DEVELOPMENT OF THE LEARNING POTENTIAL COMPUTERISED ADAPTIVE TEST (LPCAT)

MARIÉ DE BEER
Department of Industrial Psychology
UNISA

ABSTRACT

An overview of the development of a dynamic test for the measurement of learning potential - the Learning Potential Computerised Adaptive Test (LPCAT) - is provided. It was developed in South Africa with the aim of providing information on the present as well as potential future level of general nonverbal figural reasoning ability for persons from different backgrounds in a way that is fair to all concerned. Multicultural samples were used in its development and standardisation. Item Response Theory (IRT) principles and computerised adaptive testing technology address many of the earlier measurement problems concerning dynamic assessment of learning potential and made possible the construction of a psychometrically sound yet time efficient and practically useful tool for the measurement of learning potential in multicultural contexts.

The unique and complex South African context has provided new challenges and opportunities in the field of psychological assessment and specifically also for the development of psychological tests. In a country with 11 official languages, many social and educational problems and great disparity in socio-economic and educational background of individuals, psychological assessment or the development of psychological measurement tools will never be a simple matter (Claassen, 1997; De Beer, 2000a; Foxcroft, 1997; Huysamen, 2002; Owen, 1998). Much criticism has been leveled against psychometric tests because many of them do not allow for diversity among candidates. Appeals have been made to address the need for the development and adaptation of culturally appropriate measures (HPCSA, 1998; PSYSSA, 1998a; PSYSSA, 1998b). Recent legislation, such as the Employment Equity Act (1998) and the Skills Development Act (1998) also influenced psychological assessment practices. The present situation provides unique opportunities to try to find psychometrically justifiable yet practically useful solutions to the challenges.

Background

Worldwide, the measurement of cognitive ability has featured prominently in the establishment of psychology as a science and in particular in the development of psychological measuring instruments. Although Cattell, is credited for inventing the term *mental test*, the French psychologist Binet and his colleague Simon in 1905 introduced the first test of intelligence using higher mental processes and are considered the forerunners in the development of intelligence tests as we know them today (Gould, 1981; Wolf, 1973). Binet's first attempts to measure intelligence psychometrically were

aimed at using tasks that required cognitive reasoning to identify pupils in need of special education (Binet & Simon, 1905/1916). The aim was to distinguish, in a group of retarded school children, those who seemed likely to benefit from further instruction or training from those who would probably not. Binet introduced the concept of *mental level,* which has formed the basis of most intelligence tests developed since (Binet & Simon, 1915; Gregory, 2000; Wolf, 1973). Binet's interest was in identifying present ability with a view to providing developmental opportunities to improve functioning at whatever level the individual performed. Similar principles can be seen in the concept of dynamic assessment and measurement of learning potential today, which challenges the static and unchangeable view of intelligence.

In 1911, the German psychologist Stern proposed the popular concept of *mental age*, which led to the relation between mental age and chronological age, expressed as a single number, the intelligence quotient (IQ) (Thorndike & Lohman, 1990). Unfortunately, the fact that the intelligence quotient remains essentially constant over the years of the child's development tended to be incorrectly interpreted as meaning that the IQ measured a relatively innate and seemingly constant general ability. From personal interviews with Theodore Simon, Wolf (1973, p. 203) reported that Simon "continued to think of the use of IQ as a betrayal of the scale's objective". According to Thorndike and Lohman (1990, p. 35), the fact that the interpretation of IQ scores are the same, regardless of the child's age

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

Changes in IQ scores over time

Binet was the first person to focus attention on the possibility of changes in intelligence test scores over time. One of the reasons for the more recent development of learning potential as a theoretical concept and continued efforts to measure it, is that research results indicate that IQ scores are subject to change (Grigorenko & Sternberg, 1998). Changes in IQ test scores are usually linked to educational opportunity, language proficiency and general socioeconomic level, with differential changes in test scores between cultural groups (Claassen, 1997; Vincent, 1991). Where certain culture- or other subgroups are disadvantaged, an improvement in the socioeconomic and educational opportunities of the disadvantaged group results in increases in the mean group score which are beyond the normal population increases over time – leading to smaller differences between the mean scores of subgroups. Further evidence is that generally, larger differences between black and white groups are found in adults, with differences between younger people – for whom smaller differences in socioeconomic and educational opportunities are usually the trend - becoming smaller (Van de Vijver, 1997; Vincent, 1991).

In South Africa, Verster and Prinsloo (1988) reviewed the changes found between different cultural and language groups over time. Earlier distinct differences in socioeconomic and educational background between English-speaking and Afrikaans-speaking groups were, for instance reflected in differences in mean group cognitive test results. On comparing the results of different generations, larger differences were found between the (English-speaking and Afrikaans-speaking) adults than for the younger group. These results indicate that an improvement in socioeconomic and educational circumstances can affect performance in one generation, and is similar to what has been reported internationally (Plomin, 1997; Van de Vijver, 1997; Vincent, 1991).

Practical dilemmas of the present South African context

It is clear that improvement in socioeconomic and educational opportunities are reflected in changes in the mean cognitive ability scores of groups. At present there are still large differences between the cultural groups in South Africa in terms of socioeconomic and educational opportunities as well as general living conditions, with the African group in particular being the poorest off in all respects (Central Statistical Services of South Africa (CSS), 1996). Only in the last eight to ten years has South Africa embarked on the road to improving socioeconomic conditions and educational opportunities of the disadvantaged groups. These changes are certain to impact on future test results and differences in means scores between groups should decrease as conditions for the disadvantaged groups improve.

In South Africa the history of psychometric tests has largely followed international trends but also reflects the sociopolitical history of the country (Claassen, 1997). Given the recent changes in South Africa and the increasing integration in schools, universities, the workplace and society in general, there is an urgent need for culture-fair instruments that can be used for different cultural and language groups. However, the large differences in socioeconomic and educational background with which people presently still come to the assessment situation should be taken into account in the development and use of cognitive ability tests in particular. The focus needs to be on the identification of areas for development and to improve the status of the disadvantaged (Loehlin, 1992), reflecting Binet's (1905) intention that test scores should be used to identify those who can benefit from help and to use the results to plan further training and development. A focus on development by means of the measurement of potential can address both legislative and practical requirements in a country such as South Africa with its diversity of people, allowing for continuous changes of a developing society. According to Pyryt (1996), the most productive way to view group differences is as indicators of equity in society, and rather than blaming the tests for revealing differences, they should be used to promote greater awareness of the need to invest in people through social and educational programs.

While the need for the construction of culture-fair and unbiased instruments can easily be understood, the operationalisation and practical implementation of such an endeavour is hampered by many problems. Although standard intelligence tests have for many years

been widely used in the selection and placement of people and for prediction of academic and work performance, they are not always suitable for cross-cultural testing, and language proficiency as well as socioeconomic and educational disadvantage have been shown to affect results (Van Eeden, 1993; Claassen, 1997). The fact that South Africa has 11 official languages has to be taken into account, leading to a preference for instruments with nonverbal, figural content for testing cognitive ability cross-culturally. Nonverbal tests can make a useful contribution as culture-fair measures of general reasoning ability, although they should not be used in isolation since they cannot reflect all characteristics of importance. For culture-fair assessment in the South African context, the use of nonverbal figural reasoning content with items such as figure analogies and pattern completion has been found best for cross-cultural use (Hugo & Claassen, 1991). People from disadvantaged backgrounds often have not had the opportunities to develop their cognitive potential fully and as a result many of these individuals score poorly when assessed with standard psychometric instruments. However, these poor scores often reflect a lack of educational opportunities and not necessarily a lack of potential.

The complex South African context necessitates the use of procedures and tests that take the diversity of examinees into account. There should be a change in emphasis from measuring crystalised competencies that are largely the result of educational opportunities, towards the measurement of fluid ability – in particular undeveloped potential - which will allow for redressing of past imbalances. A focus on the measurement of learning potential, which allows for changes in measured ability following a learning opportunity within the assessment and which takes into account potential future achievement over and above current levels of achievement, will go a long way in addressing practical assessment problems in the cognitive domain.

Recent legislation (Employment Equity Act, 1998; Skills Development Act, 1998) also emphasises ongoing training and development, with the aim of redressing the imbalances of the past. The measurement of learning potential can also help to identify appropriate levels of training to be provided over a broad spectrum of ability, without necessarily relying on language proficiency or prior formal training, thus providing useful information for the purposes of selection or training and development. Murphy (2002) provides an extensive overview of South African research in the field of dynamic assessment, ie the measurement of learning potential.

Learning potential as alternative to standard cognitive assessment

Standard tests of cognitive ability measure mostly the products of prior learning and hence rely heavily on the assumption that all examinees have had comparable opportunities to acquire the skills and abilities being measured. This assumption is not true when individuals from different socioeconomic and cultural backgrounds are compared. Whereas ability refers to that which is available on demand, potential is concerned with what could be, and is based upon the possibility of change (Taylor, 1992; 1994a, 1994b; Von Hirschfeld, 1992; Zaaiman, Van der Flier & Thijs, 2001). Learning potential refers to an overall cognitive capacity and includes both present and improved

future performance. Implied in the use of the term is the assumption that intelligence – that which is measured with psychometric tests – is changeable, as indicated by changes in scores obtained with standard tests. By looking beyond current performance and acknowledging the possible influence of other factors on performance, more realistic measures and descriptions of cognitive development and reasoning ability can be obtained.

Binet's contribution – a historical perspective on the learning potential approach
Binet saw the measures obtained from his test as at best tentative, because further development and learning could lead to different diagnoses in future (Binet, 1905/1916). Inherent in this view is the changeability of intelligence – that in certain cases it could be further developed than would appear from initial measures – analogous to today's conception of learning potential.

Vygotsky's zone of proximal development – the theoretical base for learning potential assessment
In recent years, Vygotsky's (1978) theory of the Zone of Proximal Development (ZPD) has further contributed to a different approach to the measurement of cognitive functioning and is generally acknowledged as the theoretical base upon which dynamic assessment and the measurement of learning potential has been built.

Vygotsky (1978, p. 85) used a special case - "a simple example" of two children of the same age (10 year old chronologically) who initially measure at the same level of mental development (both eight years old in mental development level). These two children can therefore be considered exactly the same in terms of age and mental development levels. Based on this information alone, Vygotsky indicates that one would expect the future performance or "subsequent course of mental development and of school learning" of these two children to be the same (Vygotsky, 1978, p. 86). However, he proposes that useful *additional* information can be obtained if one does not stop there. He argues that if these two children are shown additional ways of dealing with problems, differences between them may become apparent. Suppose that, following the additional training, it hypothetically "turns out that the first child can deal with problems up to a twelve-year-old's level, the second up to a nine-year-old's. Now, are these children mentally the same?" (Vygotsky, 1978, p. 86). Of course the answer is no. This clear and extremely simple example unequivocally explains Vygotsky's theoretical concept of the ZPD as well as its practical implications. However, this simple example has often been taken as the general case, which leads to logical inconsistencies (see discussion of a new definition of learning potential further on).

Because most of the initial research on dynamic assessment and the measurement of learning potential involved low-ability disadvantaged or educable mentally retarded examinees – and these had generally performed quite poorly on initial unaided tests, the level of actual (present) ability was comparably low (Budoff, 1969; Brown & French, 1979; Campione, Brown & Ferrara, 1982; Carlson, 1989). With the initial scores of the samples all being low and approximately equal, the object of focus became the difference score (ZPD) as indicator of learning potential. In their interpretation of learning potential,

many authors have focused only on the individual's potential to *further* benefit from instruction (ie the ZPD) as the principal variable. Current intellectual ability has therefore assumed secondary importance. The special cases where individuals' current (present / actual) state are the same has thus been taken to present the general case. Vygotsky however clearly indicated that the zpd is to be used as a tool by means of which "we can take account *not only* of the cycles and maturation processes that have *already* been completed *but also* those processes that are currently in a state of formation, that are just beginning to mature and develop …. allowing not only for what already has been achieved developmentally but also for what is in the course of maturing " (Vygotsky, 1978, p. 87 [own italics]). Both of these should be the operational measures of interest in the assessment of learning potential.

Different approaches to the measurement of learning potential

Researchers have employed different approaches, procedures, techniques and measures in their use of dynamic assessment for the measurement of learning potential. The common link between all of these is that they involve some form of help or assistance to the person being assessed with a view to providing a more accurate assessment of individual differences than can be obtained with standard test scores. The pretest performance indicates the present (actual) level of development, while the post-test performance, which follows after relevant training, reflects the future (potential) level of development (Vygotsky, 1978). Broadly speaking, three approaches or categories of approaches can be identified based on both theoretical and practical concerns:

Feuerstein's enrichment approach

Feuerstein (1979) is generally regarded as the father of the cognitive enrichment approach to dynamic assessment. The focus of Feuerstein's approach is on remediation and the modifiability of cognitive functioning with the aim of developing those functions that are in the process of maturing to change the level of functioning of the individual concerned. This research has primarily focused on low-performing individuals using highly individualised clinical approaches and individuals who already function at high levels are not viewed as legitimate targets for this kind of assessment (Feuerstein, Feuerstein & Gross, 1997; Feuerstein, Rand, Jensen, Kaniel & Tzuriel, 1987). Many decisions depend upon subjective judgment of the practitioner and the actions and responses of the examinee determine the actions of the examiner. The role of the examiner is crucial, and because the training provided is individual and not standardised, comparison of individual results is problematic. The cognitive enrichment form of dynamic assessment, requires much skill, training, experience and investment in time and effort to administer (Tzuriel, 1997) and is consequently extremely expensive. Some studies reported disappointing results in improvement skills and abilities measured (Blagg, 1991; Frisby & Braden, 1992; Van Niekerk, 1991) considering the amount of time and human effort involved.

Dynamic measurement approaches using standard tests

A second category of dynamic testing approaches which utilizes standard tests has also been developed, and includes the test-centred coaching approach, graduated prompting methods, the psychometrically oriented learning test approach, and the testing-the-limits approach. These approaches share an emphasis on measurement with the aim of combining the assessment of learning potential or the measurement of the ZPD with sound psychometric principles (Budoff & Harrison, 1971; Guthke, 1992, 1993a, 1993b). The aim of the more quantitative psychometric-oriented approaches is to obtain objective, valid, reliable and quantifiable measures of learning potential. With standardisation as focus, these approaches contribute to improved psychometric properties of the assessment of learning potential. The various procedures based on the psychometric approach differ in the degree to which the tasks used are domain-specific, the degree of standardisation in the interventions and the level of prescriptive or diagnostic information obtained (Kozulin & Falik, 1995). In this approach the testing procedure is standardised to produce psychometrically defensible quantitative data, often with a focus on how much aid is needed to bring about a specified level of performance, rather than how much improvement can be made. Tasks of inductive reasoning and variants of progressive matrices problems are mostly used because performance in such tasks is known to be related to scholastic success. These researchers generally found that dynamic measures tended to be superior to static measures in their ability to predict how much children would profit from instruction (Boeyens 1989a, 1989b; Budoff, 1987a, 1987b; Campione , Brown, Ferrara & Bryant, 1984; Guthke, 1992; Shochet, 1994; Zaaiman et al., 2001). The importance of these approaches lie in the attempts to standardise procedures so that better psychometric measures can be obtained and the results of different examinees can be compared.

IRT-based psychometric approach

The aim of dynamic assessment is to modify and examinee's performance level by providing instructions as part of the assessment. While the use of classical test theory leads to measurement problems regarding difference scores, the use of IRT latent trait models provides a means of accurately comparing the pretest and post-test scores. In computerized adaptive testing (CAT), examinee ability level is estimated during test administration and items appropriate to the examinee's ability level are selected from an item bank to interactively construct an optimal test for each examinee (Meijer & Nering, 1999). Embretson (1987) and Sijtsma (1993a, 1993b) propose that IRT-based procedures and, in particular CAT, provide a solution to many of the psychometric problems that have been associated with dynamic assessment and the measurement of learning potential. The psychometric features of dynamic assessment instruments can be vastly improved if IRT and CAT procedures are used. Learning potential assessment needs a sound psychometric foundation, and IRT and CAT can solve several measurement problems associated with this field (Embretson, 1987, 1991, 1992; Sijtsma, 1993a, 1993b).

Problems with learning potential assessment

According to Grigorenko and Sternberg (1998) there is a paucity of published empirical research on the reliability and validity of dynamic assessment and according to them, dynamic assessment has not yet lived up to its promise. Some of the main practical and technical problems with dynamic assessment are:

- the time and difficulty involved in administering the tests
- the high cost because of the level of training required of the examiner
- subjective scoring of some procedures
- problems with the accuracy of the measurement of difference scores (ZPD)
- the lack of standardisation which limits generalisation and comparison
- the practice effect when the same instrument is used in both the pretest and the post-test
- problems in finding suitable criterion measures to provide predictive validity evidence for learning potential measures.

A Proposed solution

Use of modern psychometric approaches such as IRT and CAT can address a number of the problems mentioned. In the development of the LPCAT, the following factors that can contribute to fairer and more equitable assessment were taken into consideration:

A definition of learning potential for all ability levels

While Vygotsky included both the initial level of functioning and the ZPD in explaining his theory, the difference score or ZPD has often (incorrectly) been referred to as that which indicates 'potential'. Because much of the early research in dynamic assessment involved low-ability examinees with similar (low) initial levels of performance, the focus was *only* on the ZPD or difference score obtained. This makes allowance *only* for the special example used by Vygotsky (see earlier discussion), where initial levels of performance are equal and can therefore be ignored during interpretation. Vygotsky's special case can be used when the pretest (present) level of performance of a group is very similar, but it does not allow for the interpretation and comparison of scores of individuals where there are differences in the initial level of the performance and quite likely also differences in the ZPD. The use of the ZPD (difference) scores without reference to the level at which they occur provides incomplete information. Dague (1972, p. 71) noted that "learning ability is not independent of education" and indicated that educability is partly a function of previous schooling, while Jensen (1963, p. 1) drew the following conclusion:

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

learning tasks and psychometric measures of intelligence usually average close to zero.

Vygotsky was only illustrating the principles of his theory by using a special example and did not elaborate on the problematic interpretation of the majority of cases where both the initial level of performance and the ZPD are likely to differ. He did indicate that both the present (actual) and potential levels of performance should be considered. The problem with taking Vygotsky's special case as a general example and referring to the ZPD as defining "learning potential" can be illustrated by the following practical example:

If a person is already performing at a high level, performance after training may remain at approximately the same level without a dramatic improvement. The fact that the difference score or improvement is small (or even zero), does not mean that the individual has no or little learning potential because the overall level of performance is and remains high.

It is clear that both initial level of performance and improvement should be taken into account to provide a fair and equitable description of likely future performance. If it is taken as starting point that learning potential results are supposed to assess the capacity of a person to make progress in a learning or academic environment, such tests should predict "the ability to learn". This would imply that both the present level of performance and the ZPD are needed to improve prediction of performance in new learning situations. For real-life decisions to be made, the information presented seldomly reflects the convenient characteristics of special cases. Vygotsky's (1978) proposed use of both the actual developmental level (level of initial performance) *and* the ZPD is essential to achieve logical and practically useful interpretations. Using both the actual developmental level *and* the ZPD (difference score) as suggested by Vygotsky, allows for the interpretation of more general cases and generalising his theory to all ability levels.

Learning potential for the LPCAT is defined as a combination of the pretest performance and the magnitude of the difference between the post-test and the pretest scores. Since the LPCAT measures learning potential over a broad range of ability levels, it is important that the improvement score should not be used alone, but that present level of performance should also be taken into account.

Use of nonverbal figural item content

Use of non-verbal material to measure fluid ability is recommended for fair assessment of general reasoning ability in multicultural contexts and most cross-cultural tests make use of nonverbal content in order to obtain a more culture-fair measure of intellectual / reasoning abilities. This to some extent addresses concerns about language proficiency, socioeconomic factors and educational opportunities influencing test results. Claassen (1997) proposes that a realistic objective in cross-cultural testing would be to construct tests that presuppose only experiences that are common to the different cultures. This would preclude any verbal materials, as well as any material that relates directly to scholastic content both of which are typically found in standard cognitive tests. There is

some evidence that nonverbal content involving pictures of cultural artifacts such as vehicles, furniture, musical instruments or household appliances do involve cultural loading while items that are considered to be more culture-reduced include geometrical figures involving lines, circles, triangles and rectangles (Jensen, 1980). The item types chosen for the LPCAT were figure series, figure analogies, pattern completion. These item types are typical of the figural items found in most cognitive ability tests and are generally considered to provide a fairly pure measure of Spearman's g-factor (Jensen, 1981):

> *Culture-reduced tests try to minimize culture loading by not using words, letters, numbers, or even pictures of familiar common objects. They consist of only simple elements – lines, curves, circles and squares – and they involve such universal concepts as up/down, right/left, open/closed, whole/half, larger/smaller, many/few, full/empty, and the like. Quite complex problems involving relational reasoning can be made up of such elements – for example figural analogies, figure series completion, and matrices. Such tests are near the opposite extreme on the culture-loading continuum as compared with tests involving specific factual knowledge or scholastic content* (Jensen, 1981, p. 133)

Use of dynamic (test-teach-retest) assessment of learning potential

The dynamic test-teach-retest approach with the focus on measurement of learning potential has as its aim the provision of learning opportunities within the assessment situation in order to improve the opportunity for examinees to optimise their test performance (Campione & Brown, 1987; Hamers & Resing, 1993; Lidz, 1991). This approach acknowledges the differences with which examinees come to the testing situation. The pretest provides an indication of the present (actual) level of performance attained – similar to that which is typically assessed in standard tests. The training is aimed at providing further examples, hints and guidelines that will highlight important aspects of information required to help solve similar questions. The post-test then provides an indication of the potential future level of performance – that which the examinee is likely to attain if further training can be provided. The assumption is that examinees are likely to utilise real-life learning opportunities in a similar way.

Use of item response theory (IRT) to overcome measurement problems and to investigate differential item functioning (DIF)

The development of IRT over the last 30 to 40 years has brought significant changes in psychometric theory and test development (Embretson, 1996; Embretson & Reise 2000). IRT in its most basic form postulates that a single latent trait underlies examinee performance on a test and that the relation between this trait and the probability of a correct response on an item is a monotonically increasing curve (Hambleton & Slater, 1997). One of the requirements of IRT is unidimensionality of the construct or latent trait being measured. IRT models specify a function depicting the relation between the probability of correct responses of an individual to a test item and the individual's level on the latent trait. In IRT the item parameters are not dependent upon the ability level of the examinees responding to the item – reflecting that the item parameters are a property

of the item and not of the group that responded to the item. Another basic feature of IRT is that the examinee's ability is invariant with respect to the items used to determine it. Different sets of items will yield values of estimated ability near the examinee's actual ability level. Furthermore the difficulty level of items and the ability level of examinees are on the same scale, which allows for computerised adaptive testing – see next section (Weiss, 1983a, 1983b). In dynamic testing, the comparison of pretest and post-test scores forms a crucial part of the interpretation of results. Comparison of test scores in the classical test theory model is problematic, with measurement of difference scores identified as one of the key problems experienced in dynamic testing when viewed from the classical test theory perspective. The main advantage of IRT for learning potential measurement lies in the improved accuracy of measurement of difference scores, as well as improved means to compare scores of the same or of different examinees, since in IRT, measures are on the same scale. IRT overcomes measurement problems typically encountered when obtaining pre- and post-test measures and allows a modern-day solution to ensure both fair and accurate measurement of learning potential.

A further very useful application of IRT is differential item functioning (DIF) analysis to investigate bias (Osterlind, 1983; Wainer, 1993). Separate item characteristic curves can be drawn for different subgroups, thereby allowing for visual representation of item characteristics per subgroup, allowing for comparison of subgroups and investigation of item bias. In South Africa in particular, special attention needs to be given to bias analysis – an aspect which is also recommended by the Psychometrics Committee of the Health Professions Council for the development and evaluation of new tests (HPCSA, 1998).

Use of CAT to ensure time efficiency without forfeiting measurement accuracy

Computerised adaptive testing (CAT) is one of the most exciting developments that has flowed from IRT. It is based on the premise that "an examinee is measured most effectively when the test items are neither too difficult nor too easy for him" (Lord, 1980, p. 150). CAT involves the interactive selection of items during test administration so that item difficulty is matched to the examinee's (estimated) ability level throughout the test session. The item selected each time is the one that provides the best information at the examinee's current estimated level of ability. A test thus "adapted" to each individual examinee's ability level, results in various advantages such as more precise measurement and higher examinee motivation (Hambleton & Swaminathan, 1985; Lord, 1980; Weiss, 1983a, 1983b). An interesting link back to the past is that Binet's first test can be regarded as an individually administered adaptive test. However, this mode of testing was not fully explored until the 1960's when theoretical developments and the availability of computer technology allowed the development of IRT-based CAT. The interactive selection of appropriate items from an item bank throughout the test is possible because the difficulty level of items and the examinees' estimated ability level are on the same scale. Practical requirements for adaptive testing according to Green, Bock, Humphreys, Linn and Reckase (1984), Reckase (1989) and Weiss and Vale (1987) are:

- an adequate pool of items with well-established item parameters (based on a large and representative sample)

- an item selection procedure with rules for selecting the next most optimal item to be administered
- a scoring procedure to produce ability estimates on the same scale after each item has been administered
- stopping rules – that is, a specified level of information, a specified posterior variance for Bayesian procedures or a certain number of items administered.

In general administration of CATs, because the latent trait (theta) value is unknown at the outset, a first item of average difficulty is usually administered to the examinee. If the correct answer is given, the theta value is re-estimated (adjusted upward) and the next item will be more difficult to match the examinee's current estimated ability level. If an incorrect answer is given, the theta-value is re-estimated (adjusted downward) and the next item will be less difficult, once again to match the examinee's current estimated ability level. Based on each response and the preceding responses, the computer program estimates the theta value and its standard error and uses the information continually to select the next item to be administered. Testing is ended if the termination criterion is satisfied, for example, if the standard error of the ability estimate drops below a predetermined value or when a certain (maximum) number of items have been administered. The adaptive testing procedure quickly converges to the true theta value, using significantly fewer items than required in a traditional test to obtain the same measurement accuracy (Hambleton & Swaminathan, 1985; Reckase, 1988; Weiss, 1983a). This latter feature also addresses the concerns of testing time often mentioned with regard to the dynamic (test-train-retest) assessment of learning potential. The LPCAT is not a timed test and each examinee gets the opportunity to answer each item administered to him/her. However, for practical reasons, a time limit of three minutes per test item administered was programmed in so that individuals would not get stuck on a particular item. If a question has not been answered within three minutes, the ability level will be re-estimated (adjusted downward) and an easier question will be administered next. The assumption is that if a person has not answered a question within three minutes, it is probably too difficult for him/her. If the next question is answered correctly again, the estimated ability level will be adjusted upward. The three minute time limit was determined based on preliminary empirical research.

CAT furthermore makes possible equiprecise measurement at different ability levels, since the termination criterion can be linked to the level of accuracy of measurement that has been achieved. Another important aspect is that adaptive tests are power tests and not timed tests. With adaptive testing procedures it is possible to administer varying numbers and different sets of items to individuals while scores remain comparable – since they reflect the level of the underlying trait. This same principle allows direct comparison of pretest and post-test scores of the same examinee as well as comparison of scores of different examinees, making CAT uniquely suitable and appropriate for the measurement of learning potential. Although IRT and CAT procedures seem particularly appropriate for learning potential assessment, since it improves both measurement accuracy and time-efficiency, no previous application of CAT procedures based on IRT for learning potential assessment could be found in the literature.

In summary

A need was identified to construct an instrument for the measurement of learning potential in the domain of general nonverbal figural reasoning ability. Such a test can lessen the impact of socioeconomic, cultural or educational background on test performance so that fluid cognitive ability can be assessed in a more equitable and culture-fair manner. Since the aim of a learning potential measure is to avoid material that is related to socioeconomic or educational background and to move away from the measurement of that which has been learnt previously, use of nonverbal figural material seemed most appropriate.

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

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

The LPCAT is intended to serve as a screening instrument that can be used mainly to counter inadvertent discrimination against disadvantaged groups, since it is not dependent upon either language proficiency or prior school learning and indicates present level of reasoning performance as well as potential future level of reasoning performance after relevant training. The LPCAT was developed as a dynamic computerised adaptive test specifically for South Africa's multicultural context, with the aim of addressing the need for a fair, psychometrically sound and time-efficient measure of learning potential in the domain of general nonverbal figural reasoning. It addresses the typical concerns regarding cross-cultural assessment in terms of construct measured, methods used and investigation of item bias (Van de Vijver, 2002).

METHOD

A large multicultural sample (N=2454) was used for item analysis by means of the three-parameter IRT model, which is best for scoring multiple-choice items (McBride, 1997). Although item parameters obtained from paper-and-pencil administration may differ from those obtained in computer-administration, practical considerations – in particular

in the more rural areas – made it impossible to administer LPCAT items by computer for item analysis purposes. According to Hetter, Segall and Bloxom (1997), item parameters calibrated from paper-and-pencil administration of items can be used in power CATs of cognitive constructs without changing the construct being assessed and without reducing reliability. The number of items that needed to be administered was too large to administer to examinees in a single test. This necessitated the construction of two paper-and-pencil forms with sufficient anchor items – items answered by both groups – to calculate item parameters for all items on the same scale. Once the final LPCAT test was constructed, smaller multicultural samples were used to obtain validity information (De Beer, 2000c)

Participants

A total of forty-one schools were randomly selected from three of the ten provinces where agreement to test in the schools could be reached. At each school 60 pupils, 30 from grade 9 and 30 from grade 11 were randomly selected for testing. Furthermore, in each grade group of 30 pupils, half the examinees were boys and half girls. In each grade samle group of 30, Form A and Form B of the test were alternated, thereby ensuring an equal distribution of the two forms between both the gender and the grade groups. The cultural group and language group allocation of each school was based on the education body that had previously been responsible for that school. At the time there was limited integration in government schools. Of the four main cultural groups in South Africa (African, Indian, Coloured and White), all but the Indian group were included in the paper-and-pencil sample. The reason for the exclusion of the Indian group was threefold. Firstly, they form only 2,5 percent of the South African population (CSS, 1996). Secondly, in cognitive test performance as well as in socioeconomic status and educational attainment, they are very similar to the White croup (CSS, 1996). Thirdly, the province with the highest representation of the Indian population was not one of the three provinces included for the paper-and-pencil item analysis test administration. Indian participants were later included in the validation of the LPCAT in its computerised format. The cultural and gender composition of the sample for the item analyis is given in Table 1.

<Place Table 1 here>

According to the 1996 census information (CSS, 1996), the percentage of the different cultural groups in South Africa is 76,3 percent African, 12,7 percent White, 8,5 percent Coloured and 2,5 percent Indian. The representation of these groups in the LPCAT item analysis sample is 49 percent African, 27 percent White and 24 percent Coloured. The African group is therefore underrepresented, while the Coloured and White groups are proportionally overrepresented. This distribution does, however, provide adequate numbers of examinees of the different subgroups for item analysis purposes. There was an almost equal gender distribution with 1228 male and 1226 female pupils included. Despite the fact that the sample cannot be considered to be statistically representative of the South African population (because of the lack of both regional and full cultural representation), in practical terms, it can be regarded as being representative of groups in South Africa. The sample sizes for the different subgroups were large enough to meet the

requirements of the procedures used for item analyisis – in particular for three-parameter IRT item analysis. A total of 1277 pupils completed form A and 1173 form B. The distribution of Form A and Form B per culture group was approximately equal within each culture group.

The form per culture group composition of the item analysis sample is given in Table 2. Note that due to some missing values, the total sizes for subgroups are not always exactly the same.

<Place Table 2 here>

Measuring instruments

The newly developed LPCAT
The 270 new LPCAT items (90 of each item type) were initially administered for item analysis purpose. IRT procedures can use anchor items to combine samples for item analysis purposes, and to this end, 66 anchor items were used (22 of each of the item types figure series, figure analogies and pattern completion). Once the final version of the LPCAT in its computerised adaptive version had been constructed it was administered to different groups to obtain validity information (De Beer, 2000c).

Procedure
Both classical and IRT item analysis were performed on the 270 new items that were developed for the LPCAT. Basic IRT assumptions were empirically investigated. Furthermore, items that did not meet the set psychometric standards (based on both Classical Test Theory (CTT) and IRT requirements) and/or items that showed bias, were discarded before the compilation of the final pretest and post-test item banks.

Reliability of the LPCAT was assessed by means of the test information function associated with IRT-based computerised adaptive testing. Validity of a test should be built into it from the outset. In terms of test validity, face validity, content validity and construct validity were attended to during the development of the LPCAT. Construct validity as well as concurrent and predictive criterion-related validity were investigated once the final test had been constructed (De Beer, 2000c)

Differential item functioning (DIF) analysis was performed to identify biased items in terms of level of education, gender, culture and language group. The final test was constructed, discarding items that did not meet the required standards based on psychometric or DIF information obtained. The item allocation to the pretest and post-test item banks were done after which the final test was constructed, incorporating appropriate termination rules and linking the pretest and the post-test. The results provide four scores, namely pretest, post-test, difference score and a composite score. The results of the LPCAT reflect the level of mental reasoning – comparable to typical levels of education – irrespective of age or attained level of education.

Analysis

Classical test theory item analaysis
The ITEMAN program of MicroCAT (Assessment Systems Corporation, 1995), which was used for the classical item analysis, scores items that are not reached as incorrect, and this affects the values obtained. Therefore, for the classical test theory item analysis, the items of the two forms had to be kept separate. Consequently, for the anchor items included in both forms, two sets of values were calculated. The classical item analysis information included the item difficulty value (p-value) as well as the item discrimination value (r_{it}). Another classical test theory index that was calculated is coefficient alpha, a measure of internal consistency or test homogeneity.

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

For the LPCAT IRT item analysis, Form A and Form B were combined into a single test of 270 items by using the 66 anchor items which all examinees completed. It was possible to combine the two forms for IRT item analysis because the ASCAL program of MicroCAT (Assessment Systems Corporation, 1995) works with dichotomously scored items and makes an important distinction between items that are coded as omitted and items that are coded as not reached. Items that are not reached are excluded from the analysis for the examinee concerned. The items from the alternate form which the examinee did not complete, were therefore coded as "not reached", which made it possible to combine the two groups, thereby increasing the available sample size for the anchor items.

General IRT assumptions checked
Three of the most important general assumptions of IRT are one-dimensionality, item parameter invariance and ability parameter invariance. These three assumptions were empirically investigated for the LPCAT total item bank. The results relating to this section are discussed elsewhere (De Beer, 2000c).

- One-dimensionality

It was decided to use the entire bank of items and not only those items that were included in the final version of the test for this analysis. Exclusion of items that were eventually discarded because they failed to comply with the standards set, is likely to positively affect these results. To investigate the one-dimensionality of the LPCAT items, the

factor structure was investigated for both the total group and specific subgroups to determine whether the same constructs were being measured for the different groups. LPCAT items were constructed to measure a single domain (general nonverbal, figural reasoning). In the case of the LPCAT items, factor analysis had to be executed separately for Form A and Form B.

- Item parameter invariance

According to Lord (1980, p. 35), "the invariance of item parameters across groups is one of the most important characteristics of item response theory". He warns that it has been so customary to think of item difficulty in terms of the proportion of correct answers, that it is sometimes hard to imagine how item difficulty can ever be invariant across groups that differ in ability level. The invariance of item parameters across groups means that if we determine the item parameters for a set of items with two separate groups of examinees independently, we can expect a linear relation to exist between the item parameters. This relation can be empirically investigated by means of scatter diagrams of the parameters calculated for two separate groups and also by obtaining the correlation between the two sets of values. According to Hambleton and Swaminathan (1985), it is desirable to identify subgroups of special interest in the examinee population and use them to study item parameter invariance. The item parameters of the LPCAT items were investigated by using two sets of independent groups, namely the two gender groups (male vs female) and the two home language groups (English/Afrikaans speaking vs African languages). The item parameters for these groups were calculated separately with the MicroCAT ASCAL program (Assessment Systems Corporation, 1995) and correlated to determine the relationship.

- Ability parameter invariance

Ability parameter invariance refers to the fact that in IRT, the ability parameter of a person is not affected by the items that are used to estimate it. According to Lord (1980), ability parameters are invariant from one test to another, except for the choice of origin and scale, assuming that the two tests both measure the same ability or (latent) trait. This characteristic can be empirically investigated by calculating the ability parameters of a group of examinees with two different sets of items and comparing (correlating) the results. This was originally investigated by determining the ability parameters of the total sample by means of different subsets of items based on item type (De Beer, 2000c).

Differential item functioning (DIF) analysis

The investigation of DIF helps to identify test items that may be unfair for members of certain groups (Zieky, 1993). Bias is a technical term which indicates some systematic error in the measurement process (Osterlind, 1983) and is generally considered to be a technical matter which requires careful scrutiny and statistical investigation of test items. Fairness of a test, on the other hand, indicates whether it is an equally valid measure of ability for individuals from different groups and deals with the social consequences of test use – often involving socially-based and more subjective evaluation of information. IRT has provided a major breakthrough in the study of DIF and its more sophisticated techniques contribute to improved procedures for measuring and analysing DIF. For

example, DIF can be investigated at particular ability levels or over the entire ability spectrum, which provides distinct advantage over classical methods. The way in which the item characteristic curves (ICCs) are used to evaluate DIF is by drawing them on the same graph to compare the ICCs of two groups. In DIF analysis, the examinee group of interest is referred to as the focal group, while the group to which its performance on the item is being compared is called the reference group (Holland & Wainer, 1983) by drawing the ICCs on the same graph. "A test item is said to be unbiased when the probability for success on the item is the same for equally able examinees of the same population regardless of their subgroup membership" (Osterlind, 1983, p. 3). If there is a distinct difference between the ICCs of the two groups, the item shows DIF. Such items are flagged so that they can be further evaluated and possibly scrapped if they do not meet the requirements for inclusion into the test bank.

The most common procedure for detecting bias is by means of calculating the area between the two ICCs (Wainer, 1983). To investigate the DIF for the LPCAT items, ICCs for the following four sets of groups were compared (see Figure 1):

- Language : African home language versus English/Afrikaans
- Culture: African versus White
- Gender: Male versus female
- Level of education: Grade nine versus grade 11

The only sample that could be considered somewhat small for the three-parameter item analysis to obtain the ICCs was the White group (N=658). All the other subgroups were sufficiently large (sample sizes larger than 1000) for three-parameter IRT analysis. The area between the two ICCs was calculated as the measure of DIF.

Criteria for item selection – psychometric and DIF information used

Classical test theory (CTT), IRT and DIF analyses were used to identify items suitable for inclusion in the final LPCAT pretest and post-test item pools. For the three-parameter IRT model used, the general consensus among researchers (Baker, 1985; Hambleton & Swaminathan, 1985; Sands, Waters & McBride, 1997; Weiss, 1983a) is that a-values should be within the range of 0,8 to 2,0 and c-values within the range of 0,0 to 0,3 for items to be included in a test. Classic item parameter values were also considered and no item with r_{it} below 0,3 was included, *unless* the a-value (IRT) for the same item was above 1,00. The condition in terms of the IRT a-value was included because items that discriminate well (a > 1,00) at a high ability level may not have high item reliability values (r_{it}), since very few examinees would get the correct answer for these items. Lastly, items were discarded if the area between the ICCs of any of the four DIF comparison groups was greater than 0,5 – based on mean values and visual inspection.

Compilation of the final test

Once the items that did not meet the selection criteria had been identified and discarded, the remaining items were allocated to the final pretest and post-test item banks. Altogether 188 items remained (65 figure series, 58 figure analogies and 65 pattern

completion items). As a first step, the remaining items of each item type were arranged in ascending order of item difficulty (b-values). Thereafter the items were allocated to the pretest and the post-test item banks sequentially in a 1:2 ratio (one item to the pretest, and the next two to the post-test). This was done separately for each of the three item types to ensure an even spread of item types and item difficulties in the pretest and post-test item banks. Approximately one-third of the selected items were thus allocated to the pretest item bank (N=63) and the remainder to the post-test item bank (N=125). McBride (1997) suggests that the number of items in an item bank should exceed by a ratio of 5 or 10 to 1, the number of questions an individual examinee will encounter. For the LPCAT, the number of items in the respective item banks exceeded (by a ratio of between 5 and 8 for the pretest and by a ratio of between 7 and 10 for the post-test), the number of questions and individual will encounter. Fewer items are administered in the pretest (between 8 and 12) than in the post-test (between 12 and 18). The pretest provides an initial general level of nonverbal figural reasoning performance. In the post-test, the pretest level of performance is used as entry level, and therefore a more accurate measure of performance is possible. This requires more items at all difficulty levels in the post-test.

Instructional screens and practice examples

Items were computerised with the MicroCAT Testing System (Assessment Systems Corporation, 1989). Screens to introduce the examinees to the test, to familiarise them with the keyboard and the two keys that will be used (space bar and Enter key) and to explain the answering procedure were computerised. Computer literacy is not required of examinees. After the initial introduction, practice examples are administered to familiarise the examinee with the types of items included in the test. Three screens were prepared to show the format of each of the three item types together with two practice examples for each item type to be aministered before the pretest. These examples give the examinees an opportunity to practice the answering procedure, and also to familiarise themselves with the strategies used to find the correct answer. For the practice examples, feedback is provided after each answer to inform the examinee whether the answer he or she chose was the correct one. The correct answer is also indicated and an explanation provided as to why that answer is the correct one. The screens to accomplish this were also prepared and computerised.

Training section

The dynamic test-train-retest format of the LPCAT involves a training section between the pretest and the post-test. The training is administered as part of a single test-train-retest administration session. In the training section, the screens for the three item types are repeated again, followed by information highlighting specific aspects that should be noted in finding the correct answers to these types of questions. More practice examples and additional training screens were prepared for this section of the test.

Language choice for test administration

Two versions of the LPCAT were constructed – one where all instructions, feedback and explanation is provided on the screen, and the second one, where instructions do not appear on the screen and where all instructions, feedback and explanation is read to the examinee by an examiner. Test instructions for the latter version are available in all 11

official South African languages in the LPCAT User's Manual (De Beer, 2000b). The text on screen version is available in English and Afrikaans. A reading proficiency level of grade 6 for English or Afrikaans is recommended to administer this version. In this latter version, four items to check the understanding of both the concepts and the terms used in the explanation are administered after the initial training screens. These "language" items are scored and a percentage mark allocated. If an individual answers more than one of these extremely easy questions incorrectly, it probably indicates that he or she did not understand the terms and/or concepts used in the feedback and training. Limited understanding of the instructions and feedback may consequently have affected the results negatively. It is recommended that individuals scoring below 75 percent should be retested with the version where instructions are read to him or her.

Scoring scale

The LPCAT was constructed as a computerised adaptive test using the three-parameter item response theory model (Assessment Systems Corporation, 1989; Lord, 1980). As such, the scaling of the (latent) ability level is on the theta scale with a mean of 0 and standard deviation of 1. The three-parameter model has been used most widely in CAT and can be regarded as a general model for dichotomously scored items (Assessment Systems Corporation, 1989). The Bayesian modal method is used for ability estimation and item selection in the LPCAT. The Bayesian item selection strategy selects items on the basis of minimising the Bayesian posterior variance of the ability estimate (Assessment Systems Corporation, 1989). Since the theta scale used for ability estimates in the standard three-parameter IRT model includes negative values, the final scores for the LPCAT are linearly transformed and provided in the form of T-scores with a mean of 50 and a standard deviation of 10 as well as percentile scores and stanine scores.

In adaptive testing, the entry level of difficulty of the first item to be administered can be specified when the test is constructed. In general, an item of average difficulty is usually presented first, after which the adaptive item selection process commences. In the case of the LPCAT, the difficulty of the version where instructions are given on screen was set at 0.0 (T-score 50), which is the mean value on the theta scale. Hence an item of average difficulty at a mid-secondary level is administered first. In the version where no text is given on the screen and where instructions are read to the examinee, the entry level was set at -1.0 (T-score 40). The result is that an easier first item at a senior primary difficulty level is administered at first, whereafter the adaptive testing process commences. Although the entry levels for the two versions differ, examinees with any level of ability can be tested with any of the two test forms. Because of the adaptive testing procedure, there is no floor or ceiling effect in either of the two versions. The level of reading proficiency of the examinee and the testing context will determine which version is most appropriate.

In the construction of the final adaptive testing procedures for the LPCAT pretest and post-test respectively, the following was used:

- A list of items for each item pool from which items can be selected
- The variance of ability estimation to be used as the termination criterion
- The minimum number of items to be administered

- The maximum number of items to be administered (Assessment Systems Corporation, 1989, 1995).

During the CAT procedure items are selected, based on their difficulty level, to match the estimated ability level of the examinee at that time.. Items are sampled without replacement from the specified pool and administered to the examinee until one of the termination criteria is reached.

Reliability of the LPCAT

The classical indices of reliability namely test-retest reliability, parallel forms reliability and split-half reliability do *not* apply to computerised adaptive testing. This is because of the interactive selection of items from an item bank which results in different sets of items being administered to each examinee. One classical test theory method of evaluating reliability that can be applied is the internal consistency or coefficient Alpha index, which also reflects the homogeneity of content (Anastasi & Urbina, 1997). This aspect relates to the one-dimensionality requirement for using IRT. Coefficient Alpha was determined for the entire item analysis sample as well as for various subgroups.

Using IRT, one can predict certain characteristics of a test before it is administered, since the item parameters have been previously determined. Test information is an index of the precision of measurement that a test can provide and is directly related to the measurement effectiveness of a test (Assessment Systems Corporation, 1989; 1995; De Beer, 2000c). The test information function graphically indicates the amount of information at various ability levels, when specific items are included in a test. It is furthermore possible to compare the effect of administering various numbers of items on the information levels achieved.

The IRT equivalent to test score reliability and standard error of measurement is the test information function (Assessment Systems Corporation, 1989; 1995; De Beer, 2000a). Since the information function may vary from one ability level to the next, the standard error may also vary and needs to be calculated for a specific ability level. In CAT, where the variance of the estimation of ability is incorporated as one element used for test termination for each individual, equal accuracy of measurement is more attainable than with standard tests.

Validity of the LPCAT

In the case of the LPCAT, face validity and content validity was judged by a panel of psychological test development experts. Construct validity was assessed by correlating results with standard cognitive tests, while concurrent and predictive criterion-related validity was assessed by means of correlation with standard tests and academic or other training performance (De Beer, 2000c; De Beer, 2002).

RESULTS

Classical test theory item analysis results
The mean p-value for Form A was 0.656 and that for Form B was 0.657, while the mean r_{it}-values were 0.498 and 0.476 respectively.

Based on all the items that were administered for item analysis, the coefficient alpha values ranged between 0.925 and 0.979 for the various subgroups. The alpha value for the total group was 0.981 for Form A and 0.978 for Form B, indicating high internal consistency. Coefficient Alpha is regarded as an index of reliability in standard tests, and according to Gregory (2000) can be regarded as an index of the degree to which a test measures a single factor. Table 3 provides the coefficient alpha values for the two test forms (A and B) for both the total group and various subgroups. The high values obtained for coefficient alpha provide support for the one-dimensionality of the LPCAT items – a requirement for the use of the three-parameter IRT model.

<Place Table 3 here>

IRT item analysis results
Five of the original 270 items were discarded during initial IRT analysis because of problems with some of the distractors. The item parameters are estimated through an iterative process. A descriptive summary of the values of the item bank before item selection is provided in Table 4.

<Place Table 4 here>

The mean a-value indicates that, on average, items discriminate well, while the mean b-value, being less than 0.0 indicates that most items are reasonably easy. Selection of the items to be included in the final version of the LPCAT was based on both classical and IRT item analysis (as well as DIF analysis), although greater weight was attached to IRT item parameters.

IRT general assumptions results

One-dimensionality
The factor analysis results indicate support for a one-dimensional structure for both the total group and the various subgroups. The eigenvalues for the different groups are reported in Table 5. For both Form A and Form B, the eigenvalues for the first factor were between 6.54 and 8.92 times larger than the eigenvalue for the second factor for the toal, African and Coloured groups. The eigenvalues of subsequent factors were significantly closer to each other. The exception to the above ratios of eigenvalues was for the White group where for Form A the first eigenvalue was only 2.65 times the size of the second. This ratio was 3.33 for Form B. Considering the item types and item content used and the similarity between strategies required to solve the items, the above results provide support for the expected one-dimensionality of the LPCAT items. Scree tests also provided support for the one-dimensional nature of the LPCAT item domain (De Beer, 2000a).

<Placed Table 5 here>

Item parameter invariance

The item parameter invariance was investigated by obtaining scatter diagrams and correlation coefficients of the two sets of values when independent groups were used to obtain the item parameters. For the LPCAT the two gender groups and two language subgroups were used. The correlation results are reported in Table 6. The results indicate support for the item parameter invariance for the LPCAT items. All correlations are highly significant, with the b-value correlations which indicate item difficulty, the highest.

<Place Table 6 here>

Ability parameter invariance

In the case of the LPCAT, ability parameter invariance was investigated for three different sets of item combinations by using the separate item types to independently calculate the ability estimates for the total item analysis sample group of examinees and to obtain scatter diagrams and correlations for these values (De Beer, 2000a). The correlation results are reported in Table 7. The distributions and correlations found here are similar to those found in other such studies (Gierl & Hanson, 1995).

<Place Table 7 here>

DIF analysis

An example of the ICCs for one item for the different comparison groups is provided in Figure 1. Deciding how large an area would justify scrapping an item is somewhat subjective since no clearcut indices are provided in the literature. The general consensus is that a combination of visual inspection and empirical estimation of cutoffs should be used for flagging DIF items to be scrapped from the item pool. Considering the nature or the LPCAT items, no bias was expected for any of the subgroups involved.

<Place Figure 1 here>

For the LPCAT, an item was considered to show DIF (ie to be biased) if the area between the two curves exceeded 0.5. DIF items were discarded purely on the magnitude of the DIF indices, irrespective of the particular group against which it was considered biased. The mean values of the areas calculated for the four sets of comparison groups are provided in Table 8.

<Place Table 8 here>

Criteria for item selection

Altogether 47 items (of the original 270) were discarded on the basis of the IRT and CTT item criteria and an additional 35 items were discarded on the basis of DIF, bringing the total of discarded items to 82, or approximately 30 percent. This percentage is comparable to the findings of similar research projects. Adaptive testing demands higher quality (more discriminating) test items than conventional testing as well as more variability in item difficulty level, and in practice only about one in three items is useful for adaptive testing (McBride, 1997). A summary of the number and types of items that were discarded is provided in Table 9.

<Place Table 9 here>

Construction of the final test

The distribution of the final pretest and post-test items is indicated in Table 10.

<Place Table 10 here>

Because of the way in which items were allocated to the pretest and the post-test item banks respectively, the mean b-parameters (difficulty values) in the pretest and post-test are very similar and cover a wide range of difficulty levels. Having items available over a wide range of difficulty levels, and administering items in an adaptive manner, means that the LPCAT can provide information over a wide spectrum of ability levels.

For the pretest, between eight and 12 items are adaptively administered in the pretest from an item bank of 68 items, while in the post-test, between 12 and 18 items are adaptively administered from an item bank of 125 items. The cutoff in terms of the variance was put at 0,10 for the pretest and 0,05 for the post-test respectively, based on empirical values obtained during the LPCAT standardisation.

The LPCAT results consist of four different scores:

- The pretest score (T-score, stanine and percentile score) which reflects the level of performance at the end of the pretest, which is indicative of the actual development level
- The post-test score (T-score, stanine and percentile score) which represents the potential level of performance
- The difference score (T-score) which represents the ZPD or the magnitude of undeveloped potential
- The composite score (single score on the T-score scale) which is a conservative, reasoned combination of the present level of performance together with proportional credit for improvement shown at that level. The advantage of the composite score is that people at different levels of initial performance and with different ZPDs can be compared in a systematic manner.

Reliability of the LPCAT

As indicated earlier, the standard reliability indices do not apply in the case of computerised adaptive tests. The reliability indices available for the LPCAT are the

coefficient alpha values (see Table 3). The IRT equivalent to test score reliability is the test information function, which allows for the calculation of the standard error at specific ability levels, since it is dependent upon the level of test information at that particular level. The standard error of measurement at any level of ability is the reciprocal of the square root of the amount of test information at that level. Since the accuracy of the ability estimation is used as one of the termination criteria in adaptive testing, equal or similar accuracy of measurement at different ability levels is more attainable than with standard tests. Based on the test information available in the LPCAT pretest and post-test item banks respectively, in the pretest, roughly 68 percent of the T-score estimates will fall between -2.4 and +2.4 T-scores from the estimated ability level, and for the post-test, roughly 68 percent of the T-score estimates will fall between -1.7 and +1.7 T-scores from the estimated post-test ability level. The fact that the information levels at the extremes of ability is lower than in the centre region, means that more items will have to be administered to examinees who fall close to either of the extremes in their ability level to reached the required levels of accuracy of ability estimation.

Validity of the LPCAT

Although the focus of this article is not the validity of the instrument, some validity information is provided for the sake of completeness. Face and content validity of the LPCAT was judged to be good by a panel of test development experts involved in reviewing the development of the test (De Beer, 2000c). In terms of construct validity as indicated by comparison of LPCAT results with results of existing cognitive instruments, correlations were statistically highly significant and ranged between 0.400 and 0.645 for comparison with the Paper-and-Pencil Games (Claassen, 1996), and between 0.567 and 0.691 for comparison with the General Scholastic Aptitude Test (Claassen, De Beer, Hugo & Meyer, 1991) at secondary school level (De Beer, 2000c). In terms of criterion-related validity the results were compared with ABET training results for an adult group and with academic results for secondary and junior tertiary groups. For the low literate adult group, correlations between LPCAT results and training results ranged between 0.398 and 0.610, while for a secondary level school sample, correlations of academic results and LPCAT performance ranged between 0.439 and 0.543 (De Beer, 2000c). In a separate study with a group of bridging students, correlation of LPCAT and academic results ranged between 0.313 and 0.525 (De Beer, 2002). These results provide support for the validity of the LPCAT in the multicultural South African context.

DISCUSSION

Although the LPCAT was not developed with the Employment Equity Act of 1998 in mind – since its development started a number of years before the EEA was tabled – it does comply with the requirements of the EEA regarding psychological tests. It was developed with the multicultural context of the South African society in mind and provides for culture-fair, yet psychometrically sound assessment of learning potential in the nonverbal figural reasoning domain. Typical concerns in cross-cultural assessment regarding similarity of the construct measured were investigated by means of factor

analysis and construct validity, the dynamic test-teach-test approach was used to lessen method bias and IRT based DIF analysis was used to eliminate biased items, contributing to the development of a psychometrically sound and practically useful tool for the measurement of learning potential in multicultural contexts. The LPCAT can provide useful information for training and development, so that training can be matched with present and potential future levels of reasoning ability. In this way it helps to provide optimal developmental opportunities for individuals over a wide spectrum of ability, while taking into account that prior learning opportunities may have been very different.

REFERENCES (start on a separate page)

Anastasi, A. & Urbina, S. (1997). *Psychological Testing* (7th Edition). Upper Saddle River: Prentice-Hall.

Assessment Systems Corporation. (1989). *User's manual for the MicroCAT testing system (Version 3.0)*. St Paul: Assessment Systems Corporation.

Assessment Systems Corporation. (1995). *User's manual for the MicroCAT testing system (Version 3.5).* St Paul: Assessment Systems Corporation.

Baker, F.B. (1985). *The basics of Item Response Theory.* Portsmouth, N.H.: Heinemann.

Binet, A. & Simon, T. (1905/1916). *The intelligence of the feeble-minded*. Baltimore: Williams and Wilkins.

Binet, A. & Simon, T. (1915). *A method of measuring the development of the intelligence of young children*. Chicago: Chicago Medical Book Co.

Blagg, N. (1991). *Can we teach intelligence? A comprehensive evaluation of Feuerstein's instrumental enrichment program*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Boeyens, J. (1989a). *Learning potential: A theoretical perspective. Report PERS-432,* Human Sciences Research Council. Pretoria: Human Sciences Research Council.

Boeyens, J. (1989b). *Learning potential: an empirical investigation. Report PERS-435*, Human Sciences Researches Council. Pretoria: Human Sciences Research Council.

Brown, A.L. & French, L.A. (1979). The zone of potential development: Implications for intelligence testing in the year 2000. *Intelligence, 3*, 255-273.

Budoff, M. (1969). Learning potential: A supplementary procedure for assessing the ability to reason. *Seminars in Psychiatry, 1(3)*, 278-290.

Budoff, M. (1987a). The validity of learning potential assessment. In C.S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 52-81). New York: The Guilford Press.

Budoff, M. (1987b). Measures for assessing learning potential. In C.S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 173-196). New York: Guilford Press.

Budoff, M. & Harrison, R.H. (1971). Educational tests of the learning-potential hypothesis. *American Journal of Mental Deficiency, 76(2)*, 159-169.

Campione, J.C. & Brown, A.L. (1987). Linking dynamic assessment with school achievement. In C.S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82-115). New York: The Guilford Press.

Campione, J.C., Brown, A.L. & Ferrara, R.A. (1982). Mental retardation and intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 392-492). Cambridge: Cambridge University Press.

Campione, J.C., Brown, A.L., Ferrara, R.A. & Bryant, N.R. (1984). The zone of proximal development: Implications for individual differences in learning. In B. Rogoff & J.V. Wertsch (Eds.), *Children's learning in the 'zone of proximal development'* (pp. 77-91). San Francisco: Jossey-Bass.

Carlson, J.S. (1989). Advances in research on intelligence: The dynamic assessment approach. *The Mental Retardation and Learning Disability Bulletin, 17(1)*, 1-20.

Central Statistical Service of South Africa (CSS) (1996). *Census in Brief - Report number 1:03-01-11 (1996)*. Pretoria: Central Statistical Service.

Claassen, N.C.W. (1996). *Paper and pencil games (PPG): Manual*. Pretoria: Human Sciences Research Council.

Claassen, N.C.W. (1997). Cultural differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47(4)*, 297-307.

Claassen, N.C.W., De Beer, M., Hugo, H.L.E. & Meyer, H.M. (1991). *Manual for the General Scholastic Aptitude Test (GSAT) Senior Series*. Pretoria: Human Sciences Research Council.

Dague, P. (1972). Development, application and interpretation of tests for use in French-speaking black Africa and Madagascar. In L.J. Cronbach & P.J.D. Drenth, (Eds.), Mental tests and cultural adaptation (pp. 64-74). The Hague: Mouton.

De Beer, M. (2000a). The construction and evaluation of a dynamic computerised adaptive test for the measurement of learning potential. Unpublished D.Litt et Phil dissertation. University of South Africa, Pretoria.

De Beer, M. (2000b). *Learning Potential Computerised Adaptive Test (LPCAT): User's Manual*. Pretoria: Production Printers (Unisa)

De Beer, M. (2000c). *Learning Potential Computerised Adaptive Test (LPCAT): Technical Manual*. Pretoria: Production printers (UNISA)

De Beer, M. (2002). *Utility of learning potential computerised adaptive test (LPCAT) scores in predicting academic performance of bridging students: A comparioson with other predictors.* Paper presented at the 5th Annual Industrial Psychology Conference, Pretoria, CSIR, 13-14 June 2002.

De Beer, M. & Van Eeden, R. (1997). *Selection criteria for students in engineering and other science and technology courses at M.L. Sultan Technikon*. Unpublished research report.

Embretson, S.E. (1987). Toward development of a psychometric approach. In C.S. Lidz (Ed.), *Dynamic assessment. An interactional approach to evaluating learning potential* (pp. 141-170). New York: Guilford Press.

Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56(3)*, 495-515.

Embretson, S.E. (1992). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement, 29(1)*, 25-50.

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8(4)*, 341-349.

Embretson, S.E. & Reise, S.P. (2000). Item response theory for psychologists. Mahwah, Lawrence Erlbaum Associates.

Employment Equity Act, No 55 (1998). *Government Gazette, 400 (19370). Cape Town, 19 October 1998.*

Feuerstein, R. (1979). *The dynamic assessment of retarded performance*. Baltimore, MD: University Park Press.

Feuerstein, R., Feuerstein, R. & Gross, S. (1997). The learning potential assessment device. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 297-313). New York: The Guilford Press.

Feuerstein, R., Rand, Y., Jensen, M.R., Kaniel, S. & Tzuriel, D. (1987). Prerequisites for testing of learning potential: The LPAD model. In C.S. Lidz (Ed.), *Dynamic testing* (pp. 35-51). New York: Guilford Press.

Foxcroft, C.D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment, 13(3)*, 229-235.

Frisby, C.L. & Braden, J.P. (1992). Feuerstein's dynamic assessment approach: A semantic, logical and empirical critique. *The Journal of special Education, 26(3)*, 281-301.

Gierl, M.J. & Hanson, A. (1995). *Evaluating the goodness-of-fit between Alberta education achievement test data and model assumptions in unidimensional item response theory.* Unpublished research report prepared for the Alberta education, student evaluation branch.

Gould, S.J. (1981). *The mismeasure of man*. New York: W.W. Norton.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21(4)*, 347-360.

Gregory, R.J. (2000). *Psychological testing: History, principles, and applications*. (Third ed.). Boston: Allyn & Bacon.

Grigorenko, E.L, & Sternberg, R.J. (1998). Dynamic testing. *Psychological Bulletin, 124(1)*, 75-111.

Guthke, J. (1992). Learning tests - the concept, main research findings, problems and trends. *Learning and Individual Differences, 4(2)*, 137-151.

Guthke, J. (1993b). Developments in learning potential assessment. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 43-68). Amsterdam: Swets & Zeitlinger.

Guthke, J. (1998). *Validity of learning test versions of the Raven test*. Paper presented at the 24th International Congress of Applied Psychology, San Francisco, 9-14 August, 1998.

Hambleton, R.K. & Slater, S.C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment, 13(1)*, 21-28.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhof Publishing.

Hambleton, R.K. & Zaal, J.N. (Eds.). (1991). *Advances in educational and psychological testing: Theory and applications*. Boston: Kluwer Academic Publishers.

Hamers, J.H.M. & Resing, W.C.M. (1993). Learning potential assessment: Introduction. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (Eds.), *Learning potential*

assessment: Theoretical, methodological and practical issues (pp. 23-42). Amsterdam: Swets & Zeitlinger.

Health Professions Council of South Africa (HPCSA) (1998). *Policy on the classification of psychometric measuring devices, instruments, methods and techniques. (18/9/B)*. Pretoria: South African Professional Board for Psychology.

Hetter, R.D., Segall, D.O. & Bloxom, B.M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W.A. Sands, B.K. Waters & J.R. McBride, *Computerized adaptive testing: From inquiry to operation* (pp. 161-168). Washington, DC: American Psychological Association.

Holland, P.W. & Wainer, H. (1993). Preface. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. xiii - xv). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hugo, H.L.E. & Claassen, N.C.W. (1991). *The functioning of the GSAT Senior for students of the Department of Education and Training*. Pretoria: Human Sciences Research Council.

Huysamen, G.K. (2002). The relevance of the new APA standards for educational and psychological testing for employment testing in South Africa. *South African Journal of Psychology, 32(2), 26-33.*

Jensen, A.R. (1963). Learning ability in retarded, average, and gifted children. *Merrill-Palmer Quarterly, 9(2)*, 123-140.

Jensen, A.R. (1980). *Bias in mental testing*. London: Methuen.

Jensen, A.R. (1981). *Straight talk about mental tests*. London: Methuen.

Kozulin, A. & Falik, L. (1995). Dynamic cognitive assessment of the child. *Current Directions in Psychological Science, 4(6)*, 192-196.

Lidz, C.S. (1991). *Practitioner's guide to dynamic assessment*. New York: The Guilford Press.

Loehlin, J.C. (1992). Should we do research on race differences in intelligence? *Intelligence, 16,* 1-4.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McBride, J.R. (1997). Technical Perspective. In W.A. Sands, B.K. Waters & J.R. McBride, *Computerized adaptive testing: From inquiry to operation* (pp. 29-44). Washington, DC: American Psychological Association.

Meijer, R.R. & Nering, M.L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23(3),* 187-194.

Murphy, R. (2002) A review of South African research in the field of dynamic assessment. Unpublished MA thesis. Pretoria: University of Pretoria.

Osterlind, S.J. (1983). *Test item bias*. Beverly Hills: Sage.

Owen, K. (1998). *The role of psychological tests in education in South Africa: Issues, controversies and benefits*. Pretoria: Human Sciences Research Council.

Plomin, R. (1997). Genetics and intelligence: What's New? *Intelligence, 24(1)*, 53-77.

Psychological Society of South Africa (PSYSSA). (1998a). *Guidelines for the validation and use of assessment procedures for the workplace*. Pretoria: Society for Industrial Psychology.

Psychological Society of South Africa (PSYSSA). (1998b). *Code of practice for psychological assessment for the work place in South Africa*. Pretoria: Society for Industrial Psychology.

Pyryt, M.C. (1996). IQ: Easy to bash, hard to replace. *Roeper Review, 18(4)*, 255-258.

Reckase, M.D. (1988). *Computerized adaptive testing: a good idea waiting for the right technology*. Paper presented at the meeting of the American Educational Research Association, New Orleans, April 1988.

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8(3)*, 11-15.

Sands, W.A., Waters, B.K. & McBride, J.R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

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.

Sijtsma, K. (1993a). Classical and modern test theory with an eye toward learning potential testing. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars, *Learning Potential Assessment: Theoretical, methodological and practical issues* (pp. 117-134). Amsterdan: Swets & Zeitlinger.

Sijtsma, K. (1993b). Psychometric issues in learning potential assessment. In J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars, *Learning Potential Assessment:*

Theoretical, methodological and practical issues (pp. 175-194). Amsterdan: Swets & Zeitlinger.

Skills Development Act, No 97 (1998). *Government Gazette, 401 (19420). Cape Town, 2 November 1998.*

Taylor, T.R. (1992). *Beyond competence: Measuring potential in a cross-cultural situation fairly: Potential in Psychometrics (part two).* Paper presented at the 1992 South African Psychometrics Congress, June 1992.

Taylor, T.R. (1994a). A review of three approaches to cognitive assessment, and a proposed integrated approach based on a unifying theoretical framework. *South African Journal of Psychology, 24(4)*, 184-193.

Taylor, T.R. (1994b). *Learning potential: Theory and practical assessment*. Paper presented at the Annual Department of Industrial Psychology Conference on Fairness in Testing and Assessment (30 August 1994). Pretoria: Holiday Inn.

Thorndike, R.M. & Lohman, D.F. (1990). *A century of ability testing*. Chicago: The Riverside Publishing Company.

Tzuriel, D. (1997). A novel dynamic assessment approach for young children: Major dimensions and current research. *Educational and Child Psychology, 14(4)*, 83-108.

Van de Vijver, F. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-cultural Psychology, 28(6)*, 678-709.

Van de Vijver, F. (2002). Cross-cultural Assessement: Value for money? *Applied Psychology: An international Review, 51(4),* 545-566.

Van Eeden, R. (1993). *The validity of the Senior South African Individual Scale - Revised (SSAIS-R) for children whose mother tongue is an African language: Private schools.* Pretoria: Human Sciences Research Council.

Van Niekerk, H.A. (1991). *Evaluation of Feuerstein's instrumental enrichment programme for culturally different senior secondary students*. Unpublished Master's Thesis, University of Pretoria.

Verster, J.M. & Prinsloo, R.J. (1988). The diminishing test performance gap between English speakers and Afrikaans speakers in South Africa. In S.H. Irvine & J.W. Berry, (Eds.), *Human abilities in cultural context* (pp. 534-560). Cambridge: Cambridge University Press.

Vincent, K.R. (1991). Black/White IQ differences: Does age make the difference? *Journal of Clinical Psychology, 27(2)*, 266-270.

Von Hirschfeld, S. (1992b). *The psychological concept of potential*. Paper presented at the South African Psychometrics Congress, June 1992.

Vygotsky, L.S. (1978) *Mind in society: The development of higher-order psychological processes*. Cambridge, MA: Harvard University Press.

Wainer, H. (1993). Model-based standardised measurement of an item's differential impact. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D.J. (Ed.) (1983a). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Weiss, D.J. (1983b). Introduction. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Weiss, D.J. & Vale, C.D. (1987). Adaptive testing: *Applied Psychology: an International review, 36(3&4)*, 249-262.

Wolf, T. (1973). *Alfred Binet*. Chicago: University of Chicago Press.

Zaaiman, H., Van der Flier, H.R. & Thijs, G.D. (2001). Dynamic testing in selection for an educational programme: Assessing South African performance on the Raven Progressive Matrices. *International Journal of Selection and Assessment, 9(3),* 258-269.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

ATTACHMENTS

FIGURE 1

EXAMPLES IF ITEM CHARACTERISTIC CURVE (ICC) BIAS ANALYSIS GRAPHS FOR ONE ITEM

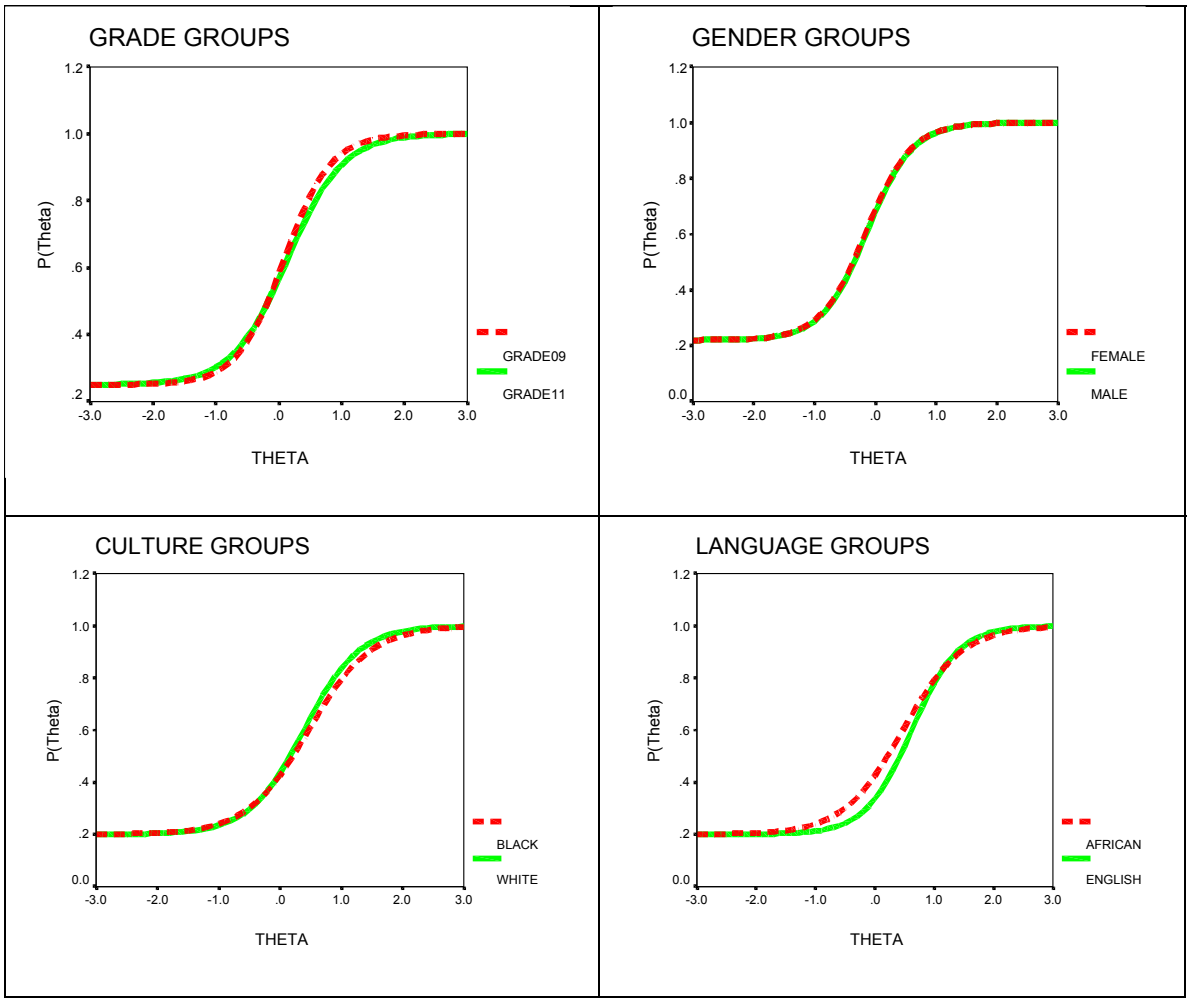

TABLE 1
CULTURE AND GENDER COMPOSITION OF THE ITEM ANALYSIS SAMPLE

Group	African pupils	Coloured pupils	White pupils	Total
Male	600	300	328	1228
Female	597	299	330	1226
Total	1197	599	658	2454

TABLE 2
FORM AND CULTURE DISTRIBUTION OF THE ITEM ANALYSIS SAMPLE

Cultuer group	Form A only	Form B only	Total
African	639	554	1193
Coloured	303	296	599
White	335	323	658
Total	1 277	1 173	2 450

TABLE 3
COEFFICIENT ALPHA VALUES FOR THE TWO TEST FORMS FOR DIFFERENT GROUPS

GROUP	Form A (168 items)		Form B (168 items)	
	N	Alpha	N	Alpha
Total group	1277	0.981	1173	0.978
African	639	0.975	554	0.971
Coloured	303	0.969	296	0.970
White	335	0.925	323	0.926
African language group	639	0.975	554	0.971
English/Afrikaans language group	638	0.973	619	0.971
Male	636	0.981	589	0.979
Female	640	0.980	584	0.978
Grade 7	622	0.980	600	0.977
Grade 9	653	0.981	572	0.979

TABLE 4

DESCRIPTIVE STATISTICS OF ITEM PARAMTERS OF THE ITEMS SUBJECTED TO IRT ITEM ANALYSIS

IRT parameter	N	Mean	SD	Minimum	Maximum
a-value	265*	1.435	0.486	0.442	2.500
b-value	265	-0.231	0.829	-1.558	3.000
c-value	265	0.179	0.085	0.000	0.470

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

TABLE 5
EIGENVALUES AND PERCENTAGE OF VARIANCE FOR DIFFERENT GROUPS FOR FORM A AND FORM B ITEMS

Group	Factor 1 Eigenvalue	Factor 1 Variance	Factor 2 Eigenvalue	Factor 2 Variance	Factor 3 Eigenvalue	Factor 3 Variance
Form A Total	44.552	26.519	5.721	3.406	3.488	2.076
Form B Total	41.736	24.843	4.678	2.784	3.329	1.982
Form A African	37.264	22.181	4.537	2.701	3.961	2.358
Form B African	33.645	20.027	3.784	2.252	3.473	2.067
Form A Coloured	34.012	20.245	5.200	3.095	3.991	2.376
Form B Coloured	33.990	20.232	5.054	3.008	4.238	2.522
Form A White	17.618	10.487	6.659	3.964	6.343	3.776
Form B White	18.032	10.734	5.418	3.225	4.532	2.698

TABLE 6
CORRELATIONS BETWEEN ITEM PARAMETERS FOR DIFFERENT SUBGROUPS (N=265 ITEMS)

Subgroups compared	b-parameter	a-parameter	c-parameter
Gender groups (Male vs female)	0.948*	0.813*	0.715*
Language groups (English/Afrikaans vs African)	0.945*	0.558*	0.454

* $p < 0.001$

TABLE 7
CORRELATIONS BETWEEN ABILITY PARAMETERS OBTAINED FROM DIFFERENT ITEM TYPES (N=2450)

Item types used for comparison	Correlation
Figure series vs Figure analogies	0.859*
Figure series vs Pattern completion	0.836*
Figure analogies vs Pattern completion	0.873*

* $p < 0.001$

TABLE 8
DESCRIPTIVE STATISTICS FOR DIF AREAS BETWEEN ICCs FOR DIFFERENT COMPARISON GROUPS (N=265 ITEMS)

DIF comparison groups	Mean	SD	Minimum	Maximum
Level of education (grade 9 versus grade 11)	0.1789	0.1471	0.0025	1.2338
Gender groups (male versus female)	0.1672	0.1616	0.0089	1.4375
Culture groups (African versus White)	0.3307	0.2081	0.0254	1.4050
Language groups (African versus English/Afrikaans)	0.2336	0.1570	0.0083	0.9762

TABLE 9
NUMBER AND TYPES OF ITEMS DISCARDED AS A RESULT OF ITEM ANALYSIS AND DIF ANALYSIS

Procedure	Figure Series	Figure Analogies	Pattern Completion	Total
Item analysis (IRT and CTT)	17	15	15	47
DIFF analysis	8	17	10	35
Total	25	32	25	82

TABLE 10
NUMBER AND TYPES OF ITEMS OF DIFFERENT TYPES ALLOCATED TO THE PRETEST AND POST-TEST

Item Type	Pretest	Post-test	Total
Figure series	21	44	65
Figure analogies	20	38	58
Pattern completion	22	43	65
Total	63	125	188