed and whether developmental modules are necessary, either as prerequisites to, or concurrent with higher educational courses (The College Board and Educational Testing Service, 1997).

The Accuplacer Computerized Placement Tests (Accuplacer CPTs), which were described in chapter four, were developed using Item Response Theory (IRT) and are adaptive in nature which implies that the sequence of questions presented to each student and the questions themselves will differ because they are based on responses to earlier questions. Each test begins with a randomly selected item of medium difficulty, and follows it with a question that is either easier or more difficult, depending on whether the first answer was correct or not. Each question is automatically selected to yield the maximum amount of information about the testee, based on the skill level indicated by their answers to previous questions. This type of testing permits a great deal of precision within a fairly short test (The College Board and Educational Testing Service, 1997; Murphy & Davidshofer, 1991). There are certain constraints to guide the selection of questions administered in order to balance content across ability levels. In addition, all item properties are considered when selecting subsequent items. The algorithm utilized allows support of item sets and controls overlap of items in individual tests. The latest methodology is used for controlling item security and ensuring the most efficient use of items in the pools (The College Board and Educational Testing Service, 1997).

As was discussed in chapter four, the main advantages of adaptive testing are that students are tested more quickly and do not become frustrated or bored by questions which are either too easy or too difficult. The difficulty of questions is quickly and automatically tailored to the ability of the individual. In addition, the tests are self-paced because they are not timed. One person taking the same test twice in succession will almost always receive different questions, thus practice effects are greatly reduced. Finally, there are savings in time because the computer scores and displays results, which means they are available almost immediately.

There are eight tests that comprise the computerized battery, namely, Reading Comprehension, Sentence Skills, Arithmetic, Elementary Algebra, College-Level Mathematics, and the Levels of English Proficiency (LOEP) that incorporates three subtests, which are Reading Skills, Sentence Meaning, and Language Use. Each test and subtest in the battery contains an item pool of 120 items; these were selected using item response theory calibrations. For each of the language related tests and College-Level Mathematics, 20 items are administered to testees, whereas for Arithmetic and Elementary Algebra, 17 and 12 items are administered respectively (The College Board and Educational Testing Service, 1997).

For the purposes of this research, only the Reading Comprehension, Arithmetic and Elementary Algebra tests were used. The reason for this was that the Sentence Skills test was not initially included in the test battery used by the institution where the study was conducted. Furthermore, the College-Level Mathematics test was found to be too difficult and the LOEP tests were found to be too easy to be of value by the institution concerned.

Reading Comprehension measures aspects such as the ability to identify ideas, to make inferences, to read critically, to evaluate strategies used by the writer, and to interpret a graphical illustration. Arithmetic measures number sense, ability to understand and perform operations with whole numbers, fractions, decimals and percentages, to reason out problems and interpret data, all with an emphasis on problems in context. Elementary Algebra measures number sense, operations using real numbers, algebraic concepts, understanding of mathematical relationships, and the ability to make connections among different representations (The College Board and Educational Testing Service, 1997).

Scores for these tests are reported on a 120-point scale. Three scores are obtained: a) the Total Right Score, which indicates performance with respect to all

the questions in the pool from which a test was drawn, thus providing an absolute measure of the testee's skills that is independent of the distribution of skills among all test takers; b) the Range, which indicates the accuracy of the score obtained, being the confidence interval that is equal to the testee's total right score, plus or minus one standard error of measurement (SEM); and c) the Percentile Rank, which indicates student performance in relation to a normative sample of test takers (The College Board and Educational Testing Service, 1997). Only the Total Right scores were used in the present study as no South African norms have yet been established for the Accuplacer CPTs.

As was mentioned in chapter three, reliability of IRT based tests is inherent in statistical estimation procedures and standard errors associated with these. Item and test analyses were conducted to provide an indication of the reliability of the reported scores. Reliability coefficients range from .87 for Reading Comprehension to .92 for Arithmetic and Elementary Algebra. This suggests that the Accuplacer CPTs can be considered reliable.

Furthermore, score comparability studies between the Accuplacer CPTs and the New Jersey Basic Skills Placement Test (NJBSPT) were conducted using item equating techniques, briefly explained in chapter three, and these yielded satisfactory results (The College Board and Educational Testing Service, 1997). Another study conducted at a community college in Florida looked at how the Accuplacer CPTs scores related to scores on the SAT, ACT and E-ACT, and used the equipercentile method, which links comparable scores with the same percentile rank. The conclusion reached was that, despite content differences, the resulting concordance tables were adequate for use to determine comparable scores in order to set standards and provide placement advice for students, but that each institution should create their own database to develop individual tables (Smittle, n.d., in The College Board and Educational Testing Service, 1997).

As was discussed in chapter four, the validity of computerized adaptive tests is established through statistical and content specifications, which need to be balanced in order to ensure content validity. The balancing of these specifications was an important factor in the development of the Accuplacer CPTs. Also, differential item functioning (DIF) statistics were calculated to ensure that items did not function differently for different groups of people (The College Board and Educational Testing Service, 1997).

One study conducted with participants from two community colleges discovered that age had a differential effect on the predictive validity of the Reading Comprehension subtest of the Accuplacer CPTs, and suggested that this be investigated for the other subtests of the Accuplacer CPTs (Cole, Muenz & Bates, 1998).

As was discussed in chapter five, it was considered logical to utilize a test that was developed in the international arena to achieve the purpose of placement of learners into degree programmes. The Accuplacer CPTs were considered to be suitable for use in the South African context as they can be used for the purpose of admissions and placement and because the item wording is linguistically appropriate, especially for second-language English speakers. In addition, a sensitivity review committee evaluated the appropriateness of their use for different cultural groups.

In order to counteract whether familiarity with a computer would be problematic for learners, applicants were prepared for the experience by working through practice examples. As only two keys on the keyboard need to be used, lack of computer familiarity did not pose a serious problem.

Some South African studies have been conducted into the validity of the Accuplacer CPTs. Multiple correlation coefficients between performance on the Accuplacer CPTs and matriculation performance, and academic performance have been found to range from .44 to .58 across different faculties (Foxcroft, 2001). However, when cultural group membership was considered, trends suggested that the Accuplacer CPTs predict academic performance better for white than black learners (Seymour, Foxcroft, Koch, Watson, & Cronje, 2000).

The present study forms part of a larger research project investigating the usefulness and psychometric properties of the Accuplacer CPTs in the South African context, and contributes to the establishment of the validity of this computerized adaptive test battery for a South African population.

Matriculation Results. Matriculation results were also utilized. Every faculty computes the Swedish point differently, their respective formulae weighting the grades and symbols for certain subjects as more important than others. In order to equalize matriculation results, the grades and symbols were assigned standard values across subjects (see table 9), including languages. The weighted standard values were then totaled to form a Composite Matriculation Score, which was used in the calculations performed in the present study.

Table 9: Weighted Standard Values for Matriculation Results

Grade level	Symbol	Weighted Standard Value
Higher	A	8
	B	7
	C	6
	D	5
	E	4
	F	3
	FF	2
	G	1
Standard	A	6
	B	5
	C	4
	D	3
	E	2
	F	1
Lower	A	4
	B	3
	C	2
	D	1

Academic Performance. First year performance was operationalised in two ways in the present study. Firstly, the average mark for the modules that the learners took in the first year of their degree programmes was computed. This average mark was used when the relationships between performance on the Accuplacer CPTs, matriculation and academic performance were computed. The principal disadvantage

that could impact negatively on the average mark is that modules may differ in difficulty from faculty to faculty, and this was not controlled in the study. Secondly, the percentage of modules passed was categorized into three classes, namely less than 50 percent of modules passed, 50 percent and more passed, and all modules passed. This was used when the clusters derived were internally validated.

Procedure

The data for this study was collected as part of an admissions and placement assessment programme that took place between November 1998 and February 1999 at a tertiary institution in the Eastern Cape. Prior to commencing testing, the students who were assessed during that period were informed about the purposes of the assessment, and that the placement assessment programme would be thoroughly researched. They then signed consent forms to the effect that their results could be used for research purposes.

There were 193 participants extracted from an existing database, namely, the placement assessment records of a tertiary institution in the Eastern Cape. The sample was comprised of a group of first year learners who were registered for degree programmes. Only those learners who had completed at least their first year modules, whether successfully or not, were included in the sample. The sample was then divided into two groups, depending on the degree programme for which students were registered, the distinction being made on the basis of whether or not Mathematics as a matriculation subject was a prerequisite for admission to the relevant programmes.

The test scores used in this study were comprised of numerical scores on certain tests of the Accuplacer CPTs, namely, Arithmetic, Elementary Algebra, and Reading Comprehension. Matriculation performance was operationalised by computing a Composite Matriculation Score. In addition, academic performance was operationalised by calculating an average mark across all modules as well as the number of modules passed. Data was analyzed using the Statistica statistical software package.

Statistical Analysis

Separate statistics were computed and reported for the two groups of learners utilised. This study was primarily exploratory and descriptive in nature. Therefore, descriptive statistics were computed and reported. Specifically, means, medians,

standard deviations, ranges, and the skewness of the distributions of the Accuplacer, matriculation, and academic performance variables were examined.

Prior to conducting further analyses, the use of parametric and nonparametric techniques was considered. Parametric techniques require that assumptions be made about the population sampled relative frequency distributions whereas nonparametric techniques require few, or no, assumptions to be made in this regard (Huysamen, 1997; Mendenhall, 1993). Specifically, the assumptions that need to be met in order for parametric techniques to be employed are that a) a probability sampling strategy was applied in data collection; b) scores on the dependent variable are independent; c) population distributions are normally distributed; and d) populations involved have equal variances (Huysamen, 1997). Parametric techniques tend to be more powerful than non-parametric techniques because they make maximum use of all information contained in normally distributed data sets (Runyon & Haber, 1991).

Although not all of the assumptions for employing parametric techniques were met in totality, it was decided that these statistical methods would be utilized. There is no nonparametric equivalent for cluster analysis and a number of the other statistics based on the cluster analysis have no nonparametric equivalents. Therefore, for the purposes of consistency, it was decided to utilize parametric procedures throughout this study.

Relationships between Accuplacer CPTs scores and matriculation results, and between Accuplacer CPTs scores, matriculation results and academic performance were investigated by calculating correlation coefficients. Pearson product moment correlation coefficients, also known as Pearson r, were considered appropriate when two variables were correlated with each other, as all the measures were on an interval scale of measurement, the underlying distributions are continuous, and all contain linear characteristics. In addition, the coefficient is independent from particular scale values, allowing the investigation of relationships among a variety of variables. Furthermore, although this coefficient is a parametric technique, it may legitimately be computed if the distributions are unimodal and fairly symmetrical (Runyon & Haber, 1991). Where the Accuplacer CPTs scores and matriculation performance were jointly correlated with academic performance, multiple correlations were computed.

One consideration when interpreting correlation coefficients is that they reflect only the linear relationship between variables. Although, low correlations could mean that the variables are unrelated, it is also possible that the variables are related nonlinearly (Mendenhall, 1993; Runyon & Haber, 1991).

In addition to calculating Pearson product moment correlation coefficients and multiple correlations, coefficients of determination (i.e., the ratio of explained variation to total variation, also known as r^2) were calculated to investigate the percentages of variation not accounted for by the variables utilized, and the degree to which the information supplied by variables overlap (Mendenhall, 1993; Runyon & Haber, 1991).

In order to identify patterns of performance of learners entering tertiary studies on the admissions and placement assessment battery, an exploratory-descriptive multivariate cluster analytic procedure was considered the appropriate technique to employ.

Cluster analysis is the statistical procedure of classifying observations into groups on the basis of certain predetermined selection characteristics. Ultimately, successful cluster analysis yields groups or clusters that exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity (Hair, Anderson, Tatham & Black, 1995; Statsoft, 1984-1995). Specifically, a non-hierarchical partitioning technique (Everitt, 1974), known as K-means cluster analysis, was utilized (Anderberg, 1973; Everitt, 1974; Hair et al., 1995; Statsoft, 1984-1995).

The term “K-means “ originated with MacQueen (1967, in Anderberg, 1973 and Everitt, 1974), and denotes the process of assigning every unit of data to the cluster with the nearest centroid (mean). The purpose is thus to divide N observations with P dimensions (variables) into K clusters in order to minimise variability within clusters and maximise variability between clusters (Everitt, 1974; Statsoft, 1984-1994). The K-means algorithm (i.e., a completely defined, finite set of steps, operations, or procedures that will produce a particular outcome) was considered the most appropriate clustering procedure because previous research (e.g., Foxcroft, 1999 and Koch, Foxcroft & Watson, 1999) had already indicated the number of clusters that could be expected from the cases in the sample (Statsoft, 1984-1995).

The optimising procedure was utilised for determining which initial centre or starting point (cluster seed) should be selected to begin the assignment of

observations to clusters. This method allows for reassignment of objects to a more similar or closer cluster if it becomes evident that, in the course of assigning objects, an observation moves closer to another cluster that is not the one to which it was originally assigned (Hair et al., 1995). Early research by MacQueen (1967, in Anderberg, 1973) revealed that the ordering of data units has only a marginal effect on cluster groupings when clusters are well separated, and that differences from one ordering to the next arise from ambiguities created by data units that fall between clusters. Initial cluster centres were computed by sorting the distances between all the observations, then selecting observations at constant intervals.

Inferential statistical analyses were conducted to provide some internal validation for the clusters identified. A Multivariate Analysis of Variance (MANOVA) allowed comparisons of the cluster profiles. MANOVAs are computed on the basis of certain underlying assumptions, two important ones being that the dependent variables are normally distributed within the groups, and that individual group variance-covariance matrices of the dependent variables are equal, thus preliminary tests, were conducted to determine that these assumptions were viable.

Investigating whether there were deviations from normality included conducting statistical tests and examining graphical plots. One of the most common statistical tests utilized for this purpose is a modification of the Kolmogorov-Smirnov, the basic procedure involving calculation of the significance level for the differences from a normal distribution. Such tests are not as useful for smaller samples and tend to be very sensitive in large samples, therefore they require the use of graphical plots to enhance the accuracy of assessing deviations from normality (Hair, Anderson, Tatham & Black, 1995).

Equal variance dispersion across groups is most commonly assessed on two levels. First, by using the Levene test, which examines the variance of one variable across any number of groups. This was done for each of the variables on which the cluster analysis was conducted. Second, by calculating Box's M test, which is multivariate and incorporates the comparison of the equality of variance/covariance matrices. Violation of this assumption requires that adjustment be made for their effects. The variance-covariance matrix is then examined to determine the group that contains the greatest variance. If the greater variances are within the larger group sizes, the alpha level is overstated and differences must be assessed using a lower value. The opposite is true when the greater variance is identified in the smaller

groups; the power of the test is reduced and the significance level must then be increased (Hair, Anderson, Tatham & Black, 1995).

The multivariate statistic known as Wilks' Lambda was then computed; this statistic is the multivariate extension of R-squared, which is associated with multiple regression. An F-approximation to Wilks' Lambda (i.e., a transformed value), known as Rao's R, was used to determine significance. In terms of interpreting Wilks' Lambda, it must be noted that the values range between zero and one, where values near one are not significant, and values near zero tend towards significance. It deserves mention that the F-statistic is remarkably robust to deviations from normality and violations of the assumption of homogeneity of variance, though this is not necessarily the case for Wilks' Lambda, thus significant univariate effects were carefully scrutinized.

Single-factor ANOVAs were computed to determine on which specific test variables the clusters differed significantly from each other. For each variable included in the cluster analysis, the variation among the means for the clusters was compared with random variation of the scores within the groups (Mendenhall, 1993).

Post hoc analyses, using Scheffe's test, were also undertaken in order to establish how clusters differed from each other on each of the variables. Although such post-hoc tests simplify comparisons, their power is quite low, but in comparison of with other post-hoc tests of significance, Scheffe's test is most conservative with respect to Type I error (Hair, Anderson, Tatham & Black, 1995).

Demographic variables (i.e., age, gender, culture group and home language) and cluster groups were cross-tabulated for each of the sample subsets in order to obtain a comprehensive description of cluster groups. Academic performance was categorized in two ways (i.e., percentage of subjects passed and average mark for the first academic year of the degree programme), and cluster groups and academic performance was cross-tabulated for each sample subset.

Cross-tabulation of categorical data basically yields a contingency table, which contains frequencies of scores that fall into each cell of the matrix (Hair, Anderson, Tatham & Black, 1995). The purpose of such cross-tabulation is to investigate the relationship between two categorical variables. Usually, columns represent groups and rows represent categories of the measured variable. Specifically, interest is in whether proportions falling in the categories for one variable are dependent upon the categories of the second variable. The actual cell count, called the observed cell

count, is then compared with the count that would be expected if the categorical variables were independent, called the expected cell count, for each cell. The greater the difference between observed and expected cell counts, the greater the evidence that the variables are dependent (Mendenhall, 1993; Siegel & Castellan, 1988). However, in this study, the cross-tabulations were used solely for yielding a more comprehensive description of cluster groups and no inferential statistics were computed.

The final chapter presents the results of the research and a discussion of the implications of the findings, despite the limitations of the design, sampling procedure, measures and statistics computed. Suggestions are also made for future research.

CHAPTER EIGHT: RESULTS AND DISCUSSION

This chapter incorporates the presentation and discussion of the findings of the present study. Performance on the Accuplacer CPTs, matriculation results in the form of a composite matriculation score (CMS), and academic performance will be described, and thereafter follows a presentation of the relationships among Accuplacer test scores, matriculation results, and academic performance. The results of the cluster analysis are presented together with relevant inferential statistics. The results from the investigation into the relationship between cluster groupings and academic performance, and cluster groupings and demographic variables, are also presented and discussed.

As described in chapter seven, prior to conducting the analyses, the sample was divided into two groups, namely, non-mathematics based degree programmes, comprised of learners registered for programmes in Health Sciences (excluding Pharmacy), Arts, Education, Law, and mathematics based degree programmes, comprised of learners registered for programmes in the Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy. This distinction was made on the basis of the different admissions requirements for various degree programmes relating to mathematics. The aims of the study, as presented in chapter five, were formulated for each of these groups separately. Consequently, the findings, as outlined in the first paragraph, will be reported separately for each group. The findings for the non-mathematics based degrees group are presented first (aims 1 to 3), followed by those for the mathematics based degrees group (aims 4 to 6).

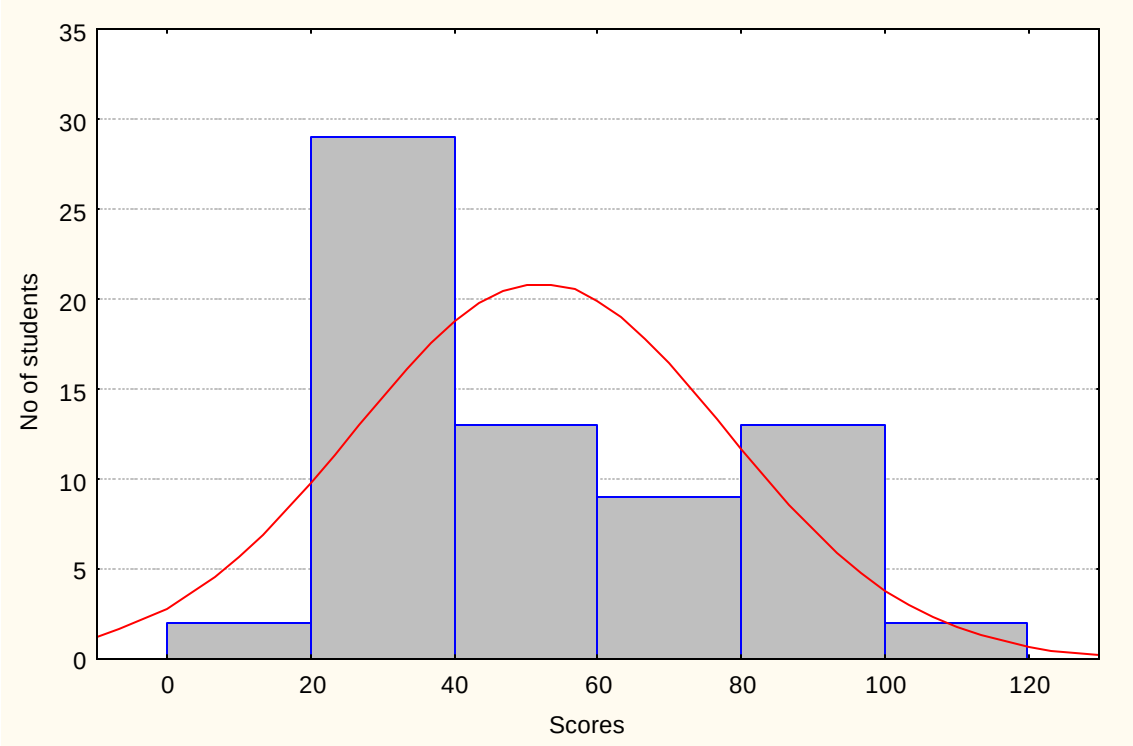
In the final section of the chapter the limitations of the study are examined and suggestions are provided for future research.

Findings for the Non-Mathematics Based Group

Descriptive Statistics for the Non-Mathematics Based Group

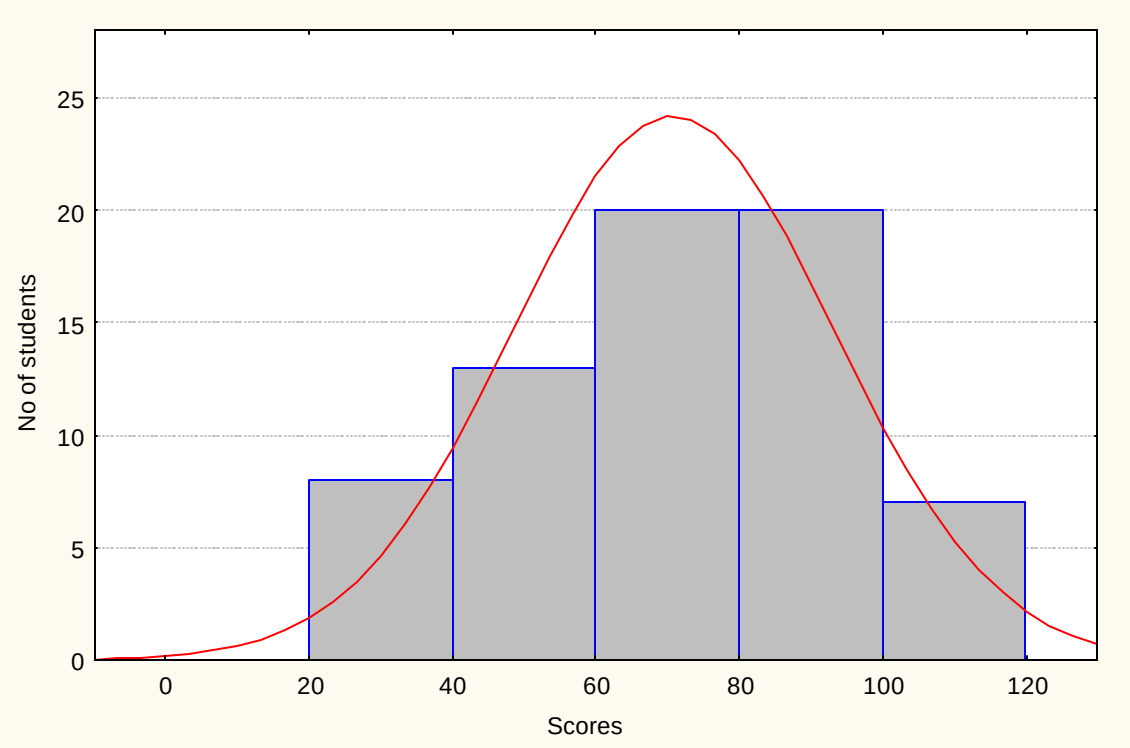
The distribution of scores on the Arithmetic subtest of the Accuplacer for the non-mathematics based group was moderately positively skewed, with a mean of 52 (SD = 25.99) and a median of 48. The scores ranged between 20 and 104. Figure 10 provides a graphical representation of the distribution of scores.

Figure 10: Distribution of Arithmetic Scores for the Non-Mathematics Based Group



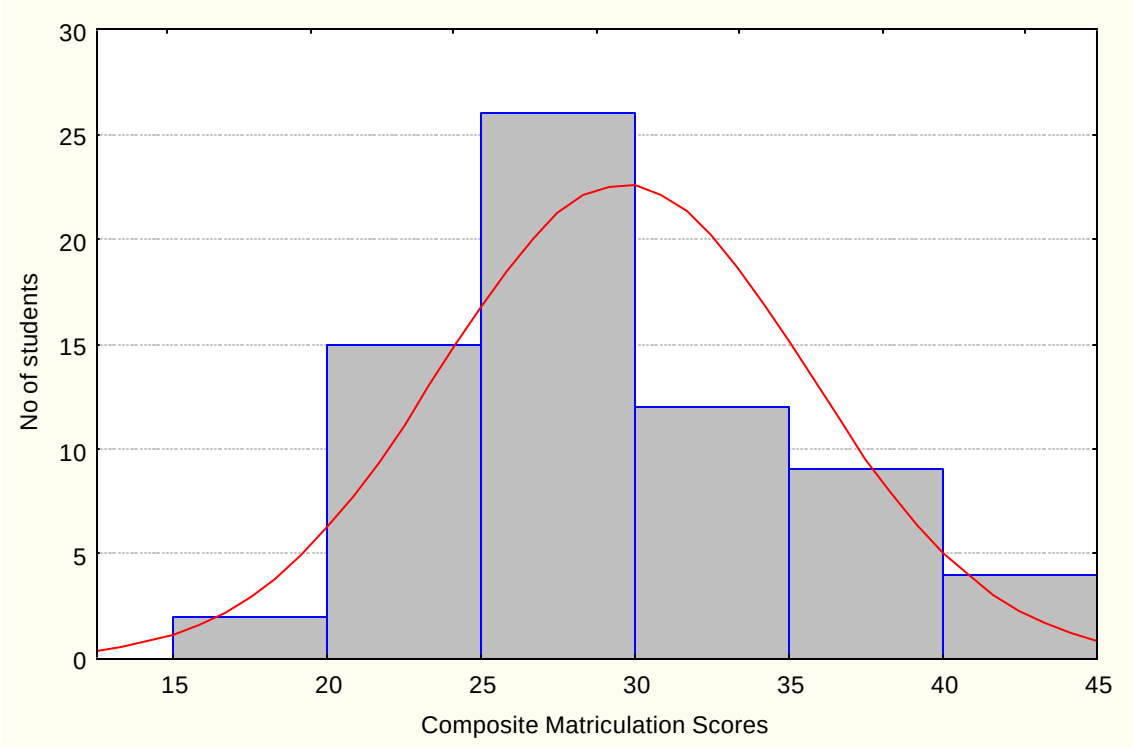
The distribution of scores on the Reading Comprehension subtest of the Accuplacer for the non-mathematics based group was slightly negatively skewed, with a mean of 70.76 (SD = 22.42) and a median of 72. The scores ranged between 27 and 113. Figure 11 provides a graphical representation of the distribution of scores.

Figure 11: Distribution of Reading Comprehension Scores for the Non-Mathematics Based Group



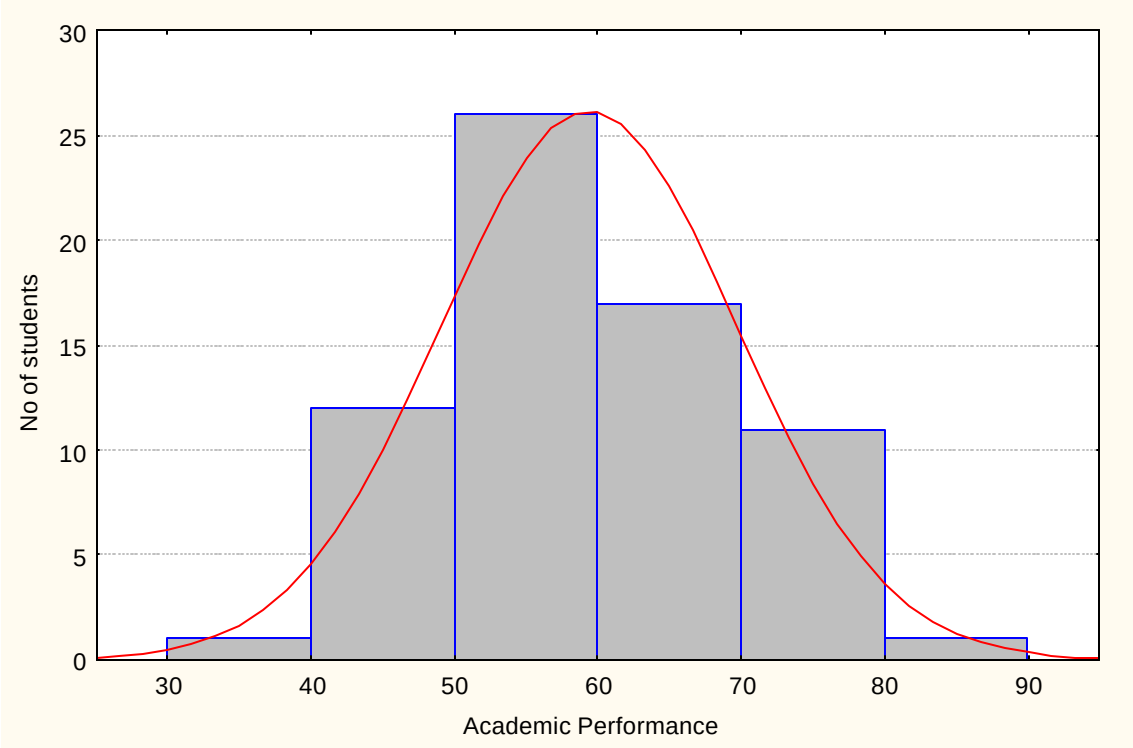
The distribution of scores on the Composite Matriculation Score for the non-mathematics based group was moderately positively skewed, with a mean of 29.63 (SD = 5.99) and a median of 29. The scores ranged between 18 and 44. Figure 12 provides a graphical representation of the distribution of scores.

Figure 12: Distribution of Composite Matriculation Scores for the Non-Mathematics Based Group



The distribution of scores on academic performance for the non-mathematics based group was slightly positively skewed, with a mean of 59.37 (SD = 10.36) and a median of 58.59. The scores ranged between 37.74 and 83.64. Figure 13 provides a graphical representation of the distribution of scores.

Figure 13: Distribution of Academic Performance for the Non-Mathematics Based Group



Correlational Analyses for the Non-Mathematics Based Group

Table 10 depicts the results of the correlational analyses between Accuplacer scores and matriculation results for the non-mathematics based group. The subtest of Elementary Algebra is not administered to applicants for non-Mathematics based degree programmes as information on applicants' mathematical knowledge is not required. Rather the information that the Arithmetic test provides on basic numeracy proficiency is sufficient for development and placement purposes.

Table 10: Correlations Between Accuplacer Scores and Composite Matriculation Scores (CMS) for the Non-Mathematics Based Group (n = 68)

	Arithmetic		Reading Comprehension	
	r	r^2	r	r^2
CMS	.45*	.2	.49*	.24

*p < .05

Table 10 indicates that there is a significant moderately positive relationship

between the Accuplacer Arithmetic and Reading Comprehension scores and the Composite Matriculation Scores obtained by learners in the faculties of Health Sciences (excluding Pharmacy), Arts, Education, and Law. The overlap between matriculation results and Arithmetic and Reading Comprehension was 20 percent and 24 percent respectively. This suggests that the Accuplacer tests, while showing something in common with matriculation results, contribute unique information on incoming learners.

Table 11 depicts the results of the correlational analyses between Accuplacer scores, the Composite Matriculation Scores (CMS) and average first-year performance for the non-mathematics based group.

Table 11: Correlations Between Accuplacer Scores, Composite Matriculation Scores (CMS) and First Year Academic Performance for the Non-Mathematics Based Group (n = 68)

	Arithmetic		Reading Comprehension		CMS	
	r	r^2	r	r^2	r	r^2
Academic Performance	.23	.05	.40*	.16	.51*	.26

*p < .05

Table 11 shows that there is a non-significant small positive relationship between Arithmetic and academic performance, and a significant moderately positive relationship between Reading Comprehension and Composite Matriculation Scores and academic performance respectively for learners in the faculties of Health Sciences (excluding Pharmacy), Arts, Education, and Law. The overlap between academic performance and Arithmetic, Reading Comprehension and matriculation results was five percent, sixteen percent and 26 percent respectively, indicating that each of these provides some information about first year academic performance.

Table 12 depicts the results of the multiple correlational analysis between Accuplacer scores and composite matriculation scores, and average first-year performance for the non-mathematics based group. Although scores on the Arithmetic test were not found to correlate significantly with academic performance for this group, this test was included in the multiple correlational analysis as it was considered important to include a numerical aspect in the calculations and this test is

used to make admissions, placement, and development decisions for this group.

Table 12: Multiple Correlation for Accuplacer Scores and Composite Matriculation Scores (CMS) with First Year Academic Performance for the Non-Mathematics Based Group (n = 68)

	Arithmetic, Reading Comprehension and CMS	
	R	R^2
Academic Performance	.54*	.30

*p < .05

The multiple correlation indicates that for this group, the relationship between the combination of Accuplacer scores and matriculation results with academic performance is moderately positive and accounts for 30 percent of the variation in academic performance. When compared with the bivariate correlations between academic performance and Accuplacer CPTs and matriculation performance respectively reported in table 11, four percent more of the variation in academic performance is explained when test and matriculation results are combined. This supports the literature reported in chapter two, which suggested that the combination of test and school results provide a better prediction of academic performance than either of the two on their own (Badenhorst, Foster & Lea, 1990; Burke, 1982; Calitz, 1997; Kotze, 1994; Venter, 1993).

Cluster Analysis Results for the Non-Mathematics Based Group

The results of the cluster analysis that was performed for the non-mathematics based group are reported in table 13.

Table 13: Number of Observations Per Cluster for the Non-Mathematics Based Group

Cluster	Number of observations	Percentage of Sample (n = 68)
1	25	36.76
2	18	26.47
3	25	36.76

Three clusters were thus identified in the data set. Table 14 presents the means on each variable for each of the cluster groupings.

Table 14: Average Scores on the Variables for the Cluster Groupings of the Non-Mathematics Based Group

	Cluster 1		Cluster 2		Cluster 3	
Variable	M	SD	M	SD	M	SD
Arithmetic	81.88	12.92	37.5	11.83	32.56	11.79
Reading Comprehension	86.12	14.12	83.72	10.7	46.08	10.82
CMS	33.48	6.12	28.17	5.88	26.84	3.66

Initial Descriptions of Cluster Groups

Upon examination of the descriptive information for each cluster (see table 14), the clusters were labelled and described as follows:

Cluster 1. This group is a low risk group, having performed well on the tests of the Accuplacer CPTs. Their arithmetic skills are average. While their receptive language skills are also average, they are slightly better than their numerical skills. Their Composite Matriculation Scores (CMS) indicate that they were average achievers during their scholastic careers.

Cluster 2. This is an average risk group of learners whose performance on the tests of the Accuplacer CPTs is mixed. Specifically, their arithmetic skills are far below average whereas their receptive language skills are average. Their Composite Matriculation Scores fell below the criteria for automatic admission.

Cluster 3. This is a high risk group who performed far below average on the tests of the Accuplacer CPTs. As a group, they lack proficiency insofar as arithmetic and receptive language skills are concerned. Their Composite Matriculation Scores indicate that they performed less well during their scholastic careers.

Internal Validation of Clusters

A MANOVA was performed to explore whether there were differences between the cluster means. Prior to this, however, preliminary investigations were conducted to determine that the assumptions underlying MANOVA computations were viable. First, it was determined whether the dependent variables were normally distributed within the groups, and this was accomplished by examining graphical plots to assess the actual degree of departure from normality. Graphical analysis indicated that the distributions of Arithmetic and Reading Comprehension deviated from normality whereas that of the Composite Matriculation Score was approximately normal.

However, Monte Carlo studies have shown that the violation of this particular assumption does not necessarily constitute as severe a problem as previously thought.

Second, and perhaps more important, statistical tests were conducted to determine whether individual group variance-covariance matrices of the dependent variables are equal. Results indicate that the assumption of homogeneity of variance was not violated. Box's M, an extremely sensitive test for homogeneity of the variance/covariance matrices (Hair, Anderson, Tatham & Black, 1995) was not significant, Box's M = 15.51, $p > 0.05$. Although there was a significant difference for the Composite Matriculation Score, there were no significant differences on the remaining variables on Levene's univariate test for this assumption. Table 15 presents the results for the tests for the univariate test for homogeneity of variance.

Table 15: Levene's Test for Homogeneity of Variance for Dependent Variables in the Non-Mathematics Based Group (df = 2, 65)

Variable	F	p
Arithmetic	0.28	0.76
Reading Comprehension	1.77	0.18
CMS	3.45	0.04*

Note. *$p < .05$.

Statistical analysis proceeded because the F-statistic (although not necessarily Wilks' Lambda) is actually remarkably robust to deviations from normality and violations of the assumption of homogeneity of variance (Statsoft, 1994-1995), and deviations were minimal.

The results of the MANOVA were significant, F = 0.07 (6, 126), $p < 0.05$. In order to provide further internal validation of the clusters, in terms of determining on which specific test variables clusters differed significantly from each other, single factor ANOVAs were conducted on each of the variables.

Table 16 presents the ANOVA results, where it can be seen that significant differences were found on all the variables among the cluster groups.

Table 16: Analysis of Variance for the Clusters in the Non-Mathematics Based Group

Variable	F-ratio	Prob F=0
Arithmetic	118.83	0.00*
Reading Comprehension	82.25	0.00*
CMS	10.85	0.00*

Note. *p < .05.

Based on the significant results of the ANOVAs, it was decided to conduct post-hoc analyses, using Scheffe's test, to determine how each cluster differed on each variable where significant differences were identified. The results of the post-hoc analyses are presented in Table 17.

Table 17: Probability Values at p < 0.05 for Cluster Differences on Each Variable for the Non-Mathematics Based Group

	Cluster 1	Cluster 2	Cluster 3
		Arithmetic	
Cluster 1		0.00	0.00
Cluster 2	0.00		0.43
Cluster 3	0.00	0.43	
		Reading Comprehension	
Cluster 1		0.82	0.00
Cluster 2	0.82		0.00
Cluster 3	0.00	0.00	
		Composite Matriculation Score	
Cluster 1		0.01	0.00
Cluster 2	0.00		0.72
Cluster 3	0.00	0.72	

Examination of the results overall indicates that the Arithmetic subtest of the Accuplacer and the Composite Matriculation Scores discriminated effectively between the low and average risk learners (clusters 1 and 2) and low and high risk learners (clusters 1 and 3), but less effectively between the average and high risk learners (clusters 2 and 3). The Reading Comprehension subtest of the Accuplacer did not discriminate as well between low and average risk learners (clusters 1 and 2)

as it did between low and high risk (clusters 1 and 3) and between average and high-risk (clusters 2 and 3) learners.

It appears that the Accuplacer CPTs and Composite Matriculation Scores are able to discriminate among groups of learners. Consequently, the information obtained from the Accuplacer CPTs and the Composite Matriculation Score could be a valuable adjunct to guide admissions and development decisions.

Demographic Descriptions of Cluster Groups

Table 18: Cross-tabulation of Cluster Grouping and Age for the Non-Mathematics Based Group

Age in years	Cluster 1	Cluster 2	Cluster 3
17 – 19	23 (33.82)	14 (20.59)	21 (30.88)
20 – 35	2 (2.94)	4 (5.88)	4 (5.88)

As can be seen from table 18, the majority of the learners who were in the age category of 17 to 19 years (75.86 percent), fell into either the low or high risk clusters (clusters 1 and 3). Although there were few representatives of the older age groups (i.e., 20 years and older), the majority of them (80 percent) fell into the average and high risk clusters (clusters 2 and 3).

Table 19: Cross-tabulation of Cluster Grouping and Gender for the Non-Mathematics Based Group

Gender	Cluster 1	Cluster 2	Cluster 3
Male	6 (8.82)	4 (5.98)	6 (8.82)
Female	19 (27.94)	14 (20.59)	19 (27.94)

Table 19 shows that the number of males and females in each cluster were fairly evenly spread. Although there were more females than males in each cluster, there were more females than males in this non-mathematics based group.

Table 20: Cross-tabulation of Cluster Grouping and Culture for the Non-Mathematics Based Group

Cultural Group	Cluster 1	Cluster 2	Cluster 3
Black	2 (2.94)	4 (5.98)	16 (23.53)
Coloured	5 (7.35)	4 (5.98)	5 (7.35)
Indian	2 (2.94)	0	0
White	16 (23.53)	9 (13.24)	4 (5.98)
Chinese	0	1 (1.47)	0

The majority of the Black learners (90.91 percent) fell into the high and average risk groups (clusters 3 and 2). The Coloured learners seemed to be fairly evenly spread across the clusters. The Indian representatives fell into the low risk group only (cluster 1). The majority of the White learners (86.21 percent) fell into the average and low risk groups (clusters 2 and 1). The only Chinese representative fell into the average risk group (cluster 2).

Table 21: Cross-tabulation of Cluster Grouping and Home Language for the Non-Mathematics Based Group

Home Language	Cluster 1	Cluster 2	Cluster 3
English	12 (17.65)	8 (11.76)	2 (2.94)
Afrikaans	8 (11.76)	3 (4.41)	6 (8.82)
English/Afrikaans	3 (4.41)	3 (4.41)	1 (1.47)
Xhosa	2 (2.94)	3 (4.41)	11 (16.18)
Other African Language	0	1 (1.47)	5 (7.35)

As can be seen from table 21, the majority of the English speaking learners (90.91 percent) fell into the average and low risk groups (clusters 2 and 1), and the same was the case (85.71 percent) for those who indicated that their home language was mixed (i.e., both English and Afrikaans). Where the Afrikaans speaking learners were concerned, 47.05 percent of them fell into the low risk group (cluster 1) and 35.29 percent of them fell into the high-risk group (cluster 3). The majority of the Xhosa-speaking learners (87.5 percent) fell into the average and high-risk groups (clusters 2 and 3), and the majority of the learners from the other African languages

grouping (83.33 percent) fell into the high-risk group (cluster 3).

Table 22: Cross-tabulation of Cluster Grouping and Percentage of First Year Modules Passed for the Non-Mathematics Based Group

Percentage of Subjects Passed	Cluster 1	Cluster 2	Cluster 3
Less than half	1 (1.47)	2 (2.94)	4 (5.88)
More than half	11 (16.18)	9 (13.24)	19 (27.94)
All	13 (19.12)	7 (10.29)	2 (2.94)

Table 22 shows that the majority of learners in the low risk (96 percent) and average risk (88.89 percent) clusters (1 and 2 respectively) passed most or all of their modules. However, the majority of those learners in the high-risk cluster group (76 percent) (cluster 3) passed most of their modules. In order to investigate this aspect further, academic performance was approached from the perspective of average performance categories.

Table 23: Cross-tabulation of Cluster Grouping and First Year Academic Performance for the Non-Mathematics Based Group

Categories of Averages	Cluster 1	Cluster 2	Cluster 3
Less than 50 percent	4 (5.88)	4 (5.88)	5 (7.35)
50 – 59 percent	5 (7.35)	5 (7.35)	16 (23.53)
60 – 69 percent	8 (11.76)	6 (8.82)	3 (4.41)
70 – 74 percent	4 (5.88)	1 (1.47)	1 (1.47)
75 percent and above	4 (5.88)	2 (2.94)	0

As can be seen in table 23, the majority of learners in the in the low risk cluster (84 percent) (cluster 1) passed overall, but only a small percentage of these (16 percent) can be said to have excelled. Also, the majority of learners in the average risk cluster (77.78 percent) (cluster 2) passed overall, and a small percentage of them (11.11 percent) excelled. The majority of learners in the high-risk cluster group (80 percent) (cluster 3) passed overall, but none can be said to have excelled.

Summary Comments on the Cluster Groups

Cluster 1: This low risk group of learners produced average scores on the tests of the Accuplacer and in their matriculation results. The majority passed most or all of their subjects and obtained an average final mark for their first year that was over 50 percent. These first year results indicate that the learners in this group are indeed predominantly average achievers. Certain cognitive and non-cognitive factors not considered in this study may have contributed to the spread of final averages across this group of learners.

Cluster 2: This average risk group of learners produced mixed results on the tests of the Accuplacer, where their numeracy skills were far below those of their reading and comprehension skills, and their matriculation results were below average. The majority of these learners passed most or all of their subjects and obtained an average final mark for their first year that was over 50 percent. These first year results indicate that the learners in this group would probably benefit from developmental assistance specifically to improve their numeracy skills. Also, other cognitive and non-cognitive factors not included in this study may have contributed to the spread of final averages across this group of learners.

Cluster 3: This high-risk group of learners produced far below average scores on the tests of the Accuplacer and in their matriculation results. The majority passed most of their subjects and obtained an average final mark for their first year that was over 50 percent. Of these learners who passed their first year at university, most of them can be considered borderline learners, indicating that this group is comprised of learners who would most likely benefit from intensive developmental assistance to improve their language and numeracy skills, although other cognitive and non-cognitive factors not included in this study may play a role in these results.

The fact that three cluster groupings were identified in the data set and that they were described as being low, average, and high risk clusters respectively, corroborates previous findings by Foxcroft (1999) at the same tertiary institution using the same measures.

Summary Comments on the Findings for the Non-Mathematics Based Group

As regards the first aim of the study, a significant, moderate relationship was found between matriculation performance, as operationalised by the Composite Matriculation Score, and scores on the Arithmetic and Reading Comprehension tests of the Accuplacer CPTs for learners in the faculties of Arts, Health Sciences,

Education and Law. By examining the percentage of overlap between the variables, these findings further suggest that the test results and matriculation performance contribute a certain degree of unique information, which could be useful when trying to predict learner performance and could assist in identifying development needs.

In terms of the second aim of the study, scores on the Reading Comprehension test and matriculation performance were found to correlate significantly with average first-year academic performance. Furthermore, when scores on the Arithmetic and Reading Comprehension tests together with the Composite Matriculation Score were correlated with average first-year academic performance, a significant moderate relationship was found and more of the variation in academic performance was explained than when each of the predictor variables were used on their own. These findings are encouraging. Not only do they add to the predictive validity data being gathered on the Accuplacer CPTs in South Africa, but they corroborate previous research findings which suggest that school and test performance together provide a better prediction of academic performance than test or school performance on their own (Badenhorst, Foster & Lea, 1990; Burke, 1982; Calitz, 1997; Kotze, 1994; Venter, 1993).

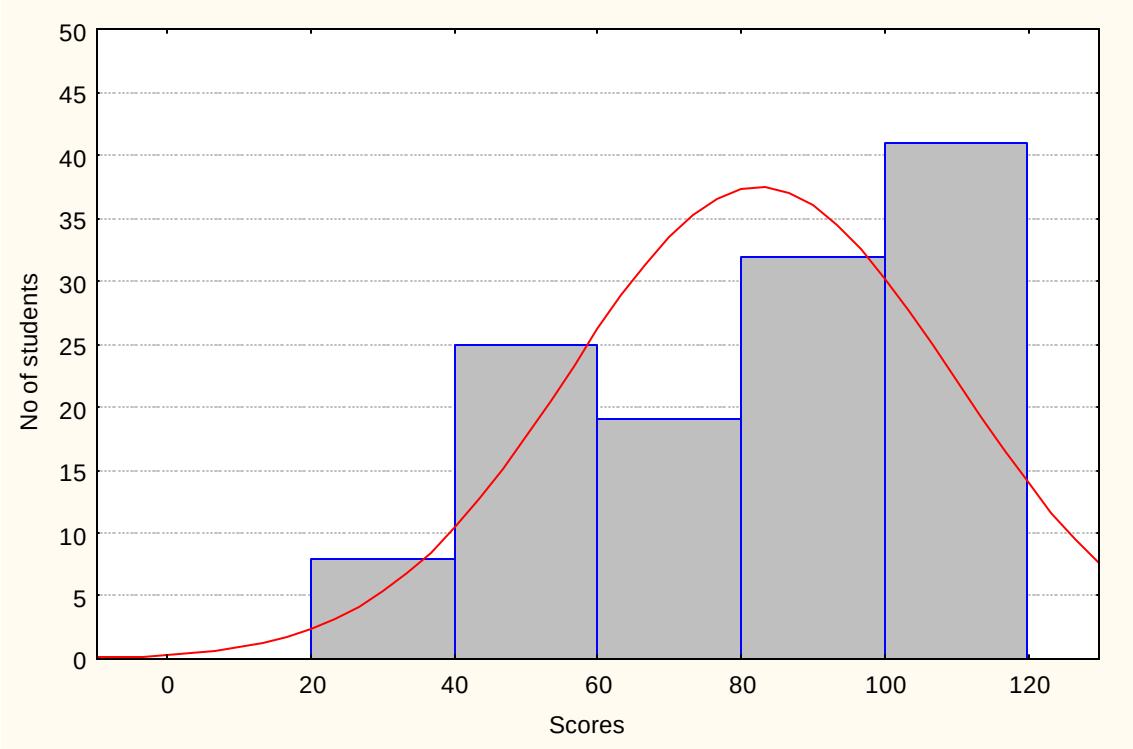
In terms of the third aim of the study, underlying patterns of performance were identified in the Accuplacer CPTs and matriculation performance, which was found to be related to academic performance for learners in the faculties of Arts, Health Sciences, Education and Law. By classifying learners' performance using these underlying patterns, valuable information regarding the development needs of first-year learners can be provided.

Findings for the Mathematics Based Group

Descriptive Statistics for the Mathematics Based Group

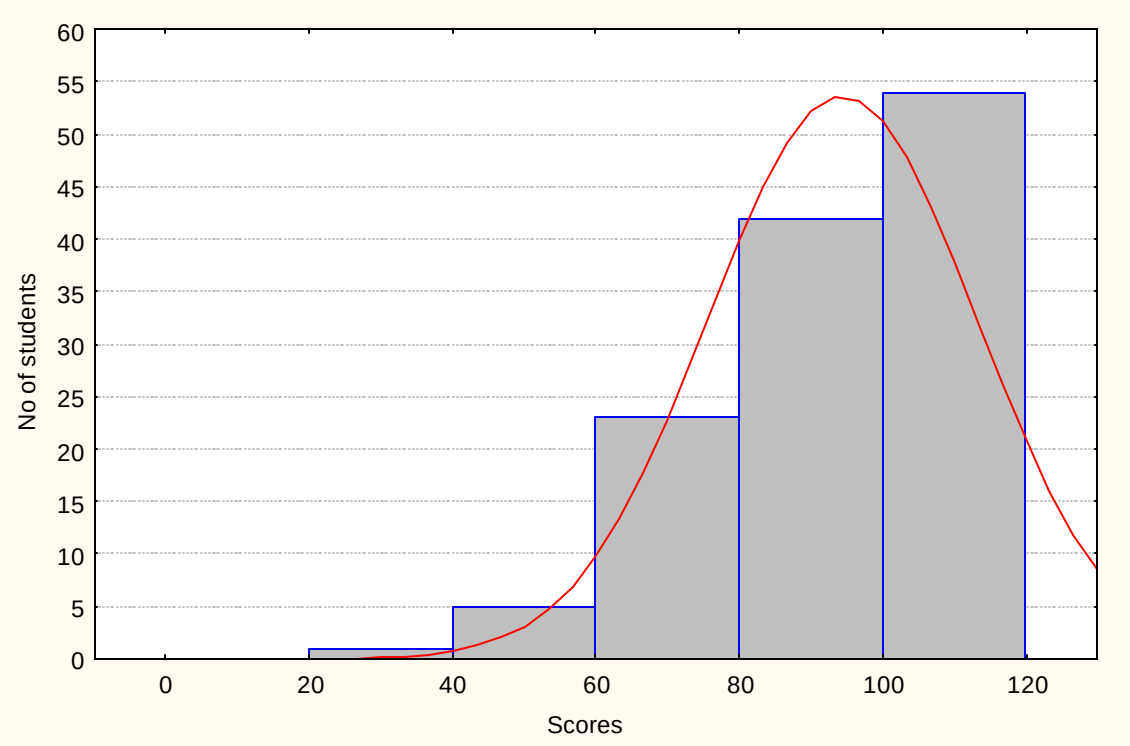
The distribution of scores on the Arithmetic test of the Accuplacer for the mathematics based group was moderately negatively skewed, with a mean of 82.59 (SD = 26.64) and a median of 85. The scores ranged between 24 and 120. Figure 14 provides a graphical representation of the distribution of scores.

Figure 14: Distribution of Arithmetic Scores for the Mathematics Based Group



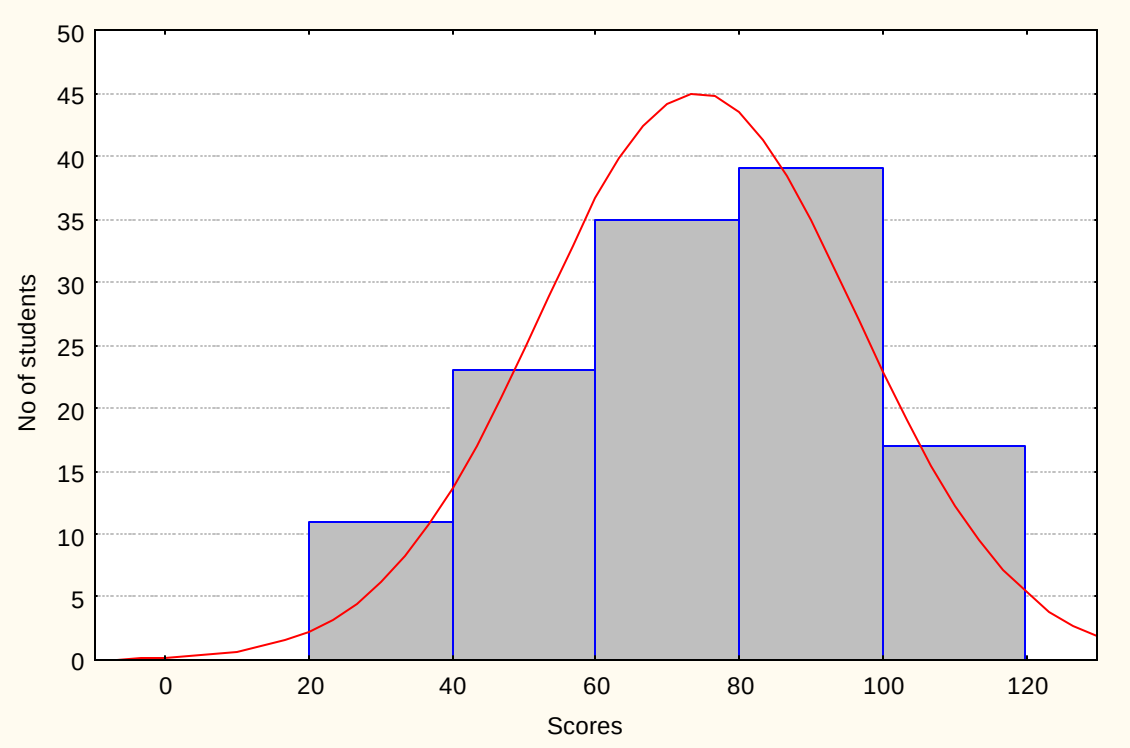
The distribution of scores on the Elementary Algebra test of the Accuplacer for the mathematics based group was negatively skewed, with a mean of 94.36 (SD = 18.59) and a median of 97. The scores ranged between 29 and 120. Figure 15 provides a graphical representation of the distribution of scores.

Figure 15: Distribution of Elementary Algebra Scores for the Mathematics Based Group



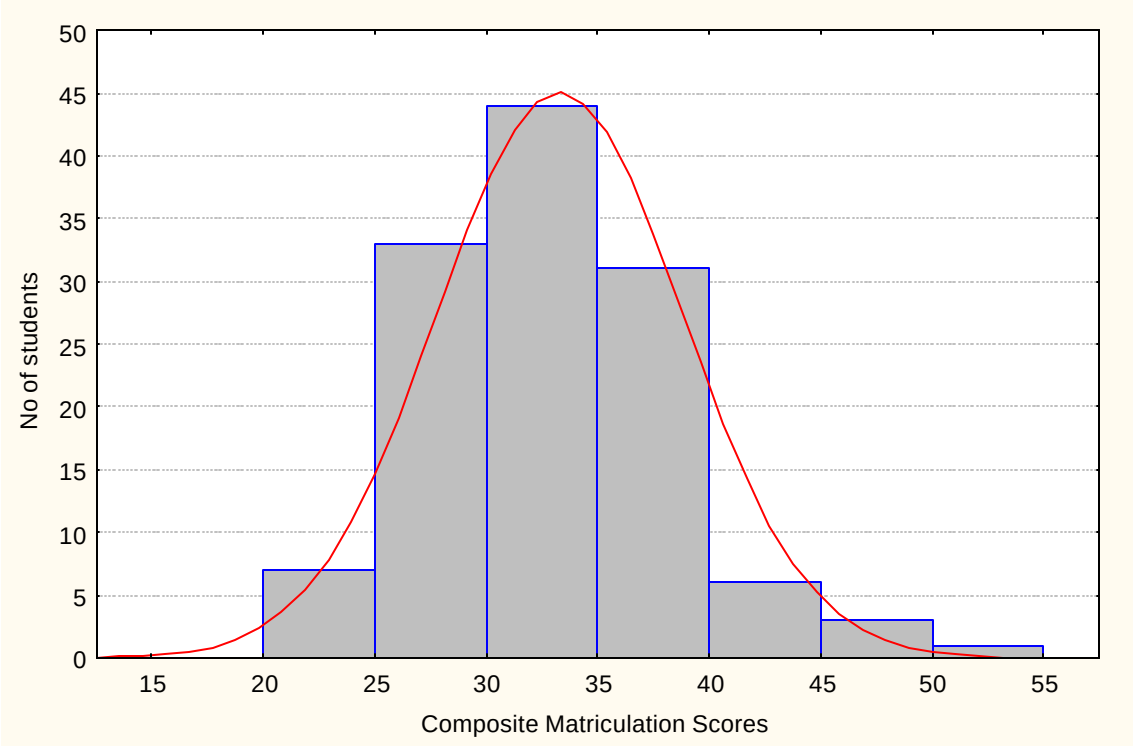
The distribution of scores on the Reading Comprehension test of the Accuplacer for the mathematics based group was slightly negatively skewed, with a mean of 74.22 (SD = 22.17) and a median of 75. The scores ranged between 28 and 120. Figure 16 provides a graphical representation of the distribution of scores.

Figure 16: Distribution of Reading Comprehension Scores for the Mathematics Based Group



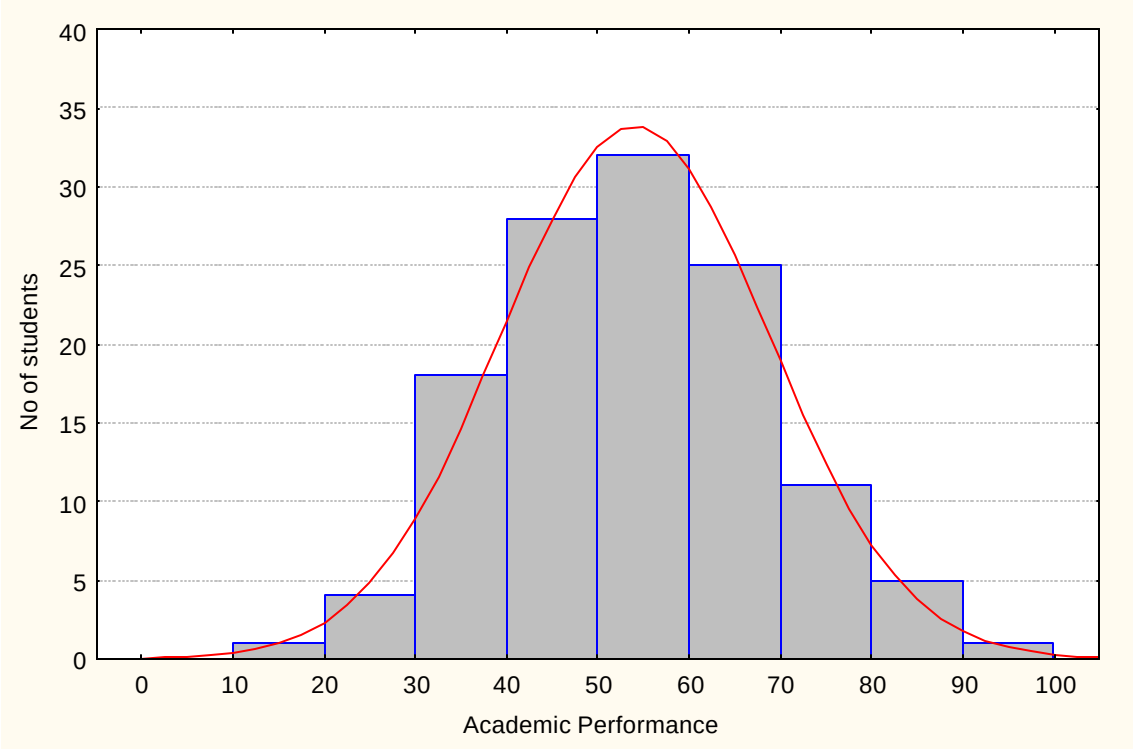
The distribution of scores on the Composite Matriculation Score for the mathematics based group was moderately positively skewed, with a mean of 33.29 (SD = 5.54) and a median of 32. The scores ranged between 21 and 53. Figure 17 provides a graphical representation of the distribution of scores.

Figure 17: Distribution of Composite Matriculation Scores for the Mathematics Based Group



The distribution of scores on academic performance for the mathematics based group was slightly positively skewed, with a mean of 54.08 (SD = 14.75) and a median of 54.88. The scores ranged between 15.5 and 92.52. Figure 18 provides a graphical representation of the distribution of scores.

Figure 18: Distribution of Academic Performance for the Mathematics Based Group



Correlational Analyses for the Mathematics Based Group

Table 24 depicts the relationship between performance on the Accuplacer CPTs and matriculation performance for the mathematics based group.

Table 24: Correlations Between Accuplacer Scores and Composite Matriculation Scores (CMS) for the Mathematics Based Group (n = 125)

	Arithmetic		Elementary Algebra		Reading Comprehension	
	r	r^2	r	r^2	r	r^2
CMS	.42*	.18	.45*	.20	.54*	.29

*p < .05

Table 24 indicates that there is a significant moderately positive relationship between the Accuplacer scores of Arithmetic, Elementary Algebra and Reading Comprehension, and the matriculation results obtained by learners in the faculties of Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy. The overlap between matriculation results and Arithmetic, Elementary Algebra and

Reading Comprehension was eighteen percent, 20 percent and 29 percent respectively. This suggests that the Accuplacer tests, while showing something in common with matriculation results, contribute unique information about learners entering mathematics based programmes.

In table 25, the results of the separate correlations between test and matriculation performance with academic performance are provided.

Table 25: Correlations Between Accuplacer Scores, Composite Matriculation Scores (CMS) and First Year Academic Performance for the Mathematics Based Group (n = 125)

	Arithmetic		Elementary Algebra		Reading Comprehension		CMS	
	r	r^2	r	r^2	r	r^2	r	r^2
Academic Performance	.57*	.33	.58*	.34	.47*	.22	.60*	.35

*$p < .05$

Table 25 shows that there is a significant moderately positive relationship between Arithmetic, Elementary Algebra, Reading Comprehension and matriculation results (Composite Matriculation Score) and academic performance respectively for learners in the faculties of Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy. The overlap between academic performance and Arithmetic, Elementary Algebra, Reading Comprehension and matriculation results was 33 percent, 34 percent, 22 percent and 35 percent respectively, indicating that each of these provides some unique information about first year academic performance.

In Table 26, the results of the multiple correlation between test and matriculation performance jointly with academic performance is provided.

Table 26: Multiple Correlation for Accuplacer Scores and Composite Matriculation Scores (CMS) with First Year Academic Performance for the Mathematics Based Group (n = 125)

	Arithmetic, Reading Comprehension and CMS	
	R	R^2
Academic Performance	.72*	.52

*p < .05

The multiple correlation indicates that for this group, the relationship between the combination of Accuplacer scores and matriculation results with academic performance is strongly positive and explains 52 percent of the variation in academic performance, which is considerable. When compared with the bivariate correlations between academic performance and Accuplacer CPTs and matriculation performance respectively reported in table 25, 17 to 30 percent more of the variation in academic performance is explained when test and matriculation results are combined. This supports the literature reported in chapter two, which suggested that the combination of test and school results provide a better prediction of academic performance than either of the two on their own (Badenhorst, Foster & Lea, 1990; Burke, 1982; Calitz, 1997; Kotze, 1994; Venter, 1993).

Since all correlations with average academic performance were positive and significant for this group, it was considered acceptable to utilize the three Accuplacer tests and the Composite Matriculation Scores in the cluster analysis, the results of which are reported in the next section.

Results for the Cluster Analysis for the Mathematics Based Group

The results of the cluster analysis that was performed for the mathematics based group are reported in Table 27.

Table 27: Number of Observations Per Cluster for the Mathematics Based Group

Cluster	Number of observations	Percentage of Group (n = 125)
1	47	37.6
2	48	38.4
3	30	24

Three clusters were thus identified. Table 28 presents the means on each variable for each of the cluster groupings.

Table 28: Average Scores on the Variables for the Cluster Groupings of the Mathematics Based Group

	Cluster 1		Cluster 2		Cluster 3	
Variable	M	SD	M	SD	M	SD
Arithmetic	108.77	8.03	79.96	12.88	45.8	11.89
Elementary Algebra	109.45	8.14	90.35	13.51	77.13	19.35
Reading Comprehension	91.72	13.21	73	17.24	48.73	13.1
CMS	36.4	6.06	32.42	4.26	29.8	3.7

Initial Descriptions of Cluster Groups

Upon examination of the descriptive information for each cluster (see table 28), the clusters were labelled and described as follows:

Cluster 1. This group is a low risk group, having performed well on the tests of the Accuplacer CPTs. Their arithmetic and algebraic skills are above average, while their receptive language skills are average. Their matriculation performance can be classified as being above average.

Cluster 2: This is an average risk group of learners. Specifically, their arithmetic and algebraic skills are average, with the latter being somewhat better than the former. Their receptive language skills are low average. Their matriculation results indicate that they were average achievers.

Cluster 3: This is a high-risk group whose arithmetic skills are far below average whereas their algebraic skills are low average. Their receptive language skills are also far below average. Their matriculation results indicate that their performance was low to below average.

Internal Validation of Clusters

To internally verify the existence of three clusters, a MANOVA was performed to explore whether there were significant differences between the cluster means. Prior to this, however, preliminary investigations were conducted to determine that the assumptions underlying MANOVA computations were viable. First, it was determined whether the dependent variables were normally distributed within the groups, and this

was accomplished by examining graphical plots to assess the actual degree of departure from normality. Graphical analysis indicated that that the distributions of Arithmetic, Elementary Algebra and Reading Comprehension deviated from normality whereas that of the Composite Matriculation Score was approximately normal. However, Monte Carlo studies have shown that the violation of this particular assumption does not necessarily constitute as severe a problem as previously thought.

Second, and perhaps more important, statistical tests were conducted to determine whether individual group variance-covariance matrices of the dependent variables are equal. Results indicate that the assumption of homogeneity of variance was violated. Box's M, an extremely sensitive test for homogeneity of the variance/covariance matrices (Hair, Anderson, Tatham & Black, 1995) was significant, Box's M = 83.22, $p < 0.05$. There were significant differences for the Arithmetic, Elementary Algebra and Composite Matriculation Scores, but there was no significant difference on Reading Comprehension on Levene's univariate test for this assumption. Table 29 presents the results for the tests for the univariate test for homogeneity of variance.

Table 29: Levene's Test for Homogeneity of Variance for Dependent Variables in the Mathematics Based Group (df = 2, 122)

Variable	F	p
Arithmetic	3.58	0.03*
Elementary Algebra	10.55	0.00*
Reading Comprehension	2.28	0.11
CMS	4.22	0.17*

Note. *$p < .05$.

Statistical analysis proceeded because the F-statistic (although not necessarily Wilks' Lambda) is actually remarkably robust to deviations from normality and the violation of the assumption of homogeneity of variance could simply be reflecting the sensitivity of the Box M test (Statsoft, 1994-1995).

The results of the MANOVA were significant, F = 0.11 (8, 238), $p < 0.05$. In order to provide further internal validation of the clusters, in terms of determining on which specific variables the clusters differed significantly from each other, single factor ANOVAs were conducted on each of the variables.

Table 30 presents the ANOVA results, where it can be seen that significant differences were found on all the variables among the cluster groups.

Table 30: Analysis of Variance for the Clusters in the Mathematics Based Group (df = 2, 122)

Variable	F-ratio	Prob F=0
Arithmetic	300.22	0.00*
Elementary Algebra	55.27	0.00*
Reading Comprehension	76.81	0.00*
CMS	17.79	0.00*

Note. *$p < .05$.

Based on the significant results of the ANOVAs, it was decided to conduct post-hoc analyses, using Scheffe's test, to determine how each cluster differed on each variable where significant differences were identified. The results of the post-hoc analyses are presented in table 31.

Table 31: Probability Values at $p < 0.05$ for Cluster Differences on Each Variable for the Mathematics Based Group

	Cluster 1	Cluster 2	Cluster 3
		Arithmetic	
Cluster 1		0.00	0.00
Cluster 2	0.00		0.00
Cluster 3	0.00	0.00	
		Elementary Algebra	
Cluster 1		0.00	0.00
Cluster 2	0.00		0.00
Cluster 3	0.00	0.00	
		Reading Comprehension	
Cluster 1		0.00	0.00
Cluster 2	0.00		0.00
Cluster 3	0.00	0.00	
		Composite Matriculation Score	
Cluster 1		0.00	0.00
Cluster 2	0.00		0.08
Cluster 3	0.00	0.08	

Examination of the results overall indicates that the Accuplacer tests (i.e., Arithmetic, Elementary Algebra and Reading Comprehension) discriminated effectively between the low and average risk learners (clusters 1 and 2), the high and low risk learners (clusters 3 and 1), and between the average and high-risk learners (clusters 2 and 3). The Composite Matriculation Score did not discriminate as well between low and average risk learners (clusters 1 and 2) as it did between low and high-risk (clusters 1 and 3) and between average and high-risk learners (clusters 2 and 3).

Demographic Descriptions of Cluster Groups

Table 32: Cross-tabulation of Cluster Grouping and Age for the Mathematics Based Group

Age in years	Cluster 1	Cluster 2	Cluster 3
16 - 19	44 (35.2)	47 (37.6)	26 (20.8)
20 - 39	3 (2.4)	1(0.8)	4 (3.2)

As can be seen from table 32, the majority of individuals in the age category of 16 to 19 years (77.78 percent) fell into the low and average risk groups (clusters 1 and 2). The representatives for older age groups are limited in number for each of the clusters, although the majority of this age group (87,5 percent) fell into the high and low risk groups (clusters 3 and 1).

Table 33: Cross-tabulation of Cluster Grouping and Gender for the Mathematics Based Group

Gender	Cluster 1	Cluster 2	Cluster 3
Male	30 (24)	19 (15.2)	7 (5.6)
Female	17 (13.6)	29 (23.2)	23 (18.4)

Table 33 shows that the majority of males (87.5 percent) fell into the low and average risk groups (clusters 1 and 2), whereas the majority of females (75.36 percent) fell into the average and high-risk groups (clusters 2 and 3) for the mathematics based group.

Table 34: Cross-tabulation of Cluster Grouping and Culture for the Mathematics Based Group

Cultural Group	Cluster 1	Cluster 2	Cluster 3
Black	5 (4)	17 (13.6)	26 (20.8)
Coloured	3 (2.4)	13 (10.4)	2 (1.6)
Indian	3 (2.4)	3 (2.4)	1 (0.8)
White	33 (26.4)	15 (12)	1 (0.8)
Chinese	3 (2.4)	0	0

As can be seen from table 34, the majority of the Black learners (89.58 percent) fell into the average and high-risk groups (clusters 2 and 3). The majority of the Coloured learners (72.22 percent) fell into the average risk group (cluster 2). The majority of the Indian (85.71 percent) and White (97.96 percent) learners fell into the low and average risk groups (clusters 1 and 2). All the Chinese learners fall into the low risk group (cluster 1).

Table 35: Cross-tabulation of Cluster Grouping and Home Language for the Mathematics Based Group

Home Language	Cluster 1	Cluster 2	Cluster 3
English	27 (21.6)	14 (11.2)	3 (2.4)
Afrikaans	12 (9.6)	14 (11.2)	0
English/Afrikaans	1 (0.8)	4 (3.2)	1 (0.8)
Xhosa	2 (1.6)	12 (9.6)	23 (18.4)
Other African Language[a]	2 (1.6)	4 (3.2)	3 (2.4)
Other[b]	3 (2.4)	0	0

Table 35 shows that the majority of the English speaking (93.18 percent) and all the Afrikaans speaking learners fell into the low and average risk groups (clusters 1 and 2). The majority of those who indicated that their home language is mixed (i.e., both English and Afrikaans) fell into the average risk group (cluster 2). The majority of the Xhosa speakers (94.59 percent) and those who speak one of the other African languages (77.78 percent) fell into the average and high-risk groups (clusters 2 and 3). All those learners who indicated that they speak some language other than the official languages of South Africa fell into the low risk group (cluster 1).

Table 36: Cross-tabulation of Cluster Grouping and Percentage of First Year Modules Passed for the Mathematics Based Group

Percentage of Subjects Passed	Cluster 1	Cluster 2	Cluster 3
Less than half	2 (1.6)	17 (13.6)	17 (13.6)
More than half	19 (15.2)	24 (19.2)	13 (10.4)
All	26 (20.8)	7 (5.6)	0

Table 36 shows that the majority (95.74 percent) of learners in the low risk

cluster (cluster 1) passed most of their modules whereas half the learners in the average risk cluster (cluster 2) passed most of their modules and more than a third of them (35.42 percent) passed less than half their modules. More than half the learners (56.67) in the high-risk cluster (cluster 3) passed less than half their modules and none of them passed all of their subjects. This aspect was further investigated by approaching academic performance from the perspective of dividing their average performance into categories.

Table 37: Cross-tabulation of Cluster Grouping and First Year Academic Performance for the Mathematics Based Group

Categories of Averages	Cluster 1	Cluster 2	Cluster 3
Less than 50 percent	5 (4)	23 (18.4)	23 (18.4)
50 – 59 percent	9 (7.2)	17 (13.6)	6 (4.8)
60 – 69 percent	17 (13.6)	7 (5.6)	1 (0.8)
70 – 74 percent	7 (5.6)	1 (0.8)	0
75 percent and above	9 (7.2)	0	0

As can be seen from table 37, the majority of learners (89.36 percent) in the low risk cluster (cluster 1) passed overall, but only a minority of them (19.15 percent) can be said to have excelled. A little over half the learners (52.08 percent) in the average risk cluster (cluster 2) passed overall, and none can be said to have excelled. Only a minority of learners (23.33 percent) in the high-risk cluster (cluster 3) passed overall, and none of them excelled.

Summary Comments on the Cluster Groups

Cluster 1: These learners obtained above average scores for numeracy and algebraic proficiency, and average scores on receptive language proficiency on the tests of the Accuplacer. They demonstrated above average performance in matriculation subjects. The majority of these learners passed most of their modules and obtained an average final mark for their first year that was over 50 percent. Certain cognitive and non-cognitive factors not considered in this study may have contributed to the spread of final year results across this cluster group.

Cluster 2: These learners obtained average scores for numeracy and algebraic proficiency, and low average scores on receptive language proficiency on the tests of the Accuplacer. They demonstrated average matriculation performance. Half of this group passed most or all of their modules whereas the opposite occurred for more

than a third of these learners. A little more than half of these learners obtained an average final mark for their first year that was over 50 percent. Certain cognitive and non-cognitive factors not included in this study may have played a role in the spread of final year results evident across this group. It is possible that these learners require some developmental assistance specifically to improve their receptive language proficiency or discipline-specific assistance to facilitate better academic results.

Cluster 3: These learners obtained low to far below average results on the tests of the Accuplacer and in their matriculation performance. More than half of this group passed less than half of their modules and only a minority obtained a final average mark of over 50 percent for their first year of tertiary studies. Certain cognitive and non-cognitive factors not included in this study may have played a role in the final year results evident across this group. It appears that these learners may require intensive developmental preparation to better prepare them for tertiary education.

The fact that three cluster groupings were identified in the data set and that they were described as being low, average, and high risk clusters respectively, corroborates previous findings by Foxcroft (1999) at the same tertiary institution using the same measures.

Summary Comments on the Findings for the Mathematics Based Group

As regards the fourth aim of the study, a significant, moderate relationship was found between matriculation performance, as operationalised by the Composite Matriculation Score, and scores on the Arithmetic, Elementary Algebra and Reading Comprehension tests of the Accuplacer CPTs for learners doing programmes in the Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy. By examining the percentage of overlap between the variables, these findings further suggest that the test results and matriculation performance contribute a certain degree of unique information, which could be useful when trying to predict learner performance and could assist in identifying development needs.

In terms of the fifth aim of the study, scores on the Arithmetic, Elementary Algebra and Reading Comprehension tests and matriculation performance were found to correlate significantly with average first-year academic performance. Furthermore, when scores on the Arithmetic, Elementary Algebra and Reading Comprehension tests together with the Composite Matriculation Score were correlated with average first-year academic performance, a significant moderate relationship was found and considerably more of the variation in academic

performance was explained than when each of the predictor variables were used on their own. These findings are encouraging. Not only do they add to the predictive validity data being gathered on the Accuplacer CPTs in South Africa, but they corroborate previous research findings which suggest that school and test performance together provide a better prediction of academic performance than test or school performance on their own (Badenhorst, Foster & Lea, 1990; Burke, 1982; Calitz, 1997; Kotze, 1994; Venter, 1993).

In terms of the sixth aim of the study, underlying patterns of performance were identified in the Accuplacer CPTs and matriculation performance, which were found to be related to academic performance for learners doing programmes in the Building Disciplines, Economic Sciences, the Natural Sciences and Pharmacy. By classifying learners' performance using these underlying patterns, valuable information regarding the development needs of first-year learners can be provided.

A Summary of the Present Findings

This study has contributed to the body of knowledge on admissions and placement testing research in the following ways:

1. It has indicated that there is a significant relationship between matriculation results and the scores on the Accuplacer CPTs for Arithmetic, Elementary Algebra and Reading Comprehension for both mathematics and non mathematics based degree programmes.
2. It has indicated that there is a significant relationship, albeit small to moderate, between the Accuplacer scores and academic performance for the two groups of degree programmes investigated.
3. It has confirmed that there is a moderate significant relationship between matriculation results and academic performance across all degree programmes.
4. It has indicated that there is a significant moderate relationship between the Accuplacer scores and matriculation performance with academic performance for the two groups of degree programmes investigated.
5. It has confirmed the existence of the cluster groups identified by Foxcroft (1999) for two groups of degree programmes. The cluster groups were described and internally validated in terms of performance on the Accuplacer tests, matriculation results in the form of Composite Matriculation Scores, certain biographical variables and academic performance.

Integration and Discussion of Findings

The correlations between the Accuplacer scores and matriculation performance was generally good for both of the groups of learners investigated, and demonstrated that each contributes unique information about learners entering academic programmes. This supports the view that test results could be used as an adjunct to matriculation results when admission to higher education institutions in South Africa is considered (Foxcroft, 1999, 2001). Provided that the proficiencies to be assessed are carefully considered, test results can provide additional information to that provided by matriculation results, which can in turn enhance admissions, placement, and development decisions. A further aspect also needs to be touched on here. As was discussed in chapter six, advances in test theory necessitate consideration of weighing up the method of test and item development used when selecting measures to include in an admissions and placement battery. This study chose to use and research a measure developed using Item Response Theory (IRT) and that is computer adaptive in nature. The findings of the present study are encouraging for the use of computer adaptive tests in South Africa and suggest that assessment practitioners should not be afraid to research and use measures that are adaptive and based on IRT as opposed to those developed using a more classical approach.

Correlations between Accuplacer tests, matriculation results and academic performance were generally moderate. If this finding is cross-validated in future studies, regression equations could be derived to predict academic performance on the basis of Accuplacer and matriculation performance. The resultant predictions could then be used to guide admissions, placement and development decisions, supporting the regression model approach to multi-stage admissions to tertiary institutions advocated by Foxcroft (1999) and Huysamen (1996).

The fact that higher correlations were found for the mathematics based group than the non-mathematics based group and that more of the variation in first year academic performance could be explained on this basis for the mathematics based group requires further investigation. Other factors that were not investigated in this study (e.g., non-cognitive factors, quality of previous schooling) could have played a role in academic performance, especially for the non-mathematics based group. This needs to be investigated in further studies as it was beyond the scope of the present study to do so. Should additional predictor variables be identified, these need to be built into the regression model used in a multi-stage admissions process.

The three clusters that emerged in both groups of learners investigated are in accordance with those identified by Foxcroft (1999), namely, a low risk group, a mixed medium risk group and a high-risk group. Should these findings be further validated in future studies, classification functions can be derived to predict cluster group membership from test and matriculation performance. This predictive information could then be used to guide admissions, placement and development decisions, as an adjunct to the information obtained from the regression equations mentioned above.

When internally validating the clusters, interesting information came to the fore, which would be fruitful to pursue in future studies. As regards culture, the majority of Black learners in both groups fell into the average and high-risk clusters. On the other hand, the majority of White and Indian learners in both groups fell into the average and low risk clusters. There was no obvious cluster distinction for the Coloureds in the non-mathematics based group, although the majority of Coloureds fell into the average risk cluster in the mathematics based group. As regards the factor of home language, the majority of English learners for both groups fell into the average and low risk clusters. For both groups, learners who indicated a mixed home language of English and Afrikaans fell into the low and average risk groups. The majority of Afrikaans speakers in the non-mathematics based sample fell into either the high or low risk clusters, whereas Afrikaans speakers in the mathematics based sample fell into the low and average risk clusters. Learners in both groups who indicated that their home language was one of the African languages spoken in South Africa fell into the average and high-risk cluster groups. Only the mathematics based sample contained learners who indicated that their home language did not fall into one of the eleven official languages of this country, and they all fell into the low risk cluster group. The way in which the factors of culture and home language impact on test and matriculation performance and how they relate to the prediction of academic performance and the underlying patterns of performance in the data set need further investigation, preferably with larger samples of learners. Should these factors be found to exert a differential impact for different groups, this will have to be taken into account when formulating admissions criteria and developing regression equations and classification functions on the basis of Accuplacer and matriculation performance.

An American study discovered that differential validity exists for different ages on the Accuplacer Reading Comprehension test (Cole, Muenz & Bates, 1998), and the study by Seymour, Cronje, Foxcroft, Koch and Watson (2000) found that Accuplacer scores may predict academic performance better for White than Black learners. Given the findings of the present study in relation to cluster group membership, the potential differential validity of the Accuplacer CPTs needs to be thoroughly investigated for culture and language groups in South Africa. It is, however, possible that what is detected on the Accuplacer tests reflects the residual effects of Apartheid education, which continues to have some influence on tertiary achievement (Dlamini, 1995; Pavlich, Orkin & Richardson, 1995). It is, in fact, preferable that the Accuplacer tests reveal any residuals that may exist, as this would facilitate the developmental focus that forms part of assessment interpretations and long term testing programmes could map whether progress is being made toward real solutions of social and educational issues (Anastasi, 1988).

This research demonstrated that the Accuplacer tests and matriculation results do discriminate to an extent among groups of learners. It has been shown previously that matriculation results do discriminate among learners, as this has been the traditional basis for admission (Badenhorst, Foster & Lea, 1990; Louw, 1992; Nunns & Ortlepp, 1994; Sharwood & Rutherford, 1994). The finding that the Accuplacer tests are able to discriminate among groups of learners supports previous research (e.g., Foxcroft, 1999; Koch, Foxcroft & Watson, 1999). It is possible that their discriminatory power increases as the level of difficulty of the degree programme increases, specifically when such programmes require mathematical knowledge.

Limitations and Suggestions for Further Research

Limitations of the Present Study

Methodologically, the study was limited by the non-probability convenience sampling procedure utilized and the fact that learners at only one tertiary institution were assessed. This was the only viable alternative available as the cost and time that would be involved in probability sampling and testing of all applicants to universities in South Africa would be astronomical, and was beyond the scope of the present study. This study should be replicated, both at the tertiary institution used as well as at other tertiary institutions, with larger samples to see whether similar findings emerge.

Statistically, although the assumptions for the use of parametric tests were not totally met, Pearson product moment correlation coefficients were utilized to investigate relationships rather than the parametric equivalent. This was done because it is legitimate to compute these coefficients if the distribution is unimodal and fairly symmetrical (Runyon & Haber, 1991). In addition, further statistics conducted in this study, including the cluster analyses and methods of investigating their internal validity have no non-parametric equivalents. Thus, consistency was maintained in the use of parametric techniques, but it is acknowledged that the stringent assumptions for the use of those techniques were not guaranteed.

It is also noted that interpretation of low correlations in particular was influenced by two possibilities, namely that the variables might be nonlinearly related and the existence of a restricted range for certain variables (Mendenhall, 1993; Runyon & Haber, 1991).

Also, cluster analysis is essentially intuitive in terms of identifying the numbers of clusters contained in a set of data (Anderberg, 1973). Many of the procedures are poorly defined, the methods are not based on a generally accepted foundation, and different algorithms could yield different results (Edelbrock and Achenbach, 1980; Fletcher & Satz, 1985). However, the procedure and method of cluster analysis was thoroughly studied and clearly stated in this research, and it was only after it was established that the Accuplacer subtests and matriculation results actually correlated significantly with academic performance, that cluster analysis proceeded and an attempt was made to validate the clusters internally. The clusters should now be externally validated on other samples.

The description of the clusters in terms of such factors as age, gender, culture and home language is tentative as cell sizes for many of these factors were very small for some of the variables.

When the clusters were internally validated, hints that the test and matriculation results might differentially predict academic performance were detected. However, the sample used in each of the two groups was too small to permit an investigation into the relationships between the variables for different cultural and language groups. Such studies need to be undertaken in the future.

While it was indicated in chapters two and six that a variety of factors impact on academic performance, only linguistic, numerical and mathematical proficiency and matriculation performance were investigated in the present study. This is partly

because this study forms part of a larger investigation in which a host of variables are being explored and thus the scope of the present study was limited to certain of the Accuplacer CPTs and matriculation performance.

Suggestions for Future Research

As previously mentioned, this study forms part of a larger ongoing research project. There are many different facets of research that could flow from the present findings.

Firstly, the predictive validity of the Accuplacer CPTs needs to be investigated for different language and culture groups, and this could be done utilizing IRT techniques and computer technology. It is important to identify if any bias exists for South African cultural groups so that results can be used fairly and the items can be adapted if necessary.

Secondly, other cognitive, non-cognitive and biographical variables, especially for non-mathematics based programmes, which enhance the prediction of academic performance need to be researched and added to the Accuplacer test results and matriculation performance. Ultimately, being able to explain the maximum amount of variance in academic performance on the basis of predictor variables will lead to more accurate admissions, placement, and development decisions being made. Once the most useful predictor variables have been identified, regression equations should be generated and used as part of the admissions process.

Thirdly, cluster analysis could be conducted on a larger sample size. Although it is unlikely that the cluster groupings would be vastly different from those yielded by this study, it is possible that their internal validation and external verification against academic performance could be more certainly established and that the groupings could be refined. Once the clusters have been refined, classification functions could be derived using discriminant analysis to predict cluster group membership and could be used to aid admissions, placement and development decisions in an empirical way. Also, future cluster analysis conducted on a larger sample size could facilitate the setting of standards for the test battery (Sireci & Robin, 1999), especially for the Accuplacer CPTs.

In conclusion, it appears that the use of test information about entry-level proficiencies of prospective learners at tertiary institutions, in combination with matriculation results can enhance decision-making about the admissions, placement, and development of learners by identifying the individual's strengths and possible

developmental needs. Such developmentally focused assessment is also in line with the new educational policy for higher educational institutions within our transforming society. The usefulness of assessment tools to obtain information about the proficiency levels of learners has been highlighted in this study, as learners are provided with an opportunity to demonstrate their potential on measures that are constructed using the most recent advances in psychometric theory and computer technology, which has the potential to reduce bias and increase the fairness of assessment procedures. Also, by constantly researching the best predictor variables of academic success and combining these into regression equations and deriving cluster groups, more accountable, empirically verified admissions criteria to higher education institutions will be established in the long term.

REFERENCES

AARP. (1996). Alternative Admissions Research Project annual report: 1996. Cape Town, South Africa: University of Cape Town, Author.

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan.

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

Anderberg, M.R. (1973). Cluster analysis for applications. New York: Academic Press.

Angoff, W.H. (1993). Perspectives on differential item functioning. In P.W. Holland and H. Wainer (Eds), Differential item functioning (pp.3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Badenhorst, F.D., Foster, D.H., & Lea, S.J. (1990). Factors affecting academic performance in first-year psychology at the University of Cape Town. South African Journal of Higher Education, 4 (1), 39-45.

Badsha, N., & Yeld, N. (1991). The Alternative Admissions Research Project (AARP) at UCT. Bulletin of Academic Staff - University of Durban-Westville, 12 (2), 32-38.

Bailey, K.D. (1987). Methods of social research (3rd ed.). New York: The Free Press.

Baker, F.B. (1985). The basics of item response theory. Portsmouth, New Hampshire: Heinneman.

Baker, F.B. (1989). Computer technology in test construction and processing. In R.L. Linn (Ed). Educational measurement (3rd ed.) (pp.409-428). New York: Macmillan.

Baker, F.B. (1992). Item response theory: Parameter estimation techniques. New York: Marcel Dekker Inc.

Barnard, J.J. (n.d.). The use of item response theory in test construction (Report No. 2759). Pretoria, GP: Institute for Psychological and Edumetric Research.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. Educational Measurement: Issues and Practice, 13 (2), 12-20.

Beller, M. (1995). Translated versions of Israel's interuniversity Psychometric Entrance Test (PET). In T. Oakland and R.K. Hambleton (Eds), International

perspectives on academic assessment (pp. 207-217). Norwell, M.A.: Kluwer Academic Publishers.

Beller, M., Gafni, N., & Hanani, P. (1999, May). Constructing, adapting, and validating admissions tests in multiple languages. Paper presented at the International Conference on Adapting Tests for Use in Multiple Languages and Cultures, Washington, DC.

Bennet, R.E., Goodman, M., Hessinger, J., Ligget, J., Marshall, G., Kahn, H., & Jack, J. (1997). Using multimedia in large-scale computer-based testing programs (Research Report 97-3). Princeton, New Jersey: Educational Testing Service.

Ben-Shakhar, G., Kiderman, I., & Beller, M. (1996). Comparing the utility of two procedures for admitting students to liberal arts: An application of decision-theoretic models. Educational and Psychological Measurement, 56 (1), 90-107.

Bodibe, R.C. (1995). Tools used to assess academic potential and financial needs of students at tertiary education. SSCSA Report, Vista University, Port Elizabeth.

Boyer, S.P., & Sedlacek, W.E. (1988). Noncognitive predictors of academic success for international students: A longitudinal study. Journal of College Student Development, 29, 218-223.

Brown, F.G. (1983). Principles of educational and psychological testing (3rd ed.). New York: Holt, Rinehart & Winston.

Bugbee, A.C. (1996). The equivalence of paper-and-pencil and computer-based testing. Journal of Research on Computing in Education, 28 (3), 282-299.

Bunderson, C.V., Inouye, D.K., & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 367-405). New York: Macmillan.

Burke, M.J. (1982). A path analytic model of the direct and indirect effects of mathematical aptitude and academic orientation on high school and college performance. Educational and Psychological Measurement, 42, 545-550.

Calitz, A.P. (1997). The development and evaluation of a strategy for the selection of Computer Science students at the University of Port Elizabeth. Unpublished doctoral thesis, University of Port Elizabeth, South Africa.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P.W. Holland and H. Wainer (Eds), Differential item functioning (pp. 397-413). Hillsdale,

New Jersey: Lawrence Erlbaum Associates.

Camilli, G., & Shepard, L.A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage.

Canadian Psychological Association. (1987). Guidelines for educational and psychological testing. Ottawa, Ontario: Author.

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20 (3), 213-229.

Cole, J.C., Muenz, T.A., & Bates, H.G. (1998). Age in correlations between Accuplacer's Reading Comprehension subtest and GPA's. Perceptual and Motor Skills, 86, 1251-1256.

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

Cole, N.S., & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 147-200). New York: Macmillan.

The College Board and Educational Testing Service. (1997). Accuplacer program overview: coordinator's guide. Princeton, New Jersey: Authors.

The College Board and Educational Testing Service. (n.d.) Good practice in college entry-level placement. Princeton, New Jersey: Authors.

The College Entrance Examination Board. (1994). On behalf of educational excellence: College Board assessment programs for secondary school students. New York: College Board Publications.

Cozby, P.C. (1989). Methods in behavioural research (4th ed.). Mountain View, California: Mayfield.

Cronbach, L.J. (1990). Essentials of psychological testing (5th ed.). New York: HarperCollins.

Dahlstrom, G.W. (1993). Tests: small samples, large consequences. American Psychologist, 48 (4), 393-399.

Dana, R.H. (1996). Culturally competent assessment practice in the United States. Journal of Personality Assessment, 66 (3), 472-487.

Dancer, L.S., Anderson, A.J., & Derlin, R.L. (1994). Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. Journal of Consulting and Clinical Psychology, 62 (4), 710-717.

Dane, F.D. (1990). Research methods. Belmont, California: Wadsworth.

Davey, T., & Nering, M. (1998, September). Controlling item exposure and maintaining item security. Paper presented at the CBT Colloquium: Building the

Foundation for Future Assessments, Philadelphia, PA.

Davey, T., & Parshall, C.G. (1995, April). New algorithms for item selection and exposure control with computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Delvare, I. (1996). Addressing tertiary failure rates in South Africa (Report No. 1/96). Pretoria, GP: South African Institute of Race Relations.

Dlamini, C.R.M. (1995). The transformation of South African universities. South African Journal of Higher Education, 9 (1), 39-46.

Dodd, B.G., De Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. Applied Psychological Measurement, 19 (1), 5-22.

Dodd, B.G., & Fitzpatrick, S.J. (1998, September). Alternatives for scoring computer-based tests. Paper presented at the CBT Colloquium: Building the Foundation for Future Assessments, Philadelphia, PA.

Dooley, D. (1995). Social research methods (3rd ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Dorans, N.J., & Potenza, M.T. (1994). Equity assessment for polytomously scored items: A taxonomy of procedures for assessing differential item functioning (Research Report 94-49). Princeton, New Jersey: Educational Testing Service.

Duran, R.P. (1989). Testing of linguistic minorities. In R.L. Linn (Ed). Educational measurement (3rd ed.) (pp. 573-587). New York: Macmillan.

Edelbrock, C., & Achenbach, T. M. (1980). A typology of child behaviour profile patterns: Distribution and correlates for disturbed children aged 6-16. Journal of Abnormal Child Psychology, 8, 441-470.

Education White Paper 3: A programme for the transformation of higher education. (1997). Government Gazette, 382 (17944).

Embretson, S.E. (1996). The new rules of measurement. Psychological Assessment, 8 (4), 341-349.

Embretson, S.E. (1997). Measurement principles for the generation of tests: A quiet revolution. In R.F. Dillon (Ed.), Handbook on testing (pp. 20-38). Westport, Connecticut: Greenwood.

Everitt, B. (1974). Cluster analysis. London: Heinemann Educational Books, Ltd.

Feuer, M.J., & Fulton, K. (1994). Educational testing abroad and lessons for the United States. Educational Measurement: Issues and Practices, 13 (2) 31-39.

Fletcher, J. M., & Satz, P. (1985). Cluster analysis and the search for learning disability subtypes. In B. P. Rourke (Ed.), Neuropsychology of learning disabilities (pp. 40-64). New York: Guilford.

Folk, V.G., & Smith, R.L. (1998, September). Models for delivery of computer-based tests. Paper presented at the CBT Colloquium: Building the Foundation for Future Assessments, Philadelphia, PA.

Fouad, N.A. (1993). Cross-cultural vocational assessment. The Career Development Quarterly, 42, 4-13.

Fox, L.H., & Zirkin, B. (1986). Achievement tests. In G. Goldstein and M. Hersen (Eds), Handbook of psychological assessment (p. 119-131). New York: Pergammon.

Foxcroft, C.D. (1994). The use of tests in screening and selection. Unpublished paper, University of Port Elizabeth.

Foxcroft, C.D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. European Journal of Psychological Assessment, 13, 229-235.

Foxcroft, C.D. (1999). Placement testing at university entrance: preliminary thoughts and findings. Unpublished manuscript, University of Port Elizabeth.

Foxcroft, C.D. (2001). Proposal to revise UPE's undergraduate admissions policy and procedure for its implementation. Unpublished manuscript, University of Port Elizabeth.

Gipps, C. & Murphy, P. (1994). A fair test? Assessment, achievement and equity. Buckingham: Open University Press.

Goldstein, H. (1996). Group differences and bias in assessment. In H. Goldstein and T. Lewis (Eds). Assessment: Problems, developments and statistical issues (pp. 41-93). Chichester, West Sussex: John Wiley & Sons, Ltd.

Green, B.F. (1983). Adaptive testing by computer. In R.B. Ekstrom (Ed.) Measurement, technology, and individuality in education: New directions for testing and measurement (pp. 5-12). San Francisco: Jossey-Bass.

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.

Green, B.F. (1990). System design and operations. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp. 23-40). Hillsdale, NJ: Lawrence

Erlbaum Associates.

Greyling, J.H., & Calitz, A.P. (1997). Selecting CS/IS students in the year 2000. Unpublished manuscript, University of Port Elizabeth.

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1995). Multivariate data analysis with readings (4th ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Hambleton, R.K. (1986). The changing conception of measurement: A commentary. Applied Psychological Measurement, 10 (4), 415-421).

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 147-200). New York: Macmillan.

Hambleton, R.K. (1990). Item response theory: Introduction and bibliography. Psicothema, 2 (1), 97-107.

Hambleton, R.K. (1995). Meeting the measurement challenges of the 1990s and beyond: New assessment models and methods. In T. Oakland and R.K. Hambleton (Eds.), International perspectives on academic assessment (pp. 83-104). Norwell, MA: Kluwer Academic.

Hambleton, R.K. (1996). Advances in assessment models, methods and practices. In D.C. Berliner and R.C. Calfee (Eds.), Handbook of educational psychology (pp. 899-925). New York: Simon and Schuster.

Hambleton, R.K., Clauser, B.E., Mazor, K.M., & Jones, R.W. (1993). Advances in the detection of differentially functioning test items. European Journal of Psychological Assessment, 9 (1), 1-18.

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. Dordrecht: Kluwer-Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hambleton, R.K., Zaal, J.N., & Pieters, J.P.M. (1991). In R.K. Hambleton and J.N. Zaal (Eds). Advances in educational and psychological testing (pp. 341-391). Boston: Kluwer Academic.

Hashway, R.M. (1998). Assessment and evaluation of developmental learning: Qualitative individual assessment and evaluation models. Westport, CT: Praeger.

Helms, J.E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? American Psychologist, 47 (9), 1083-1101.

Helms, J.E. (1997). The triple quandary of race, culture and social class in standardized cognitive ability testing. In D. Flanagan, J.L. Genshaf and P.L. Harrison (Eds). Contemporary intellectual assessment: Theories tests and issues (pp. 517-532). London: Guilford.

Henry, M. (1988). ASAT and the TE score: A critique of 'objective testing'. Australian and New Zealand Journal of Sociology, 24 (2), 289-311.

Holburn, P.T. (1992). Test bias in the intermediate mental alertness mechanical comprehension, blox and high level figure classification tests (Contract Report C/PERS 453). Pretoria, GP: Human Sciences Research Council.

Holland, P.W., & Wainer, H. (1993). Preface. In P.W. Holland and H. Wainer (Eds), Differential item functioning (pp. xiii-xv). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory: Application to psychological measurement. Homewood, Illinois: Dow Jones-Irwin.

Hunter, J.E., Schmidt, F.L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychology tests. In C.R. Reynolds and R.T. Brown (Eds), Perspectives on bias in mental testing (pp. 41-99). New York: Plenum Press.

Hutchinson, T.P. (1991). Controversies in item response theory. Adelaide, South Australia: Rumsby Scientific Publishing.

Huysamen, G.K. (1979). Psychological test theory. Pretoria: J.L. van Schaik.

Huysamen, G.K. (1994). Methodology for the social and behavioural sciences. Halfway House, Gauteng: Southern Book Publishers.

Huysamen, G.K. (1995). The applicability of fair selection models in the South African context. Journal of Industrial Psychology, 21 (3), 1-6.

Huysamen, G.K. (1996). Fair and unbiased admission procedures for South African institutions of higher education. South African Journal of Higher Education, 10 (2), 199-207.

Huysamen, G.K. (1997). Introductory statistics and research design for the behavioural sciences. (Vol. 2) (3rd ed.). Cape Town: G.K. Huysamen.

Jacobs, G.J. (1995). A proposed method for the selection of university students. Bulletin for University Teachers, 19 (3), 36-50.

Jensen, A.R. (1984). Test bias: Concepts and criticisms. In C.R. Reynolds and R.T. Brown (Eds), Perspectives on bias in mental testing (pp. 507-545). New York: Plenum Press.

Jones, L.V. (1994). Perspectives on educational testing: Discussion. Educational Measurement: Issues and Practice, 13 (2), 28-31.

Jordaan, J.J. (1995). Affirmative action: Excellence versus equity. South African Journal of Higher Education, 9 (1), 53-64.

Kingsbury, G.G., & Houser, R.L. (1993). Assessing the utility of item response models. Educational Measurement: Issues and Practice, 12 (1), 21-27.

Kline, P. (1993). The handbook of psychological testing. London: Routledge.

Koch, S.E. (1997). Lecturing between hope and despair: lecturers' perceptions of academic development needs of students and lecturers at the university of Port Elizabeth, a qualitative assessment. COAD Report, University of Port Elizabeth.

Koch, S.E., Foxcroft, C.D., & Watson, A. (1999, May). Challenges posed when assessing English proficiency in a multilingual sample of South African university entrants. Poster session presented at the International Conference on Adapting Tests for Use in Multiple Languages and Cultures, Washington, DC.

Kok, F. (1992). Differential item functioning. In L. Verhoeven, & De Jong, J.H.A.L. (Eds). The construct of language proficiency: Applications of psychological models to language assessment (pp. 115-124). Amsterdam: John Benjamins.

Kolen, M.J., & Brennan, R.L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Kotze, H.N. (1994). Keuringsmodelle vir universiteitstuidierigtings: 'N psigometriese ondersoek. Unpublished doctoral thesis, University of Potchefstroom for Christian Higher Education, South Africa.

Kotze, N., Van der Merwe, D., & Nel, A. (1996). Culture fair selection procedures: The case for psychometrics. Paper presented at the SSCSA Conference, UNISA, Pretoria, GP.

Laatsch, L., & Choca, J. (1994). Cluster-branching methodology for adaptive testing and the development of the adaptive category test. Psychological Assessment, 6 (4), 345-351.

Lautenschlager, G.J., Flaherty, V.L., & Park, D-G. (1994). IRT differential item functioning: An examination of ability scale purifications. Educational and Psychological Measurement, 54 (1), 21-31.

Legg, S.M., & Buhr, D.C. (1992). Computerized adaptive testing with different groups. Educational Measurement: Issues and Practice, 11 (2), 23-27.

Le Roux, G.J. (1995). Extraordinary meeting of the CUP - 18 June 1994. Bulletin.

Linn, R. (1989). Current perspectives and future directions, In R.L. Linn (Ed). Educational measurement (3rd ed.) (pp.1-10). New York: Macmillan.

Locke, L., Silverman, S.J., & Spirduso, W.W. (1998). Reading and understanding research. Thousand Oaks, California: Sage.

Lord, F.M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lourens, P.J.D. (1984). Fairness in selection: An elusive ideal. South African Journal of Psychology, 14 (3), 101-105.

Louw, A.D. (1992). Selection of prospective tertiary students: predicting academic and professional success. Unpublished manuscript, Cape Technikon.

Louw, A. (1994). Keuring van voornemende studente. South African Journal of Higher Education, 8 (4), 156-170.

Luecht, R.M., & Clauser, B.E. (1998, September). Test models for complex computer-based testing. Paper presented at the CBT Colloquium: Building the Foundation for Future Assessments, Philadelphia, PA.

Manning, W.H., & Jackson, R. (1984). College entrance examinations: Objective selection of gatekeeping for the economically privileged. In C.R. Reynolds and R.T. Brown (Eds), Perspectives on bias in mental testing (pp. 189-220). New York: Plenum Press.

Maurelli, V.A., & Weiss, D.J. (1981). Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries. Minneapolis, MN: University of Minnesota, Department of Psychology.

McBride, J.R. (1997a). Research antecedents of applied adaptive testing. In W.A. Sands, B.K. Waters and J.R. McBride (Eds), Computerized adaptive testing: From inquiry to operation (pp. 47-57). Washington D.C.: American Psychological Association.

McBride, J.R. (1997b). Technical Perspective. In W.A. Sands, B.K. Waters and J.R. McBride (Eds), Computerized adaptive testing: From inquiry to operation (pp. 29-44). Washington D.C.: American Psychological Association.

McBride, J.R. (1997c). Dissemination of CAT-ASVB Technology. In W.A.

Sands, B.K. Waters and J.R. McBride (Eds), Computerized adaptive testing: From inquiry to operation (pp. 83-95). Washington D.C.: American Psychological Association.

McBride, J.R., Wetzel, C.D., & Hetter, R.D. (1997). Preliminary psychometric research for CAT-ASVB: Selecting an adaptive testing strategy. In W.A. Sands, B.K. Waters and J.R. McBride (Eds), Computerized adaptive testing: From inquiry to operation (pp. 83-95). Washington D.C.: American Psychological Association.

Melamed, T. (1992). Use of biodata for predicting academic success over thirty years. Psychological Reports, 71, 31-38.

Mendenhall, W. (1993). Beginning statistics: A-Z. Belmont, California: Wadsworth, Inc.

Mensh, E., & Mensh, H. (1991). The IQ mythology: Class, race, gender and inequality. Carbondale: Southern Illinois University Press.

Messick, S. (1994). Alternative modes of assessment, uniform standards of validity (Research Report 94-60). Princeton, New Jersey: Educational Testing Service.

Miller, R. (1992). Double, double, toil and trouble: The problem of student selection. South African Journal of Higher Education, 6 (1), 98-104.

Mills, C.N., & Stocking, M.L. (1995). Practical issues in large-scale high-stakes computerized adaptive testing (Research Report 95-23). Princeton, New Jersey: Educational Testing Service.

Ministry of Education. (2001). National Plan for Higher Education [On-line]. Available: http://education.pwv.gov.za.

Mitchell, T.W. (1994) The utility of biodata. In G.S. Stokes, M.D. Mumford and W.A. Owens (Eds), Biodata handbook: Theory, research an use of biographical information in selection and performance prediction (pp. 485-513). Palo Alto, CA: Consulting Psychologists Press, Inc.

Mitchell, G., & Fridjhon, P. (1987). Matriculation examinations and university performance. South African Journal of Science, 83, 555-560.

Mollendorf, J.W., & Sauer, G.F. (1990). 'n Oorsig en beskouing oor navorsingsbevinginge rakende die voorspelling van moontlike akademiese sukses aan 'n universiteit. Unpublished manuscript, IPEN.

Moreno, K.E., Segall, D.O., & Hetter, R.D. (1997). The use of computerized adaptive testing in the military. In R.F. Dillon (Ed.), Handbook on testing (pp.204-

219). Westport, Connecticut: Greenwood.

Murphy, K.R., & Davidshofer, C.O. (1991). Psychological testing: principles and applications (2nd ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Nel, C.M. (1996). Academic potential as a predictor of study success in a multicultural society. Unpublished manuscript, University of Pretoria.

Nel, C.M. (1997). Universities in transformation: The emerging picture of first year students in a multicultural context. Unpublished manuscript, University of Pretoria, South Africa.

Nelson, R, & Rainier, M. (n.d.). A holistic approach to admissions: inclusion of biographical data in the selection process. Unpublished manuscript, Rhodes University, Grahamstown, South Africa.

Nickels, B.J. (1994). The nature of biodata. In G.S. Stokes, M.D. Mumford and W.A. Owens (Eds), Biodata handbook: Theory, research an use of biographical information in selection and performance prediction (pp. 1-16). Palo Alto, CA: Consulting Psychologists Press, Inc.

Nunns, C., & Ortlepp, K. (1994). Exploring predictors of academic success in Psychology I at Wits University as an important component of fair student selection. South African Journal of Psychology, 24 (4), 201-208.

Osterlind, S.J. (1983). Test item bias. Beverly Hills, CA: Sage.

Owen, K. (1989). Bias in test items: An exploration of item content and item format (Report No. P-106). Pretoria, GP: Human Sciences Research Council.

Owen, K. (1992). Test-item bias: Methods, findings and recommendations (Report No. ED-15). Pretoria, GP: Human Sciences Research Council.

Parshall, C.G. (1998, September). Item development and pretesting in a computer-based testing environment. Paper presented at the CBT Colloquium: Building the Foundation for Future Assessments, Philadelphia, PA.

Pavlich, G.C., Orkin, F.M., & Richardson, R.C. (1995). Educational development in post-apartheid universities: A framework for policy analysts. South African Journal of Higher Education, 9 (1), 65-72.

Peirce, B.N., & Stein, P. (1995). Why the "Monkey's Passage" bombed: Tests, genres and teaching. Harvard Educational Review, 65 (1), 50-65.

Plake, B.S. (1998, September). Alternative for scoring computer-based tests. Paper presented at the CBT Colloquium: Building the Foundation for Future Assessments, Philadelphia, PA.

Plug, C. (1996). An evaluation of psychometric test construction and psychological services supported by the HSRC. Unpublished manuscript, University of South Africa.

Pickering, J.W., Calliotte, J.A., McAuliffe, G.J. (1992). The effect of noncognitive factors on freshmen academic performance and retention. Journal of the Freshmen Year Experience, 4 (2), 7-30.

Portes, P.R. (1996). Ethnicity and culture in educational psychology. In D.C. Berliner and R.C. Calfee (Eds), Handbook of educational psychology (pp. 331-357). New York: Simon and Schuster Macmillan.

Potenza, M.T., & Dorans, N.J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19 (1), 23-37.

Rainier, M. (1995). The role of a biographical approach: A review of recent research and some initial recommendations. Discourse, 6 (1),

Reschly, D. (1986). Aptitude tests. In G. Goldstein and M. Hersen (Eds), Handbook of psychological assessment (p. 132-156). New York: Pergammon.

Reynolds, C.R., & Brown, R.T. (1984). Bias in mental testing: An introduction to the issues. In C.R. Reynolds and R.T. Brown (Eds), Perspectives on bias in mental testing (pp. 1-34). New York: Plenum Press.

Richardson, S.M., & Sullivan, M.M. (1994). Identifying non-cognitive factors that influence success of academically underprepared freshmen. Journal of the Freshmen Year Experience, 6 (2), 89-100.

Runyon, R.P., & Haber, A. (1991). Fundamentals of Behavioural Statistics. (7th ed.). Singapore: McGraw-Hill, Inc.

Rust, J., & Golombok, S. (1989). Modern psychometrics: The science of psychological assessment. London: Routledge.

Rutherford, M., & Watson, P. (1990). Selection of students for science courses. South African Journal of Education, 10 (4), 353-359.

Sands, W.A., & Waters, B.K. (1997). Introduction to ASVAB and CAT. In W.A. Sands, B.K. Waters and J.R. McBride (Eds), Computerized adaptive testing: From inquiry to operation (pp. 3-9). Washington D.C.: American Psychological Association.

Saunders, S.J. (1992). Access to and quality in higher education: A comparative study with some thoughts on the future of higher education in South Africa. Unpublished manuscript, University of Cape Town.

Sawyer, R. (1996). Decision theory models for validating course placement tests. Journal of Educational Measurement, 33 (3), 271-290.

Schaeffer, G.A., Steffen, M., Golub-Smith, M.L., Mills, C.N., & Durso, R. (1995). The introduction and comparability of the computer adaptive GRE general test (Research Report 95-20). Princeton, New Jersey: Educational Testing Service.

Schmitt, N., Hattrup, K., & Landis, R.S. (1993). Item bias indices based on total test score and job performance estimates of ability. Personnel Psychology, 46, 593-611.

Schmitt, N., & Pulakos, E.D. (1998). Biodata and differential prediction: Some reservations. In M.D. Hakel (Ed.), Beyond multiple choice: Evaluating alternatives to traditional testing for selection (pp. 167-182). Mahwah, NJ: Lawrence Erlbaum Associates.

Schoonman, W. (1989). An applied study on computerized adaptive testing. Amsterdam: Swets & Zeitlinger.

Schutte, M. (1994). Veranderlikes wat die leersukses van Vistastudente beinvloed. Unpublished doctoral thesis, Potchefstroom University for Christian Higher Education, South Africa.

Sedlacek, W.E., & Webster, D.W. (1989). Noncognitive indicators of student success. The Journal of College Admissions, 1 (125), 2-9.

Segall, D.O. (1997).The psychometric comparability of computer hardware. In W.A. Sands, B.K. Waters and J.R. McBride (Eds), Computerized adaptive testing: From inquiry to operation (pp. 219-228). Washington D.C.: American Psychological Association.

Seymour, B.B., Cronje, J.H., Foxcroft, C.D., Koch, S.E., & Watson, A.S.R. (2000). The impact of research on the direction of learner-centredness at UPE. Paper presented at the 7th annual South African Association of Institutional Research conference, Port Elizabeth, South Africa.

Sharwood, D., & Rutherford, M. (1994). DET results do predict the students' chance of success: A new model for selection. Paper presented at the SAAAD Conference.

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.

Siegell, S., & Castellan, N.J. (1988). Nonparametric statistics for the behavioural sciences (2nd ed.). New York: McGraw-Hill.

Sijtsma, K. (1998). Methodology Review: Nonparametric IRT approaches to the analysis of dichotomous item scores. Applied Psychological Measurement 22 (1), 3-31.

Sireci, S.G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. Applied Measurement in Education, 12 (3), 301-325.

Skuy, M., Zolezzi, S., Mentis, M., Fridjhon, P., & Cockroft, K. (1996). Selection of advantaged and disadvantaged students for university admission. South African Journal of Higher Education, 10 (1), 110-118.

Smit, J. (n.d.). The value of General Scholastic Aptitude Test scores for university selection purposes at a historically white tertiary institution. Unpublished manuscript, Vista University.

Smith, S., & Beecham, R. (1994). Student selection, university of Durban-Westville, 1994:Context and action. Paper presented at the SAAAD Conference.

Smith, M., & Segall, R. (1994). University access and admission. Paper presented at the SAAAD Conference.

Smittle, P. (1991). Computerized adaptive testing in reading. Journal of Developmental Education, 15 (2), 2-5.

Snyders, A.J.M., (1997). Report from the O-level working group. University of Port Elizabeth, South Africa.

Somer, B., & Somer, R. (1991). A practical guide to behavioural research: tools and techniques (3rd ed.). New York: Oxford University Press.

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? Language Testing, 14 (3), 242-247.

Statsoft. (1984-1995). Statistica for Windows (Volume III): Statistics II (2nd ed.). Tulsa, OK: Author.

Stocking, M.L. (1987). Two simulated feasibility studies in computerised adaptive testing. Applied Psychology: An International Review, 36 (3/4), 263-277.

Stocking, M.L. (1994). An alternative method for scoring adaptive tests. (Research Report 94-48). Princeton, New Jersey: Educational Testing Service.

Stocking, M.L. (1997). Revising item responses in computerized adaptive tests:

A comparison of three models. Applied Psychological Measurement, 21 (2), 129-142.

Stocking, M.L., & Lewis, (1995a). Controlling item exposure conditional on ability in computerized adaptive testing (Research Report 95-24). Princeton, New Jersey: Educational Testing Service.

Stocking, M.L., & Lewis, (1995b). A new method of controlling item exposure in computerized adaptive testing (Research Report 95-25). Princeton, New Jersey: Educational Testing Service.

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17 (3) 277-292.

Stocking, M.L., & Swanson, L. (1996). Optimal design of item pools for computerized adaptive tests (Research Report 96-34). Princeton, New Jersey: Educational Testing Service.

Stokes, G.S. (1994). Introduction and history. In G.S. Stokes, M.D. Mumford and W.A. Owens (Eds), Biodata handbook: Theory, research an use of biographical information in selection and performance prediction (pp. xv-xix). Palo Alto, CA: Consulting Psychologists Press, Inc.

Taylor, T.R. (1987). Test bias: The roles and responsibilities of test user and test publisher (Special Report No. Pers-424). Pretoria, GP: Human Sciences Research Council.

Taylor, T.R. (1994). 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), 184-192.

Thissen, D., & Mislevy, R.J. (1990). Testing algorithms. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp. 103-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland and H. Wainer (Eds), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ting, S.R. (1997). Estimating academic success in the 1st year of college for specially admitted White students: A model combining cognitive and psychosocial factors. Journal of College Student Development, 38 (4), 401-409.

Trabin, T., & Weiss, D.J. (1983). The person-response curve: Fit of individuals to item response theory models. In D.J. Weiss (Ed.), New horizons in testing: Latent

trait test theory and computerized adaptive testing (pp. 83-108). New York: Academic Press.

Tracey, T.J., & Sedlacek, W.E. (1985). The relationship of noncognitive variables to academic success: A longitudinal study. Journal of College Student Personnel, 26, 405-410.

Tracey, T.J., & Sedlacek, W.E. (1987). Prediction of college graduation using noncognitive variables by race. Measurement and Evaluation in Counseling and Development, 19, 177-184.

The TTT Programme. (1993). Access, selection and educational development: A case study. Unpublished manuscript, University of Natal.

University of Stellenbosch. (2001). ADP overview: [On-line]. Available: http://www.sun.ac.za/aop/default.html.

Van der Linden, W.J. (1995). Advances in computer applications. In T. Oakland and R.K. Hambleton (Eds), International perspectives on academic assessment (pp. 105-124). Norwell, MA: Kluwer Academic.

Van der Linden, W.J. (1998, September). Models for computer-based testing: A discussion. Paper presented at the CBT Colloquium: Building the Foundation for Future Assessments, Philadelphia, PA.

Van der Linden, W.J., & Hambleton, R.K. (1997). Item response theory: Brief history, common models, and extensions. In W.J. van der Linden and R.K. Hambleton (Eds.), Handbook of modern item response theory (pp. 1-28). New York: Springer-Verlag.

Van der Vijver, F.J.R., & Poortinga, Y.H. (1991). Testing across cultures. In R.K. Hambleton and J.N. Zaal (Eds). Advances in educational and psychological testing (pp. 277-308). Boston: Kluwer Academic.

Van der Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13 (1), 29-37.

Van der Walt, J.C. (1995). Universities in a new South Africa. Bulletin.

Vane, J.R., & Motta, R.W. (1986). Group intelligence tests. In G. Goldstein and M. Hersen (Eds), Handbook of psychological assessment (p. 100-118). New York: Pergammon.

Van Ede, D.M. (1996). How to adapt a measuring instrument for use with various culture groups: A practical step-by-step introduction. South African Journal of

Higher Education, 10 (2), 153-160.

Venter, J.A. (1993). Towards a just and valid evaluation of the learning quality and potential of prospective candidates for tertiary education. Paper presented at the SAAAD Conference: Finding our voices.

Vicino, F.L., & Moreno, K.E. (1997). Human factors in the CAT system: a pilot study. In W.A. Sands, B.K. Waters and J.R. McBride (Eds), Computerized adaptive testing: From inquiry to operation (pp. 157-160). Washington D.C.: American Psychological Association.

Vispoel, W.P. (1993). Computerized adaptive and fixed-item versions of the ITED vocabulary subtest. Educational and Psychological Measurement, 53 (3), 779-788.

Wainer, H. (1990). Introduction and history. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp. 1-21). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Dorans, N.J., Green, B.F., Mislevy, R.J., Steinberg, S., & Thissen, D. (1990). Future Challenges. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp. 233-271). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Wang, T., & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. Journal of Educational Measurement, 35 (2), 109-135.

Watson, A. (1997). Survey on the admission policy and alternative admission procedures at South African universities. University of Port Elizabeth, South Africa.

Watson, A., Foxcroft, C.D., & Koch, S.E. (1999, May). The development of a non-cognitive questionnaire for a multi-cultural, multi-lingual South African student population: preliminary findings. Poster session presented at the International Conference on Adapting Tests for Use in Multiple Languages and Cultures, Washington, DC.

Watson, A.S.R., Van Lingen, J.M., & De Jager, A.C. (1997). Increasing access for black students through a special admission procedure: the UPE experience (1993-1996). Unit for Student Counselling Report, University of Port Elizabeth, South Africa.

Way, W.D., Steffen, M., & Anderson, G.S. (1998, September). Developing, maintaining, and renewing the item inventory to support computer-based testing. Paper presented at the CBT Colloquium: Building the Foundation for Future Assessments, Philadelphia, PA.

Wedman, I. (1994). The Swedish Scholastic Aptitude Test: Development, use and research. Educational Measurement: Issues and Practice, 13 (2), 5-11.

Weiss, D.J. (1983). Introduction. In D.J. Weiss (Ed.). New horizons in testing: Latent trait test theory and computerized adaptive testing (pp. 1-8). New York: Academic Press.

Weiss, D.J. (1985a). Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 53 (6), 774-789.

Weiss, D.J. (1985b). Final report: Computerized adaptive measurement of achievement and ability. Minneapolis, MN: University of Minnesota, Department of Psychology.

Weiss, D.J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. Lubiski and R.V. Dawis (Eds), Assessing individual differences in human behaviour: new concepts, methods and findings (pp. 49-79). Palo Alto, CA: Davies-Black.

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 and J.N. Zaal (Eds.). Advances in educational and psychological testing (pp. 69-95). Boston: Kluwer Academic.

Whitney, D.R. (1989). Educational admissions and placement. In R.L. Linn (Ed). Educational measurement (3rd ed.) (pp. 515-525). New York: Macmillan.

Wilkinson, W. K., & McNeil, K. (1996). Research for the helping professions. Pacific Grove, California: Brooks/Cole Publishing Company.

Williams, L. (1999). Student prospectus - alternative access programmes. [On-line]. Available: http:\\www\unp.ac.za\MP\prospectus\altac.htm.

Willingham, W.W., & Cole, N.S. (1997a). Research on gender differences. In W.W. Willingham and N.S. Cole (Eds). Gender and fair assessment (pp. 17-54). Mahwah, N.J.: Lawrence Erlbaum Associates.

Willingham, W.W., & Cole, N.S. (1997b).Summary and implications. In W.W. Willingham and N.S. Cole (Eds). Gender and fair assessment (pp. 347-365).

Mahwah, N.J.: Lawrence Erlbaum Associates.

Willingham, W.W., & Cole, N.S. (1997c). Fairness issues in test design and use. In W.W. Willingham and N.S. Cole (Eds). Gender and fair assessment (pp. 227-345). Mahwah, N.J.: Lawrence Erlbaum Associates.

Yeld, N. (1992, April). Admissions. Paper presented at Conference on Tertiary Education in a Changing South Africa, Port Elizabeth, South Africa.

Yeld, N., Haeck, W., Shall, A., & Hiscock, M. (1994). Alternative Admissions Research Project annual report: 1993. University of Cape Town, South Africa.

APPENDIX A
GUIDELINES FOR ASSESSING COMPUTERIZED ADAPTIVE TESTING

Content Considerations

1. Content specifications for items should be identical for paper-and-pencil and computerized adaptive tests.
2. Item content of items included in the item pool should match content specifications.
3. Test items should be designed be compatible with the available computer equipment.

Dimensionality

4. Goodness of fit of the relevant IRT model must be checked.
5. Items that are highly discriminating should be included.
6. Factor analysis of the inter-item tetrachoric correlations should be undertaken.
7. The assumption of local independence should be investigated.
8. Tests that are not unidimensional should be divided into subtests
9. Tests should be balanced to reflect the heterogeneity of domain content and item formats.

Reliability

10. The SEM of every test score should be reported as a function of the test score, in the metric of the reported score.
11. In addition, the SEM of every test should be reported in the ability metric.

Validity

12. The equivalence of variance-covariance matrices for paper-and-pencil and CAT tests should be evaluated.
13. Covariance-structures of the two versions should be compared.
14. The two versions should be validated against one external criterion.
15. The degree of prediction bias should be assessed for relevant subgroups.

Estimation of Item Parameters

16. The sample for item calibration should be adequate in size, (i.e., it should contain at least 1000 cases).
17. The calibration sample should be selected so that enough examinees are included from the range of ability required to estimate the lower asymptote and the point of inflection of the ICC.

18. The method of item parameter estimation should be demonstrated to be empirically consistent (i.e., large samples should result in good estimates).
19. The method of item parameter estimation should be demonstrated to be unbiased, otherwise the degree and nature of existing bias should be specified.
20. ICCs should fit the observed data.
21. Item difficulties in the paper-and-pencil and computerized adaptive tests should be compared.

Linking of Item Parameters

22. The linking procedure for placing items on the same scale should be comprehensively described.
23. Equivalence of groups should be demonstrated when linking procedures utilized equivalent groups.

Item Pool Characteristics

24. Distributions of the item parameter estimates and descriptive statistics for these should be presented.
25. Information for the entire item pool should be reported.

Item Selection and Test Scoring

26. The method of selecting items and estimating ability must be reported explicitly and in detail.
27. The procedure should include a way of varying items selected to avoid the exclusive use of only a few items.
28. The computer algorithm should be able to administer designated items and record the responses separately, without hindering the adaptive process.
29. The computer must be able to base the choice of the initial item on prior information.

Human Factors

30. The testing environment should be quiet, comfortable and distraction-free.
31. The monitor should be free of glare.
32. Legibility of the items should be empirically evaluated.
33. The monitor should be capable of displaying graphics that have fine detail.

Please note that this list was adapted from tables in Hambleton, 1989 and Hambleton, Zaal and Pieters, 1991, which summarized the guidelines of Green, Bock, Humphreys, Linn and Reckase, 1982.