May 2009

# The role of IRT in selected examination systems

UMALUSI

Council for Quality Assurance in
General and Further Education and Training

# The role of IRT in selected examination systems

SJ Howie
C Long
V Sherman
E Venter

An Umalusi Research Report
September 2008

PUBLISHED BY

UMALUSI

Council for Quality Assurance in
General and Further Education and Training

37 General Van Ryneveld Street, Persequor Technopark, Pretoria
Telephone: 27 12 3491510 • Fax: 27 12 3491511
Email: Info@umalusi.org.za • Web: www.umalusi.org.za

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| ASPA | Aggregated Subject Pairs Analysis |
| CEA | Centre for Evaluation & Assessment |
| CEVO | Central Committee for Ratification of Examination |
| Cito | National Institute for Educational Measurement |
| CTT | Classical Test Theory |
| Ebtanas | Evaluasi Belajar Tahap Akhir Nasional |
| GCSE | General Certificate of Secondary Education |
| GDP | Gross Domestic Product |
| HAVO | Higher general continued education |
| IRT | Item Response Theory |
| JMLE | Joint Maximum Likelihood Estimation |
| KS3 | Key Stage 3 |
| MAVO | Junior general secondary education |
| MBO | Middle-level vocational education |
| MOEC | Ministry of Education and Culture |
| NABTEB | National Business and Technical Examination Board |
| NAEP | National Assessment of Educational Progress |
| NBEM | National Board for Educational Measurement |
| NCES | National Center for Education Statistics |
| NECO | National Examinations Council |
| NTCE/NBCE | National Technical and Business Certificate Examinations |
| OBE | Outcomes-based education |
| RNCS | Revised National Curriculum Statement |
| SPA | Subject Pairs Analysis |
| SSCE | Senior School Certificate Examination |
| TEE | Tertiary Entrance Examination |
| TER | Tertiary Entrance Ranking |
| TES | Tertiary entrance score |
| UAN | Unjian Akhir Natsional |
| UME | University Matriculation Examination |
| UN | Unjian Natsional |
| VBO | Pre vocational education |
| VMBO | Preparatory middle-level vocational education |
| VWO | Preparatory scientific education |
| WA | Western Australia |
| WACE | Western Australian Certificate of Education |
| WAEC | West African Examinations Council |
| WASSCE | West African Senior School Certificate Examination |
| WO | Scientific education |

# Section 1:

## Introduction

The existing national examination for the matriculation certificate is being phased out and a new national set of examinations is being introduced in 2008. The 2008 national examination is the final stage of the introduction of the Revised National Curriculum Statement (RNCS). This research report is intended to support the process of maintaining and improving examination standards. It also serves as a contribution to the broader project being undertaken by Umalusi to explore the use of Item Response Theory (IRT) for linking assessments across matriculation examinations from year to year. The intention of exploring the use of IRT is that this may provide additional information for Umalusi to report more meaningfully on the standards in education. In the medium to long term, the application of IRT could provide a means for retrospective analysis of matriculation examinations and by analysing test items that have been piloted, build an item bank from which items could be used to monitor standards longitudinally.

Over the past two decades, there has been an increase in the demand for the accountability of education systems and, therefore, the monitoring of learning outcomes internationally. This has resulted in the increase of national assessments and international comparative studies focusing on learning outcomes. On the other hand, examinations have been well established in both developing and developed environments for decades and, in some cases, for centuries. These examination systems are very important to the broader society and functioning of those societies. In several countries, national examinations are implemented at the end of primary school, in addition to being offered in secondary schools. The final examination in the primary school is often for selection of the best candidates in systems that are unable to provide for all children to attend secondary school (e.g. Mozambique). The information derived from these examinations is primarily for certification purposes and selection into further phases of learning or into employment. As a consequence, the examination attempts to achieve maximum discrimination for those students for whom the probability of selection is high. This is done by excluding items that are easy or of intermediate difficulty; if most students answered an item correctly the item would not discriminate among the high-scoring students (Greaney & Kellaghan, 1996).

In primary education, core subjects are assessed in these examinations whilst at secondary level students tend to select and specialise in subject areas. Subjects offered vary from one examination authority to another, but it is not unusual to find syllabi and examinations in 30 or more subjects (Greaney & Kellaghan, 1996, p. 32).

Public examinations can often appear to be relatively unstructured and students may write extended essays where the scoring procedures are not clearly specified and therefore rely heavily on the professional judgement of individual markers. Students may also choose particular subsets of questions that they want to answer. With regard to scoring and reporting in public examinations, these usually follow norm-referenced procedures (Greaney & Kellaghan, 1996, p. 33). The emphasis is on how the candidate performed in relation to the other candidates in the examination. The grades awarded merely imply that one grade is higher than or lower than another grade, rather than specifying a specific level of knowledge and skills.

One important aspect to consider is that, with the expansion of educational provision and greater numbers sitting for public examinations, as is the case in South Africa, the characteristics of examinees change. Increased participation rates inevitably result in decreases in the average level of achievement in a variety of school subjects (Keeves, 1994, Willmott, cited in Greaney & Kellaghan, 1996, p. 35). Furthermore, in addition to changes in the population, changes occur in the public examinations themselves from year to year. It is also common practice to release the examination papers to serve as guides for future years. This practice of annually designed

and set examination papers means that the same standard may not be maintained. A clear definition of standards needs to be maintained when constructing new examination papers every year, otherwise it is not possible to make meaningful comparisons about performance from one examination to another (Greaney & Kellaghan, 1996, p. 35).
The purpose of this report is to provide an overview of the psychometric approaches used to link assessment results of high-stakes examinations across subject areas and over time. Furthermore, it provides an overview across a number of countries that differ in terms of their economic status, education systems, assessment bodies and subjects offered at school level. Three abbreviated case studies (from Europe, Australasia and Asia) are included in this desktop study. In addition three mini illustrations are used to illustrate the range of examination systems.

## 1.1 Standards and standards setting

The discussion on assessment and measurement in examination systems cannot be isolated from that of the debates in standards and standard setting. Therefore, a short discussion of the key issues on this topic is included in this section.

Debates on standards are prevalent internationally, and have been at the heart of the education quality debate in South Africa. Standard setting can be defined as the process by which a standard or cut score is established (Downing & Halaydna, 2006, p. 226). The process is a complex one and involves a panel of selected people who follow a prescribed system of rules to assign a number, which distinguishes between two or more degrees of performance (Cizek, cited in Downing & Halaydna, 2006). The examination process, which is deemed simpler and more straightforward, also involves a measure of professional judgement both in the setting of examinations or in the defining of a pass mark. **Table 1** depicts a generic examination process:

**Table 1: Examination process**

| Activity | Description[1] |
|---|---|
| **Overall plan** | Systematic guidance for all test development activities |
| **Content definition** | Sampling plan for domain, essential source of content-related validity evidence, delineation of construct |
| **Test specifications** | Operational definitions of content, framework for validity defensible sampling of content domain, norm or criterion referenced, desired item characteristics |
| **Item development** | Development of effective questions, formats, validity evidence related to adherence to evidence-based principles, training of item writers, reviewers, effective item editing |
| **Test design and assembly** | Designing and creating test forms, selecting items for specified test forms, pre-testing considerations |
| **Test production** | Publishing activities, security issues, validity issues concerned with quality control |
| **Test administration** | Validity issues concerned with standardisation, security issues, timing issues |
| **Scoring test responses** | Validity issues, quality control, key validation item analysis |
| **Passing scores** | Establishing defensible passing scores, relative vs. absolute, validity issues concerning cut scores, comparability of standards. Maintaining constancy of score scale (equating, linking) |
| **Reporting test results** | Validity issues: accuracy, quality control, timely, meaningful, misuse issues, challenges, retakes |
| **Item banking** | Security issues, usefulness, flexibility, principles for effective item banking |
| **Test technical report** | Systematic, thorough, detailed documentation of validity evidence, 12-step organisation, recommendations |

*Source:* Downing & Halaydna, 2006

[1] The descriptions have been shortened.

Standard setting is viewed as a statistical necessity as well as a procedural one. Most commonly, two standards are prevalent: performance standards and content standards. The former is the cut score or achievement level (passing score), and the latter describes the set of outcomes, objectives or specific instructional goals (Downing & Halaydna, 2006).

Standard setting methods emerged in the 1980s and new approaches were developed in the 1990s in response to calls for polytomous item formats and multiple cut scores. Whilst there are more than 50 methods, these can be clustered into relative (norm-referenced) and absolute (criterion-referenced) methods. There are 10 generic steps in setting performance standards:

1. Select a large and representative group of participants.
2. Choose a standard setting method; prepare training materials and standard setting meeting agenda.
3. Prepare descriptions of the performance categories or referent candidate or group.
4. Train participants to use the standard setting method.
5. Compile item ratings or other judgments from participants and produce descriptive or summary information or other feedback for participants.
6. Facilitate discussion among participants of initial description or summary information.
7. Provide an opportunity for participants to generate revised ratings/judgments, compile information, and repeat steps 5 and 6.
8. Provide for a final opportunity for participants to review information, arrive at final recommended performance standards, and cut scores.
9. Conduct an evaluation of the standard setting process including gathering participants' confidence in the process and resulting performance standards.
10. Assemble documentation of the standard setting process and other evidence as appropriate, bearing on the validity of resulting performance standards.

*Source: Downing & Halaydna, 2006*

The main issues around standard setting seem to revolve around the complexity of the process, changing demands of the education system and society, the human element (subjectivity), standards and accountability, and uniformity and standardisation (Howie, 2008).
In conclusion, the debate about standards is far from over in the international community and the problem of setting standards remains as much a fundamental, unsolved problem today as it was 20 years (Glass 1978; Linn 2000). Nonetheless, standard setting and systematic methods used are crucial to the integrity of the overall system and, therefore, as much as they are imperfect, they are also indispensable (Howie, 2008).

## 1.2 Structure of the report

The report is structured as follows: Section 2 comprises two parts: firstly, we have by way of introduction explained the differences between IRT and Rasch and given an outline of the essential features of the Rasch model. Then follows a summary of the methods for comparing difficulties, which fall into two categories: statistical methods and judgement methods, which are, in some cases, used in conjunction. The statistical methods include, among others, the Subject Pairs Analysis (SPA) currently used in the South African system and latent trait models such as the Rasch model. Judgement methods include, among others, social ratification, a procedure used in the adjustment of continuous assessment marks in the South African context.

Section 3 provides an overview of three examination systems, namely the Key Stage tests in the United Kingdom, the National Assessment of Educational Progress (NAEP) in the United States of America, and, by way of comparison, the examination system in Nigeria. This overview includes a description of different types of assessment, the characteristics of these systems and, where appropriate, descriptions of their methodologies, in particular where IRT and Rasch related methodologies are used. In Section 4 the examination systems in The Netherlands, Western Australia, and Indonesia are described with a specific focus on their moderation and standardisation processes. Western Australia follows an outcomes-based curriculum and has formulated assessment and standardisation strategies for instituting a system that is defensible in terms of ensuring fairness. The Netherlands system is noted for the stability of the system and the considered approach to instituting changes based on careful research. Indonesia has conducted radical reform in the time span of only ten years. This reform process is of interest to developing countries. Section 5 concludes with observations drawn across all three cases, analysing various features of these systems. Finally, the implications arising from the review and analyses are described.

# Section 2:

## Methodologies for Moderating Exit Level Examinations

### 2.1 Overview of item response theory

As the brief was to look into the role of Item Response Theory (IRT) in exit level examinations in the international arena, we provide a brief overview of IRT. This overview of IRT will provide the conceptual background and facilitate access to the information contained within the case studies. Firstly, a brief overview of IRT, with a focus on the related Rasch model, is provided. This overview outlines the core assumptions and requirements of these models and explains the differences. An explanation of the 2-parameter model and the 3-parameter model is given, however, because the mathematical structure is similar; the Rasch model, the one-parameter model, is used to illustrate the theory.

#### 2.1.1 IRT Models

IRT models the relationship between the person's level on the latent trait (or underlying construct) being measured by a test and the person's response to a test item or question (Lord, 1980). A distinction is made between a person's test performance as observed in item responses and their underlying ability or latent trait[2] (Ryan, 1983). This distinction, notably that test performance is not synonymous with ability, arises because most testing situations reflect a person's unobserved ability imperfectly.

IRT makes strong assumptions about a person's behaviour when responding to items and assumes that it is possible to describe mathematically the relationship between a person's trait level and performance on an item (Stocking, 1999, p. 55). This relationship is modelled using probability theory: both item difficulty and person (learner) performance on the items contribute to the model. A simple logistic function is used in the Rasch model to transform the difference between ability level and difficulty level into a probabilistic estimate.

The Rasch model assumes this pattern of responses. The probability is high that a low performing person would only answer the easy questions correctly. The probability is also high that the higher performing persons will be able to answer all the easy questions and in addition, some of the more difficult questions correctly. In the 'middle' area, the probability is that a person answers items correctly only half of the time. The area where the correct answers meet the incorrect answers provides for some unpredictability (Bond & Fox, 2007). When the learner 'ability' is organised hierarchically from highest total score to lowest total score, and the items from highest difficulty level to lowest difficulty level, a pattern, which approximates the Guttmann [3] structure, is formed. That test results approximate this pattern is a requirement of the Rasch model.

This relationship is modelled in Figure 1. The horizontal axis represents the underlying ability or latent trait, the vertical axis represents the probability of a correct response. For example, learners located about the zero point on the horizontal scale would have a 0.5 probability of getting the item correct. Learners located higher (to the right of zero) on the latent trait have a greater probability of getting the item correct, and learners located lower (to the left of zero) have a lesser probability of getting the item correct.

[2] A latent trait or construct is an underlying, unobservable characteristic of an individual that cannot be measured directly, but will explain scores attained on a specific test pertaining to that attribute, and can be measured along a continuum that is context specific.

[3] An explanation of the Guttmann structure is found in all Rasch texts.

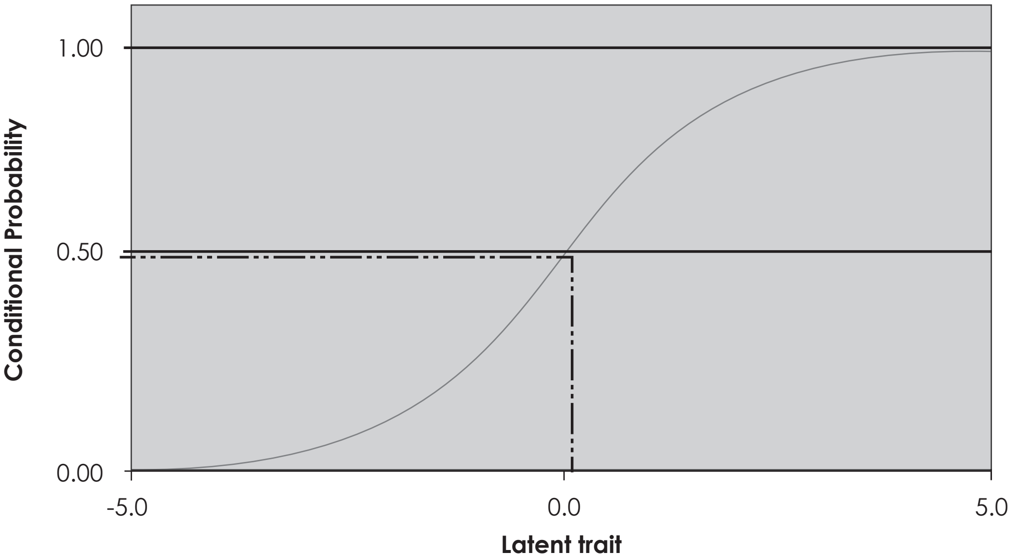

*Figure 1: The mathematical relationship between person ability and item difficulty for a particular item*

## 2.1.2 One-, two- and three- parameter models

Item Response models include the one-, two- and three-parameter models. The one-parameter model considers item difficulty as the only parameter responsible for the probability of a correct response. In the two-parameter model, an item discrimination factor is added to each item. In the three-parameter model, a further parameter, a guessing factor is added.

While the two- and three-parameter models are described as IRT models, the one-parameter model is described as a latent trait theory (LTT) model (Ryan, 1983). IRT and latent trait theory "have different philosophic bases, provide different information when applied to data, and help educators solve different types of problems" (Ryan, 1983:63). The essential difference is that latent trait analysis, also known as the one-parameter model, or Rasch model is concerned with "measuring the underlying or latent trait of an individual", while the two- and three-parameter IRT models are concerned with "describing and summarising observed performance statistically" (Ryan, 1983:63). In essence the Rasch model is concerned with measurement (of an underlying construct), IRT is concerned with modelling. In the Rasch model, anomalies in the results show up problems in defining the construct of interest or latent trait in the test construction and administration, or in the particular learner, that require further investigation. In IRT, a model that accounts for the data may be used.

The remainder of this section will deal with the one-parameter Rasch model, or Latent Trait model, as this will provide the basic information for understanding the philosophy underpinning the model and the requirements. The one-parameter Rasch model, designed initially for dichotomous data, has been extended to include the partial credit model, which has applications to polytomous data (Andrich & Marais, 2008). The Rasch-Andrich Rating Scale, based on Rasch requirements, has applications to Likert type scales and others. Andrich, in a conceptual breakthrough, comprehended that a rating scale, for example, a Likert type scale, could be considered as a series of Rasch dichotomies ((Andrich & Marais, 2008).

The partial credit model applies to achievement items where marks are allocated for partially correct answers or where a sequence of tasks has to be completed. Essentially, the partial credit model is the same as the rating scale model, with the only difference being that in the partial credit model, each item has its own threshold parameters. These models adhere to the same requirements and in essence comply with fundamental measurement (see Rasch model).

## 2.2 Rasch Model

Georg Rasch, a Danish mathematician, developed the Rasch model in the 1950s. Rasch modelling, though made public as early as 1960 through his book, Some probabilistic models for the measurement of attainment and intelligence, and promoted at the time by Ben Wright at the University of Chicago, and others subsequently, is relatively new in the context of public examinations in most countries. The reasons for this are hypothesised in the conclusion to this section.

### 2.2.1 Measurement

Social **measurement** according to Rasch needs to meet the requirements of physical measurement (1960/1980). Bond and Fox (2007) note that great care is taken in the physical world where fundamental measurement is concerned and where years of research go into the design of instruments. Finely calibrated instruments are used to measure, for example, the volume of fuel in a vehicle, the temperature at which a cake should be baked or the levels of certain vitamins in the human body (Bond & Fox, 2007). The same rigour is absent, however, when educational research or a psychological investigation is undertaken (Bond & Fox, 2007).

This requirement for social measurement is not new. In the 1920s, Thurstone (1925, cited in Andrich, 2002) laid down the requirements for measurement in the social sciences as follows:

(1) items should be located on a continuum, or scale;
(2) the locations of items should be invariant across different populations which are to be measured by the items; and
(3) the locations of items on a continuum should satisfy the requirement of additivity" (Andrich 1989, cited in Styles, 1999, p. 25)

### 2.2.2 Invariance

The requirement for invariance can be understood intuitively, in that one would expect that the difficulty value of an item should be intrinsic to the item and not dependent on the persons writing the test if the test was properly targeted at the group. Similarly, a person's ability on the underlying construct should not be dependent on or influenced by the specific test that is written. The Rasch models provide "a theoretical base for invariant comparison within a given frame of reference" (Humphry, 2005, p. 3). The specified frame of reference requires a collection of objects (test items), a collection of agents (testees) and the outcomes of interactions between these. It is important, therefore, for Rasch measurement, that the measuring instrument is appropriately targeted for the given population (Humphry, 2005, p. 3).

### 2.2.3 Interval scales

The Rasch model provides a method for calibrating ordinal data onto an interval scale. The raw scores are ranked according to both person ability and item difficulty. These scores, both person scores and item difficulties are transformed into probabilistic estimates that can be located on the same scale. The ability and difficulty are related by the logits function, the difference between ability and difficulty being equal to the log of the odds. Through the application of this model, raw scores undergo logarithmic transformations that render an interval scale where the intervals are equal, expressed as log odds units or logits. The claim that "it is an interval scale rests on the fact that the same difference between item difficulty and person ability corresponds to the same probability of success" (Coe, cited in Newton, 2007, p. 344). The Rasch process establishes a natural unit for any empirical investigation (Humphry, 2005). It is this feature that allows the process of statistical equating to be done across tests.

### 2.2.4 Undimensionality

Underlying the Rasch model is the notion that it makes sense to measure one construct, as in the physical sciences. This resonates with one of the basic assumptions of the Rasch model that a relatively stable latent trait underlies test results (Boone & Rogan, 2005). This latent trait can be exhibited in more or less quantities in a specified target audience. Unidimensionality is the term used for the measurement of one attribute or dimension at a time (Bond & Fox, 2007). In general terms, we strive for coherence in test design. A test constructor would not be satisfied if in a mathematics test, for example, an item supposed to be testing mathematics ability was answered incorrectly by a top student, when students at the lower end of the scale were getting the item correct. One would check the item for possible flaws and would investigate the item, the learners, or both for explanations of the anomalies.

### 2.2.5 Data fit the model

A distinctive feature for use of the Rasch Measurement Model, namely that the data are required to fit the model, has profound implications for assessment and research. As explained by Andrich (1989, p. 16)

> In the traditional approach, the agenda is to search for models that best account for the data. This (exploration) tends to be carried out by statisticians. In the Rasch approach, the data are required to subscribe to aspects of validity usually required by scientists (Duncan, 1984, p. 398), and in addition, to conform to the chosen model. That (verification) task needs to be carried out by researchers who understand the substance of the variables.

The requirement, therefore, for any social measurement is a clearly defined theoretical construct, which may vary in breadth, for example, it may be the more defined mathematical ability, or it may be the broader construct, the academic ability.

### 2.2.6 Rasch process

The test instrument is designed with the above criteria in mind: that is, with a clearly defined theoretical construct and an appropriate target level for the population being tested. In essence, we define the construct to be measured, create units of measurement that are independent of the construct and the persons being measured, and order persons and items along a continuum. In order to create units that fit the criteria of invariance for a particular frame of reference, we apply Rasch analysis (Rasch, 1960, 1980; Humphry & Andrich, 2008). The frame of reference for any empirical study is constituted by a class of persons who respond to a class of items in a well-defined response context (Humphry & Andrich, 2008, p. 1). The probability of a person achieving success on a particular item is entirely determined by the difference between the difficulty of the item and the person's ability. The score obtained on a test is a function of person ability and item difficulty.

### 2.2.7 Missing data

The Rasch model has no problem handling missing data, since it generates probabilistic estimates. Missing data typically occur either when a person does not respond to an item or, in the case of adaptive testing and matrix designed testing situations, where not all items are administered to all persons. In a practical test situation, where a student of high ability as measured by the test as a whole leaves a difficult item out, an item-person parameter would be estimated from the person's overall test score.

### 2.2.8 Reliability and validity

An important question in the Rasch analysis, as stated previously, is whether the data fit the model. Person and item fit are checked for anomalies. Both overfit, when too much predictability is experienced, or underfit, when the item does not conform as expected, are problematic.

The Rasch analysis is a test of the test, the test constructor and the learner. The outcomes of this process are that the theoretical construct, initially defined, is checked for unidimensionality. Items that do not fit as expected are checked for validity. For example, an item that learners of high ability, as measured by this test, get wrong and learners of low ability get correct, may indicate a problem with the item. The difficulty levels of items for this particular cohort are checked. The expected functioning of learners is investigated and anomalous learners would be investigated.

## 2.3 Summary

In summary, the IRT models provide information about the individual items and the learner responses to those items. The collation of all person-item responses certainly provides richer data with which to assess the validity of the test construct and the reliability of the instrument. The Rasch model provides a "stringent modelling tool" that is useful when the data fit the model and that will provide information about test inequity or differential item functioning when misfit is identified (Ryan & Williams, 2005). Rasch modelling is able to accommodate both dichotomous and polytomous data. Questions that have multiple parts can be analysed using the Rasch model.

While the claim is that Rasch measurement provides improved measures (Andrich & Marais, 2008; Bond & Fox, 2007), the systems using the traditional test theory that are in place do the task more or less adequately, and will therefore remain in place until the need is felt to replace the systems that are "tried and tested". Educational assessment is not alone when it comes to being slow in adopting new technologies.

Another reason for the limited use thus far is that it is thought to be too fine-grained for general examination use, and the methodology is wrongly thought to be too complex for general public accessibility. The use of Rasch measures for public high stakes examination is claimed to provide improved measures (Togonoloni & Andrich, 1996), however, the education of the stakeholders to understand the processes necessarily delays the implementation. This situation is explained in the Western Australia case study. However, in the Western Australian examination system, Rasch methodology has been used by researchers to check the validity of the assessment and certification (Andrich, 2005) at stages in the process. The Rasch model is also used in piloting and designing tests, for estimating student ability and as a basis for making comparisons (Humphry, 2005).

According to Newton (2007), there is only one reported example of an explicitly latent trait approach in England. Coe (cited in Newton et al., 2007) applied the Rasch model (sometimes referred to as the latent trait model) to GCSE data in England in order to investigate grade allocations. Two related studies using the Rasch model, namely Coe (2008a and 2008b), are referenced in the next section. Considerable criticism of the Rasch model by, among others, Goldstein (2007) has no doubt discouraged the use of the Rasch model. The major contention seems to be around the Rasch requirement that the data fit the model, rather than a model designed to explain the data. This philosophical distinction is elaborated by Andrich & Marais (2008), Humphry (2005) Bramley (2007) and Coe (2007).

While the Rasch model and IRT have up until now had very little and very specific applications in high stakes examinations, the Rasch model is entirely appropriate for constructing scales and equating test forms. Because of "the capacity to deal with missing information", and therefore link "tests through common examinees, or sets of examinees through common test items" these models are used extensively (Ryan & Williams, 2005). Coe (2008a, 2008b) uses the Rasch model for comparing the difficulty levels of subjects and the associated grade allocations. The next section examines the different purposes of linking, the types of linking that have been identified, and methods for comparing difficulties. The Rasch model is one of the methods that is used in specified cases, though as yet it is not widely recognised.

## 2.4 Test comparisons

The need for linking between different forms of tests, for making comparisons across different subjects and for maintaining standards across years has provided challenges to education departments and to statisticians over the years. The requirement to compare high stakes examinations is especially important in situations where a change in curriculum introduces new materials, different teaching strategies, and different modes of assessment. The new South African curriculum introduced into the high schools in 2004, culminating in matriculation in 2008, indicates the necessity for reviewing current techniques for moderating examinations.

In this section of the study, we introduce some conceptual issues, explain terminology that is current in the literature and discuss the different methods for comparing difficulty, noting in which countries these methods are used. Finally, we discuss how the Rasch model may contribute to establishing reliable and fair processes.

## 2.4.1 Concepts and terminology

### *2.4.1.1 Scaling, linking, equating*

The term scaling refers to the process of transforming a set of scores while maintaining the meaningfulness of the data. The mean and standard deviation are used in the scaling formulae. This may be used when two tests of differing total score need to be combined.

Linking is about monitoring comparability (Newton, 2007) between different tests of different forms of the same subject or of different subjects. An example of current interest is that of the matriculation examination. We trust the matriculation examination for mathematics and science are of comparable difficulty, and that the mathematics examination from 2007 and from 2008 is of comparable difficulty. We would also like to know that the supplementary papers, prepared for February/March 2009 are equivalent in difficulty to the papers set for the October/November 2008 examinations. Linking these examinations in some way would provide information about the comparability of these examinations. At present, professional judgement is used to match examinations.

The conceptual basis of linking requires that there is a common linking construct. The notion of a common linking construct presupposes that the tests of interest have been set to a curriculum framework and test specifications (that are made explicit). The goal of any linking exercise is to “put scores from two or more tests on the same scale – in some sense” (Kolen & Brennan, as cited in Newton, 2005, p. 423).

### *2.4.1.2 The ‘linking construct’*

Most often “linking” is used to denote situations in which two tests are designed from different frameworks and specifications. The intention is to “calibrate tests built to different frameworks” (Newton, 2007, p. 3). In this case, the linking construct might represent higher order skills/abilities that are shared by both tests. For example, when linking different subjects, such as Mathematics and English, for tertiary entrance purposes, the linking construct may be academic ability. In cases where the tests are designed to different frameworks and specifications, any inferences drawn should be made in terms of the linking construct (Newton, 2005, p. 108). Students at the same level on the scale (derived from two tests) could be said to have in common the same level of attainment in terms of the linking construct.

### *2.4.1.3 Equating and linking*

Equating, according to Newton (2007), is a special case of linking where the construct of interest is identical in each of the tests. This form of linking is between two tests that are designed to the same framework and test specifications that are testing the same construct in the same way. An equating relationship is established (Newton, 2007). The intention is to “calibrate tests built to the same content and statistical frameworks” (Newton, 2007, p. 3), or to adjust scores on equivalent test forms (Béguin, 2000).

According to Béguin, (2000) the classification of equating falls into two groups: methods that use only the observed (manifest) score distributions and methods that are based on latent variables (underlying constructs) such as the true score in classical test theory or proficiency in IRT and Rasch models. Observed-score equating methods can only be applied where randomly equivalent groups exist, whereas methods based on latent variables allow for greater flexibility in linking through the transformation of item difficulties and person abilities into probabilistic estimates.

Mean equating, linear equating and equipercentile equating are explained by Béguin (2000). In the mean equating method, Béguin (2000) explains that the mean scores for each test form are equal to each other and all other scores are subsequently transformed. Scores on the two forms are viewed as equivalent if they are the same number of score points away from their respective means. With linear equating, scores are deemed equivalent when they are an equal number of standard deviations away from their respective means (Béguin, 2000).

Béguin (2000) explains that in equipercentile equating, scores on Form 1 that have the same percentile rank as scores on Form 2 are equivalent. Therefore, for the population of examinees used to equate the test forms, the proportion of examinees below or at any equated score will be the same for either test forms, with the exception of error in sampling examinees (Kolen, 1984). Sampling error can cause large fluctuations in the results of equipercentile equating. Equipercentile equating is used to standardise examination scores in the Western Australian examination system.

#### *2.4.1.4 'Comparable outcomes' vs 'comparable performance'*

A radical change in the curriculum, as has been the case in the South African system, may disadvantage the learner in terms of performance. Factors such as unconfident teachers, new materials, and unfamiliar assessment techniques may cause general lowering of performance. Cresswell (cited in Newton, 2005) distinguishes two perspectives on the comparability of tests: "comparable performances" (where the performance on items deemed to be comparable is judged and linked) and "comparable outcomes" where the marks of comparable students are linked. The choice of which linking method to use depends on the stakeholder. For example, when monitoring schools, performance criteria are required, but in the use of public school examinations to qualify individuals, the requirement is for the outcomes to be linked. Important to note is that the "comparable outcomes" perspective would require that the first cohort of students on the new curriculum should have grades equivalent to the last cohort on the old curriculum.

#### *2.4.1.5 Different types of linking*

Different types of linking are founded on two factors: a conceptual foundation (what the two tests have in common) and methodological rigour (what methodology is most suitable). Linn (1993, p. 85) listed five types of linking, "listed in order of statistical rigour", that incorporate a conceptual and a methodological component.

- Equating – tests measure the same construct in the same way.
- Calibration – tests measure the same construct somewhat differently.
- Statistical moderation – tests do not measure the same construct but scores can be linked using an external measure.
- Prediction – tests do not measure the same construct, but an empirical relationship between scores can be estimated.
- Social moderation – tests do not have anything in common.

Many different methods have been used to investigate the comparability of examinations across years and in different subjects. In South Africa, the Subject Pairs Analysis (SPA) has been used. This method is used widely in the United Kingdom. The Average Marks Scaling method is used in Australia, although it must be noted that the territories use slightly different methods (Tognolini, 2006). The related Kelly method is used in Scotland. Recently, however, the effectiveness of Rasch measurement for linking across year groups or for tracking in longitudinal studies has been observed (Coe, 2007).

These different methods for comparing the difficulty levels of examinations across subjects (where there are some common examinees) and across years (where there may be some common items) fall into two distinct categories, namely statistical methods and judgement methods (Coe, 2007). The "use of common persons to equate assessments of relatively equivalent overall demand" is

sometimes termed horizontal equating; the use of common items across different year groups is sometimes termed vertical equating[4] (Humphry, 2005, p. 130).
A brief summary of the different methods, their assumptions, their limitations and in which countries the particular method is used, follows.

### 2.4.2 Statistical methods

Three of these methods depend on comparisons among the results of the same candidate in different examinations (Coe, 2008). These "common examinee methods" include Subject Pairs Analysis (1), common examinee linear models (2) and latent trait models (3). The other two statistical methods, the common reference test (4), and the 'value added' model (5), depend on "comparing the grades achieved in different examinations with those achieved by others who are judged to be similar on the basis of some additional information" (Coe 2008, p.17).

#### *2.4.2.1 Subject pairs analysis (1)*

According to Coe (2008a) these methods have been widely used by examination boards in England, Wales and Northern Ireland. They are used in these countries to compare grades. The basic method (SPA) considers those candidates that have taken examinations in a pair of subjects. For each candidate the difference between two subjects is compared. The proportion achieving higher grades, the same grade or lower grades is computed to form the basis of a comparison between two subjects (Coe, 2008).

A variation of this method is to compute the average difference between grades, by assigning numbers to grades, that is by converting "examination grades into a numerical scale" (Coe 2008, p. 17). The next step in the process is to "calculate the mean grade differences for all possible pairs of subjects and to average the mean differences, for each subject separately" (Coe, 2008, p. 18). This would mean, for example, that the average differences in grades in mathematics are compared with all the other subjects taken with mathematics. This approach is described as 'Aggregated Subject Pairs Analysis' by Coe (2008). The difference between the SPA and ASPA is that the samples in the simple SPA may not be representative, whereas in the ASPA, the estimate of a subjects' difficulty based on all the candidates taking the subject, is more representative.

#### *2.4.2.2 Common examinee linear models (2)*

These models "compute the relative difficulties of different subjects from a matrix of examination by candidate results", essentially by solving "a set of linear simultaneous equations" (Coe, 2008:18). Variations of these methods have been used in Scotland (Kelly's method) and in Australia (Average Marks Scaling). The Average Marks Scaling method is described in Section 4. According to Coe (2008), these methods overcome the problem of underestimating difficulties of subjects, because candidates who take a relatively difficult subject like mathematics are also likely to take chemistry, an equally difficult subject. For details of these methods see Coe (2008).

#### *2.4.2.3 Latent trait models (3).*

In a very recent publication, Coe (2008) states that there has been limited use of the Rasch model, the main method of this type. The only recorded use by examination boards is that of the Tasmanian Qualifications Authority (Coe, 2008). Work by Tognolini and Andrich (1996) proposes how the Rasch model may be used in Western Australia. This has not been taken up in Western Australia at the time of writing.
The Rasch model is similar to the common examinee linear models, but differs in that the difficulty estimates for each grade can be found independently. For details of this process see Coe (2008). Another difference is that the Rasch model requires both the subjects and the candidates to fit the model, the assumption underlying the model being that a unidimensional latent trait underpins the performance. Where a subject or candidate does not fit the model, that is, does not conform to expectations, the model will identify the 'misfits'. These anomalies can then be checked.

[4] This process has been used in Western Australia to monitor and improve systemic assessments.

#### *2.4.2.4 Common (reference) test method (4)*

The rationale behind the common test or reference test method is that if pupils of the same ability are considered, similar examination results are expected. Pupil ability is estimated through a reference test that could be a totally different test or be estimated through a common section. Regression lines are drawn to illustrate the relationship between the pupil ability and the examination results. A reference test was formerly used in Western Australia: This has been superseded by the Average Marks Scaling method. The Netherlands also uses reference testing for certain subjects. The advantage of the reference test is that "there is no requirement for the examinations being compared to have any common candidates" (Coe, 2008, p.23).

#### *2.4.2.5 'Value-added' models (5)*

According to Coe, "value-added analyses have been widely used by awarding bodies in the UK" (2008, p.23). In the general method,  "the regression model can include explanatory variables that help to explain variation in examination performance, such as candidate's prior attainment, gender, socio-economic status, type of school attended, etc" (Coe, 2008, p.23). Multivariate regression models enable one to investigate the combined influence of various different variables. Newton (2007, p. 12) states that: "If all of the 'input' variables are measured adequately then it should be possible to predict the 'outcome' measure – the examination result – with confidence." A problem, though, with this method, is that all the previously uncontrolled-for variables also have to be measured, otherwise the analysis can still be legitimately challenged.

### 2.4.3 Judgement Methods

The argument against purely statistical approaches is that educational content can be ignored. Judgement methods rely on the decisions of expert examiners to judge standards based on their experience (Coe, 2008). Judgement methods can be categorised as judgement against an explicit 'standard', and judgement against other scripts.

#### *2.4.3.1 Ratification method*

Newton (2007) makes the point that historically judgemental methods have been used, in particular, the ratification method. This method entails having different subject experts or experienced examiners in a subject field ratify scripts. Ratification is the agreement that a certain script is of a certain appropriate standard. Repudiation, on the other hand, means that the examiner views the script as either of a higher or lower standard. Frequencies of the judgements are then considered. Value judgement theory of linking is divided into a series of claims (Newton, 2005):

- Technically impossible, therefore no formal meaning
- Social obligation to make comparisons even when linking cannot be achieved technically
- Comparability can be achieved by proclamation – it must respect multiple uses.
- Comparability "necessitates a procedure through which empowered arbiters make non-specific judgements of equivalent value (where the perceived value of an examination performance is relative to each arbiter's personal take on comparability)" (Newton, 2005, p. 117).

Social moderation, similar to the Ratification method, is a process of discussion and debate, the purpose of which is to negotiate shared understanding, i.e. consensus moderation (Linn 1993). Social moderation includes a family of methods for linking standards judgementally – where teachers' judgements are brought into line. Maxwell (cited in Newton, 2005) identifies two types of social moderation, panel moderation, which is bureaucratically imposed, and peer moderation, which is socially negotiated. In many public examination systems social moderation is the primary means of establishing a linking relationship.

#### *2.4.3.2 Paired comparison method*

The paired comparison method requires judges, rather than judging one script, to judge the relative quality of two scripts. The judges then order the two scripts according to relative worth or quality. By applying the Rasch model, the 'judged difficulty' of each script can be estimated. Scripts at the point where it is most difficult to decide which grade they fall into, that is, where there is a 50-50 chance of falling into a B-grade or a C-grade, mark the threshold line.

### 2.4.4 Summary

**Rasch models** are based on the underlying assumption of measurement that requires a unit of measurement that is independent of item difficulty and person ability. In test equating, it is necessary to equate units of measurement across two or more tests. This process has in the past mostly been done through professional judgement. What the Rasch model does is enable the units of measurement in both tests to be scaled to a new measure, or for both sets of items to be put on the same scale. This new measure, in the case of tests of different underlying constructs such as Science and English, is then measuring a more general construct 'academic ability'.

When embarking on any linking process, there are conceptual, methodological and practical considerations. From the conceptual consideration, the linking construct is the organising principle that has to be considered. Linking can only be achieved if a plausible linking construct can be defined (even if only roughly and loosely defended) (Newton, 2005). For instance, in the Western Australian case study, the linking construct which enables putting different subjects, for example English and mathematics, on the same scale, is the plausible, and possibly defendable construct, "academic ability" (Partis, 1977). The only inferences that can be drawn about position on the scale are in terms of the linking construct, "academic ability". For tertiary institutions and other stakeholders, the single ranking, the Tertiary Entrance Ranking (TER), represents academic ability. According to Newton (2005), the linking construct must be made explicit in order to be fair.

This proposed rigour in measurement, advocated by Bond & Fox (2007), Andrich (2005) and others, should be extended to the field of education in South Africa. The Rasch model provides an avenue to attain this goal.

The trend, however, has been to use IRT in systemic assessments, where the focus is on monitoring education processes, rather than in public examinations, where the turnaround time between writing the examination and publishing results is short. The next section illustrates how IRT or Rasch has been used in different types of assessment, namely the National Assessment of Educational Progress (NAEP) and the Key Stage examinations.

The use of Rasch modelling in test construction and the equating of different forms of tests that test the same construct, and similar tests that test different subject matter, has been pioneered in the Western Australian examination system. This modelling strategy enables post hoc reflection on aspects of the examination system, as will be seen in the Western Australian case study.

Ryan and Williams (2005, p. 11) express the view that the Rasch model provides a "stringent modelling tool", which is useful when the data fit the model and which will provide information about test inequity or differential item functioning when misfit is identified. In addition, the Rasch model is appropriate for constructing a scale because of "the capacity to deal with missing information". It is therefore capable of linking "tests through common examinees, or sets of examinees through common test items" (Ryan & Williams, 2005, p. 14).

In a recent study entitled Relative difficulty of examinations in different subjects (Coe et al., 2008A), a comparison was conducted using each of the five statistical methods described in Section 2.4.2. Their overall conclusion was that there was "a reasonably high level of agreement across the methods for the A-level data" (Coe 2008, p. 98). For the GCSE data, there was close agreement between the methods, with the exception of the Rasch method, which provided more detailed information on some aspects of the scale. In a subsequent paper (Coe, 2008B), the Rasch model is used to make comparisons across the GCSE subjects. Coe prefaces his discussion with the statement that the Rasch model is validated on theoretical grounds, and that, though the model provides a "convenient and theoretically sound way of establishing comparability of different examinations based on a unidimensional latent trait", it is not widely used in the UK (Coe, 2008b, p.20). The fact that it is also not used widely in Australia[5] , despite theoretical motivations by Tognolini and Andrich (1995) providing the conceptual basis for doing so, is an indication of the necessity for public support and trust being in place before "new" methodologies are implemented in the field of high stakes testing.

[5] The Rasch model is used  for scaling in Tasmania and in Cyprus (Robert Coe, in conversation 23/10/08).

# Section 3:

## Examination Systems

Examinations by their very nature are high stakes endeavours as success or failure may bring serious consequences. Curriculum and teaching tend to revolve around the examinations, the preparation of which requires considerable effort in preparation by both learners and teachers. Partly as a consequence, it has been observed in many examination contexts that potential low scorers may be prevented from taking the examination in order to boost the school's overall performance (Madaus & Greaney, 1985). There are also problems of test corruption and test score pollution in places. Perhaps this is inevitable when examinations determine the pathways into further education and the workplace. Furthermore, from the schools' perspective, if the results are used to rank school districts or schools, the tests will be perceived by schools as an important indicator of what is to be valued in education (Madaus & Kellaghan, 1992). Essentially what will be taught is that which is being examined, and what is not examined will not be taught.

The first public examination took place in China more than 2 000 years ago with the purpose of selecting citizens for positions in the civil service. However, public examination systems in schools have a shorter history, being introduced as a graduation examination in 1788 in France (World Bank, 2001a). Examinations are important for the individual candidates and their families as the results could determine future educational and life choices (Bishop, 1998) as well as for the schools and teachers as their reputation may be affected by the examination performance (World Bank, 2001b).

**Table 2: Functions of examinations**

| Function | Description |
|---|---|
| **Selection** | To select individuals to the next level of education, especially in a situation where there are only a number of places available. |
| **Certification** | To provide evidence that candidates have reached a certain level of achievement. The certificates issued may then be used for employment purposes. |
| **Control** | By means of exercising control over examinations, the curriculum can be tailored to national goals and objectives. |
| **Motivation** | Motivation implies that clear goals are provided for which individuals can strive. This provides a sense of purpose as well as tangible incentives. |
| **Monitoring** | Monitoring here refers to the gauging of educational standards as well as judging the effectiveness of schooling. |

*Source:* Eckstein & Noah, 1989; World Bank, 2001b.

**Table 3: Characteristics of well-functioning examination systems**

| Characteristic | Description | Indicators |
|---|---|---|
| **Fitness for purpose** | The examination papers and the marking system provide scores that are both reliable and valid. | • Acceptance that the examinations are according to the curriculum.<br>• Statistical evidence of technical quality (for example, the reliability or level of difficulty overall and on an item level)<br>• Adequate quality control measures. |
| **Equity, integrity and public confidence** | The conduct of the public examination system is fair and acceptable to the public. | • The public has confidence in the results of the examination system.<br>• High level of trust of examination agency staff<br>• High level of trust in the supervisory staff.<br>• Little evidence of cheating<br>• Examination authority has procedures for rechecking of marks and has an appeals procedure in place.<br>• Special support for disadvantaged candidates is available.<br>• The question papers used do not contain culturally inappropriate questions, or questions in a language with which some students are relatively unfamiliar.<br>• The grading system is applied equally to all. |
| **Efficiency and cost-effectiveness** | The examinations authority should deliver the required services making the best possible use of resources. Public examinations are administered according to schedules and results are issued on time. | • Examination fees, if applicable, do not place an excessive burden on parents.<br>• Examination authority can demonstrate the cost-effectiveness through its accounting procedures.<br>• Efficient staff are available.<br>• Examination papers are printed in the most cost-effective way without compromising security.<br>• Results are issued on time and in an appropriate form for decision making.<br>• Feedback on examination performance is given to schools to influence instruction practices. |
| **Transparency** | The examination process should, as far as possible, be open to public scrutiny | • Non-confidential materials such as regulations, curriculum and sample/past examination papers are widely available.<br>• Reports, including statistical data, on examination performance are available.<br>• Marking system and criteria for award of grades are available.<br>• Examination authority maintains records of administrative practices, results and marking schemes. |
| **Beneficial effect on classroom practice** | The public examination system should promote good teaching and learning practices. | • Examinations encourage the development of higher-order thinking skills, avoids rote learning.<br>• High quality reports for teachers and other interested parties (e.g. textbook boards) are distributed regularly. |

Examination systems serve a variety of functions and an overview of these is presented in **Table 3**. However, very often within an examination system, several of these functions may be identified. For example, in South Africa, the Grade 12 examinations serve a certification purpose but at the same time, a selection purpose for higher education. Likewise, in Indonesia, especially at the lower secondary exit level, examinations serve a certification and selection function.

The importance of examinations cannot be understated as the results very often affect a large group of students and have real consequences (Bishop, 1998). For this very reason, it is important to identify what the ideal system should look like and what the characteristics of such a system should be. The World Bank (2001a) provides an overview.

The purpose of this section was to present agreed characteristics of well-functioning examination systems, and to describe three systems that have different purposes from three different continents to broaden our understanding of what is possible in different contexts. In two of the systems, not public examinations, IRT, specifically the Rasch model, is used. The aim of presenting the three examination systems (Section 3.1 - 3.3) is to provide additional information. The three case studies, however, presented in Section 4, are discussed in more depth and detail. These illustrations provide the reader with a sense of how IRT can be utilised by providing a brief description of the system.

## 3.1 Examinations in nigeria

In Nigeria, at the end of secondary school, candidates are expected to sit for a number of examinations such as:

- The West African Senior School Certificate Examination (WASSCE), conducted by the West African Examinations Council (WAEC)
- The Senior School Certificate Examination (SSCE), conducted by the National Examinations Council (NECO)
- The National Technical and Business Certificate Examinations (NTCE/NBCE) conducted by the National Business and Technical Examination Board (NABTEB).

Candidates need a minimum of five credit passes in any of the examinations in order to be able to sit for the University Matriculation Examination (UME), which is conducted by the Joint Admissions and Matriculation Board (Obioma & Salau, 2007).

Initially, the West African Examinations Council (WAEC) was responsible for the administration and management of British examinations in West Africa. According to Afemikhe (2007, p. 6) there were problems "such as public outcry about confidence in examinations conducted due mainly to examination misconduct and irregularities, long delay in release of results and certificates and unwarranted seizure of results" Due to the difficulty experienced, the National Board for Educational Measurement (NBEM) and the National Business and Technical Examinations Board (NABTEB) were established. In 1990 the National Examinations Council took over the SSCE. The aim of the examinations is to ascertain levels of proficiency as well as to provide certification (Afemikhe, 2007; Obioma & Salau, 2007).

One of the major concerns in Nigeria over examinations is cheating (Afemikhe, 2005; Afemikhe, 2007). The Joint Admissions and Matriculation Board has attempted to minimise cheating in its examinations by using different versions of the examination in the subjects tested. According to Afemikhe (2005, 2007) scaling and equating have been used to provide equivalent versions.

## 3.2 Key stage 3 in the united kingdom

Key Stage 3 (KS3) is a national examination in the United Kingdom. KS3 aims to raise standards by strengthening teaching and learning across the curriculum for all 11–14 year olds (Crown, 2008a). KS3 is an important part of the United Kingdom's agenda for transforming education (Crown, 2008b). The following subjects are included in Key Stage 3 as part of the National curriculum (Qualification and Certification Authority, 2008):

- **English**
- **Mathematics**
- **Science**
- Information and Communication Technology (ICT)
- Design Technology
- History
- Geography

- Modern Foreign Language
- Art and Design
- Music
- Physical Education
- Citizenship
- Sex Education
- Careers Education
- Religious Education

However, only the first three subjects are assessed through formal national examinations, with the assessment of ICT being proposed. The remaining subjects are taught and assessed by teachers. As part of the Key Stage 3 conceptualisation, various cut-off scores are used and reported. In order to ascertain these cut-off scores, various equating exercises are undertaken. Bramley (2006) refers to four different sources of evidence:

1. Statistical equating on pre-test data
2. Judgement exercises using practicing teachers
3. Scrutiny of scripts by senior markers
4. Impact data using raw scores of approximately 20 000 learners

The point of the equating exercise, in itself, is a standards maintenance exercise and not a standard setting exercise (Bramley, 2006). Rasch modelling is used and deemed appropriate as "the appropriate statistical tool for equating, since it yields estimates of person ability and item difficulty on an equal-interval logit scale" (Bramley, 2006, p. 6). The equating method currently being used is termed a "parallel anchor test" where two different tests both provide items that are combined into a third test[6] (Bramley, 2006, p. 10). Within the assessment design and, indeed, the pre-test cohort, a third of the learners take version A, one third take version B and one third take the parallel anchor test. The cut-scores on the parallel anchor test are known and by means of simple linear equating or equipercentile equating to the anchor test, the cut scores of version A and B are calculated.

## 3.3 National assessment of educational progress (NAEP) in the United States

The National Assessment of Educational Progress (NAEP) differs from the Nigerian and English system in that it is national assessment and not an examination system per se. However, it is also one of the most significant systems to have been developed and utilises the state of the art test theories and methods. Designs and methods developed for this project have served as models for national assessment all over the world. NAEP has served as an example for international comparative studies, while those who are involved with the development of this national assessment consult on examination systems across the world. NAEP is a congressionally mandated project of the National Center for Education Statistics (NCES) within the Institute of Education Sciences of the United States Department of Education. The project itself is undertaken by the Education Testing Service based in Princeton employing more than 3 000 specialists and researchers in evaluation and assessment, the largest of its kind in the world. The Commissioner of Education Statistics is responsible for carrying out the NAEP project, while the National Assessment Governing Board oversees and sets policy for NAEP (National Assessment of Educational Progress, 2007). NAEP collects information on a national as well as state level in subject areas such as mathematics and reading (Thomas, 2000). Matrix sampling is used to monitor educational performance. The primary goal of NAEP is to gather information about the degree to which educational goals are being met (Power & Wood, 1984).

The assessments are designed by committees of learning area specialists, teachers and concerned

[6] The 'parallel anchor' test is a form of common reference test (See Section 2.4.2.4).

citizens. The committee specifies the objectives of the assessment and these are defined in terms of content-by-process matrices. Tasks for each of the objectives are designed to cover a range of difficulty levels (Beaton & Johnson, 1992). Post assessment, IRT is used to summarise and report on findings (Beaton & Johnson, 1992; Thomas, 2000). Developmental scales are identified and scale anchoring is used to interpret the scales. Several anchor points are identified. The anchor points are set far apart to reflect performance differences.

In a second method of anchoring, a three parameter logistic model is used to scale the scores (Beaton & Johnson, 1992). The benefit, according to Mislevy, Johnson & Muraki (1992, p. 131), of making use of scaling, is that "the performance of a sample of students in a subject area or a subarea can be summarised on a single scale even when different students have been administered different exercises.

## 3.4 Summary

The different educational historical and current contexts, the type of examination system, the time factor and the amount of financial and human resources determine the sophistication of the methodology. In the section to follow, three case studies are discussed in depth to shed light on contextual information and the utilisation of statistical methods, including IRT and the Rasch model, and judgement methods.

# Section 4:

## Case Studies

In this section, three systems are presented, with a special emphasis on the process of standardisation in each examination system. The examination systems in The Netherlands, Indonesia and Western Australia are drawn from different continents in an attempt to reflect a variety of contexts and educational systems. Both The Netherlands and Western Australia have a long history of utilising IRT in educational research whilst Indonesia is a latecomer to this area, but has been investing in this field for more than a decade. Each of the systems is described in terms of the design of the education systems before focusing on the examination systems and, in particular, the technical details related to the use of IRT. These are finally discussed in terms of the impact of this utilisation on the examination system. Some aspects are reported in more detail in one case study than in another, as this is dependent on the information available[7] .

## 4.1 The netherlands

The Netherlands has a population of 16.36 million people (as of January 2007). The gross domestic product (GDP) was €529.1 billion by the end of 2006. The total expenditure on education, student loans and research constituted 5% of the GDP and 19.2% of the gross central government expenditure as expressed in Van der Ree (2007). At the end of 2006, 909 500 pupils were enrolled in Ministry of Education-funded secondary education, pupils with special needs included (De Wit, Van de Ven & Van der Mijl, 2007).

### 4.1.1 Schooling in the Netherlands

In this section an overview is given of the organisation of the education system in The Netherlands with specific reference to the position of secondary education streams within the system as well as an indication of the subjects offered in the central examination as discussed later in the document.

***Secondary education***

Three types of secondary education opportunities exist in The Netherlands and are geared to meet the needs and background of the pupil. All three types of education are aimed at children aged 12 and older and all start off with a basic secondary education. There are approximately 700 secondary schools in The Netherlands and they consist of public, (special) religious and private schools, all of which receive public funds.

**Table 4** provides a summary of the three types of education, the duration of each and the age range targeted.

**Table 4: Types of education**

| Type of education | Description | Extent | Age range |
|---|---|---|---|
| **VMBO** | Preparatory middle-level vocational education | 4 years | 12-16 years |
| **HAVO** | Higher general continued education | 5 years | 12-17 years |
| **VWO** | Preparatory scientific education | 6 years | 12-18 years |

**Pupils with special attention needs can enter practical education (PRO) or special education (VSO).*

[7] It is important to note that, with regard to test design and methods, the literature is sometimes not explicit with regard to either. Where possible, deductions have been extrapolated from the information that is given. Where this is the case, it is so noted in the text.

The first two years of the secondary education are essentially the same for all pupils, with the emphasis moving away from mere knowledge to independent learning and teaching skills (Alberts, 2001). After this there are a number of routes through secondary education. These are fairly flexible allowing for movement across tracks.

At the age of sixteen, pupils from VMBO generally choose to enter the middle-level vocational education (MBO) area or they could also enter HAVO, which is preparation for higher vocational education. HAVO-graduates can also enter VWO or MBO. VWO is a preparation for scientific education (WO). Until 2006, VMBO played out in two different systems, i.e. VBO (pre-vocational education) and MAVO (junior general secondary education) as discussed in Alberts (2001) and The Netherlands Ministry of Education, Culture and Science (2007).

***Subjects offered in Secondary education***

A wide variety of general subjects are offered at the VMBO secondary level (see **Table 5** for details). **Table 6** presents a list of all the vocational subjects offered at secondary level.

**Table 5: General subjects in VMBO Central examination**

| | | | |
|---|---|---|---|
| Dutch | Turkish | Mathematics | Drawing |
| Frise | Arabic | Physics and Chemistry 1 | 'Handenarbeid' |
| French | Economy | Physics and Chemistry C | Textiles |
| German | Economy Compex (C) | Physics and Chemistry 2 | Audiovisual studies |
| German (Digital) | Company law | Biology | Music |
| English | Geography | Biology (C) | Dance |
| Spanish | History | Visual arts | Drama |

*(Source: Alberts, 2007)*

**Table 6: Vocational subjects at secondary level**

| | | |
|---|---|---|
| Building techniques | Caring | Green issues |
| Electro technique | External caring | Flower arranging |
| Graphics media | Care and Welfare – General | Agriculture |
| Installation technique | Fashion and Commerce | ICT |
| 'Instalelektro' | Consumption – Bakery | Sport service delivery and safety |
| 'Metalelektro' | Consumption – HORECA (= Hotels, Restaurants & Cafés) | Technology and service delivery |
| Metal technique | Consumption - General | Technology and commerce |
| Transport and Logistics | Plant cultivation | Service delivery and commerce |
| Vehicle technique | Animal breeding: Production | |
| Administration | Animal breeding: Pets | |

*(Source: Alberts, 2007)*

***The Netherlands education system***

**Figure 2** provides a schematic presentation of how the secondary school system fits into the schooling system in The Netherlands.

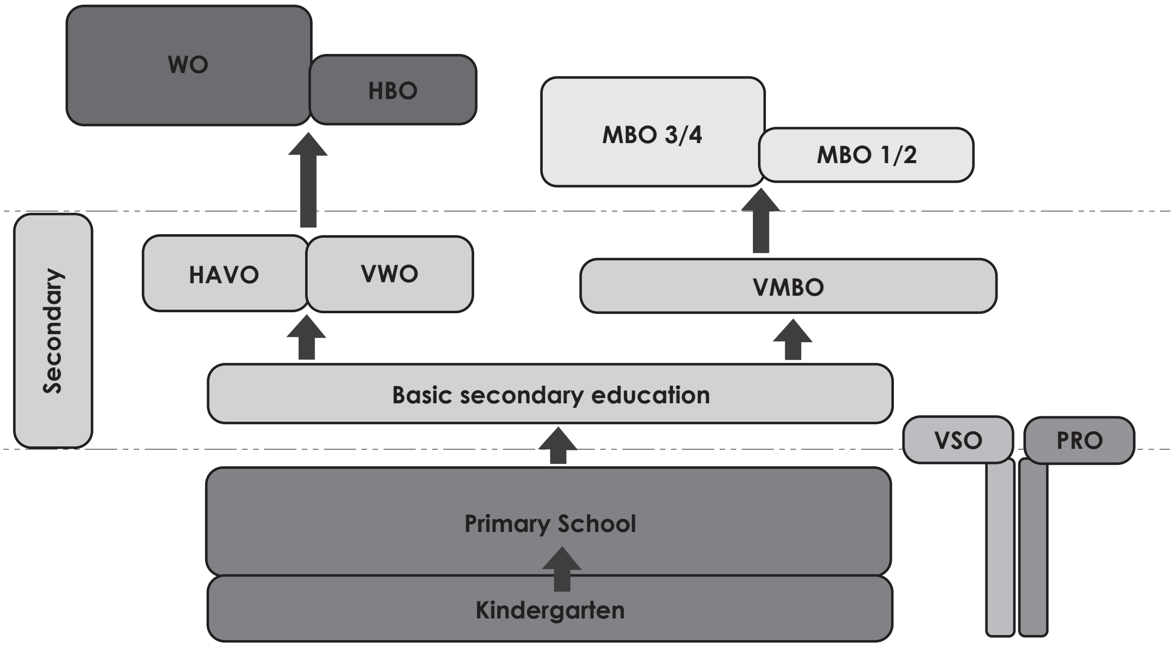

*Figure 2: Diagrammatic presentation of the schooling system in The Netherlands*

*(Source: Adapted from De Wit et al., 2007)*

## 4.1.2 The examination system

As in most countries, the national examinations in The Netherlands serve a dual role – as school exit examination as well as being a portal into tertiary education. Maintaining standards are thus crucial for the tertiary education sector to maintain confidence in the performance levels of their student intake. Alberts (2001) explains that, The Netherlands' parliament was concerned that the fact that more pupils entered higher education institutions could be an indication of dropping standards.

Prior to 1990, a cut-off score was mainly used as a standard setting device. A cut-off score is a score that distinguishes between pass and fail. The cut-off score was adjusted each year to render a constant percentage of pupils that passed each year but, according to Alberts (2001), this procedure does not guarantee equivalence of examinations.

## 4.1.3 Assessment bodies

The Minister of Education, Culture and Science appoints the Central Committee for Ratification of Examination (CEVO) directly, but the minister is ultimately responsible for the examinations. The actual execution of the Minister's responsibility is delegated to CEVO. Syllabus specifications provide the specifications for the content of the examinations and the National Institute for Educational Measurement (Cito) prepares the examinations. CEVO checks and ratifies the examinations and is legally responsible for the contents thereof.

The national school-leaving examination consists of two parts. The first part is the school internal examination set and scored by the schools. The second part or national examination is commissioned by CEVO to Cito for the construction of the examinations (Alberts, 2001). The national examinations are the same for all pupils attending the same type of school. The mean for the two examinations serves as the final mark for the school-leaving examinations.
Generally, the procedures for the central examinations entailed the following (Alberts, 2001):

- Every year the construction cycle for each examination started from scratch
- The cut-off score was set again every year
- CEVO preferred test questions to be screened, that is ratified by professionals rather than being pre-tested.

According to Alberts (2001), schools are responsible for marking the examination papers after the pupils have written, but they do not grade the papers as ‘pass’ or ‘fail’. Cito then collects all the pupil data, does the data analyses and sends the results on to CEVO. CEVO sets the cut-off score which indicates the transition or borderline between pass and fail in such a way that the same percentage of pupils pass the examinations every year, as mentioned previously. This method does not guarantee a consistency of standards over time.

### 4.1.4 Standardisation

Béguin, Alberts and Kremers (2008) make the all-important point that pupils writing school end exam-inations should be exposed to the same requirements for passing the examinations year by year. This has the implication that the examinations should be content equivalent and that one examination should not be easier or more difficult than another examination.

To ensure content equivalence, CEVO, according to Béguin et al. (2008), prescribes a so-called examination model that contains the following information:

- Duration of the examination and number of tasks
- Type of tasks: type of questions (open and closed questions), cognitive domains (knowledge, insight, skills), etc.
- Topics/domains
- Test matrix: distribution on topics/domains
- Authorised devices: atlas, dictionary or calculator.

The same examination model is used every year and this practice ensures that the examination contents are, as far as possible, comparable to each other. Teams of lecturers with experience in the setting of examinations are involved.

Three different procedures for maintaining standards are used according to Béguin et al. (2008):

- The **standard** procedure entails that, in most subjects, the section scores for approximately 2 000 pupils are collected and calculated.
- Standards comparison **after** the examinations are conducted in the subjects French, German and English, primarily examinations that consist mostly of closed-type questions, are executed supplementary to the standard procedure.
- Standards comparison **before** the examinations in subjects where the papers consist mainly of open-ended questions are also executed supplementary to the standard procedure.

Béguin et al. (2008) explain that the application of the latter two procedures is not always possible in cases where topics change a great deal over the different years. In subjects where very few cand-idates are involved, these two procedures will also not be applicable. The rationale for each procedure is discussed in the following paragraphs.

#### *4.1.4.1 The standard procedure*

In examinations where a large number of pupils write the examination, the assumption is made that the performance level of a group of pupils in one year would not differ from that of a group of pupils of another year. This assumption leads to the fact that the average symbol (grade)[8] for the two years will be the same. One way to adjust the standard through this procedure is to keep

[8] The explanation on how this statistic is calculated follows in the paragraph on Grading following in this section.

the percentage of 'fails' the same. In most cases, where a large number of pupils are participating in an examination, this method is, according to Béguin et al., (2008), a powerful and justifiable method that is simple and cost-effective to execute.

An objection to the standard procedure is the fact that the proficiency of pupils may differ from year to year is not taken into account. It could therefore make a difference to a pupil's result in which year the examination was written, because if all other pupils performed better during a certain year, a specific candidate could have obtained a lower symbol (grade) than if all other pupils performed worse.

#### *4.1.4.2 Standards comparison after the examination*

In a few subjects, as previously mentioned, the assumption is that the proficiency levels of pupils can differ from year to year. The purpose of the standardisation of examinations, according to Béguin et al., (2008) is to be able to relate the same performance to the same symbol (grade) over the years. An explicit distinction should be made between the difficulty of the test and the proficiency of the pupils in the case of differing year on year proficiency levels. To do the standardisation, additional information is collected and complex statistical models are applied. The simplest way would be to have the same cohort of students take both examinations, because if the examination conditions are the same, the difference in results should only be attributed to a difference in difficulty on the examination.

For several reasons this approach is not practical to follow:

- The questions of the previous examinations are known.
- It is not plausible to include questions that would not contribute to a pupil's end result.
- The new examination cannot, for reasons of confidentiality, be administered before the actual examination date.
- The total time to administer two examinations makes it impossible.

A procedure was subsequently developed whereby the same group of pupils did not have to take both the examinations (Béguin et al., 2008). Additional data are collected from pupils that did not take part in the examination. They get an examination paper that partly consists of some of the questions from one examination and partly of some questions from the other examination. With the help of these results, the relative performance on both examinations for an equivalent group of pupils can be estimated. This method falls into the 'common reference method' category.

Cito's point of departure is that the results of pupils depend on two factors: the ability or proficiency of the pupils (pupil ability) as well as the difficulty of the questions or item (item difficulty). The model used by Cito predicts the results that pupils will attain on the questions and, through an iterative estimation process, it is determined which parameter values fit the model the best. The model is then used to estimate or predict results on items that the pupils did not answer. The application of this technique makes it possible to predict what the results would have been if the same pupils took both examinations and, consequently, which examination was more difficult. In summary, this procedure contains basically three steps:

- Estimation of the model parameters
- Prediction of the results on questions or items that the pupils did not answer
- Judgement about which examination was the more difficult.

Additional data are collected from pupils that are deemed to be on the same proficiency level as the group they are compared to, but that do not partake in the examination. An example is that information can be collected about the VMBO examinations through administering the examination to three-year VWO students. Such a test contains old examination questions or anchor questions (see section on Linking) for which the statistical properties are known.

A reference examination is set for each subject. This reference examination should be a realistic representation of what pupils should know and it should have an appropriate difficulty value. The administration of this reference examination takes place immediately after the new examination is administered and standardisation is done to ensure that the same performance requirements are adhered to as in earlier examinations.

#### 4.1.4.3 Standards comparison before the examination

In the case of subjects with mainly open-ended questions, standards comparisons occur according to all the principles discussed in the section above on Standards comparison after the examination, but the additional information or data are gathered **before** the new examination is taken, the reason being that it takes longer to score open-ended questions than multiple-choice questions. In other words, even before a new examination is taken, the standards authorities attempt to ascertain that the difficulty level is the same as the reference examination. With open-ended questions, the probability of pupils scoring full marks on a question, as well as the probability of pupils scoring partial credits, is calculated, as explained in Béguin et al. (2008).

Béguin (2000:1) gives the reason for using the Rasch measurement model in the analyses of the reference examinations as follows:

> "The reason for this choice is that the test equating procedure using the Rasch model is both very fast and numerically robust. The properties are the more valuable since all computations for all examinations in the procedure must be carried out in a very short time."

#### *4.1.4.4 Grading*

In The Netherlands, grades from 1 (the lowest) to 10 (the highest) are given. After the scores have been awarded, a score on an examination is converted to a grade through the formula:

$$\text{Grade} = 9 \times \frac{S}{L} + N$$

where S is the obtained score, L is the maximum possible score and N is the standardisation term. The standardisation term, N, is determined by CEVO after the examination and has a value of between and including 0.0 and 2.0. An examination paper of average difficulty would have N=1.0. For a relatively difficult examination, the value of N will be 2.0 and for a relatively easy examination, the standardisation or adjustment term would be 0.0. Generally, a grade of 5.5 and higher will result in a 'pass' and a grade of 5.4 or lower constitutes a 'fail'.

### 4.1.5 Implications for South Africa

The existing Netherlands examination system functions well in a stable broader educational setting within the country. Changes to the system are only implemented after a process of thorough research.

## 4.2 Western Australia

The size of the Australian population is 20.2 million (OECD, 2004). The Gross Domestic Product is AUD 857 765 million (2004). The per capita expense for secondary school education is AUD 7923 (OECD, 2004). In Western Australia there are 280 secondary schools (Curriculum Council, 2005). Each of the Australian states (certainly until 2007/2008), have autonomous independently functioning education and examination systems. This may change in the future, as the task of aligning the results from the different states for tertiary entrance purposes proves to be very difficult. Queensland, for example, has an entirely school based examination system.

Western Australia as a case study was considered, as the education system is similar to the South African system in critical ways. Both countries claim to have instituted an outcomes-based system of education, and there are similar processes in the design of the examination system. Where Western Australia differs is that they have had a relatively stable education system for a long time, changes are instituted after due research and they have a strong tradition of psychometric research.

### 4.2.1 Western Australian schooling system

In Western Australia there is compulsory schooling from Years K to 10. Senior Secondary, Years 11 and 12, offers 21 subjects (including vocational subjects), which are school assessed. In addition, there are Year 12 Tertiary Entrance Examination (TEE) subjects. For each of these subjects, schools award a grade. Achievement of a grade counts towards the completion requirements for the Western Australian Certificate of Education (WACE). These grades are moderated by the Curriculum Council to ensure comparability.

Western Australia follows an outcomes-based curriculum (OBE). The school curriculum is structured into 8 learning areas. The outcomes (from two to four for each subject) are stipulated for these learning areas and eight levels of achievement are specified over the 12 years (Andrich, 2005). The courses, for example, English for Years 11 and 12, are organised into six units (1A, 1B, 2A, 2B, 3A, 3B), which target sequentially higher levels. Some students do two courses; most do four courses; "mainstream students" in a course will do the four highest levels (Andrich, 2005). The assessment requirements for each course have four stages of aggregation (see Andrich 2005). The requirement in the design and teaching of these courses is that the levels across outcomes in a particular course, for example, the reading outcome and the writing outcome, are equivalent and that the levels across courses, for example, mathematics and English, are equivalent (Andrich, 2005).

The Western Australian system allows students to choose any combination of TEE subjects and WACE courses. The final tertiary entrance ranking is made up of the combined score, including both a Year-12 school mark, (made up of semester examinations, class tests, class work, research assignments and practical work) and an external examination mark in each subject or course. Because the tertiary entrance ranking, the tool used for selection into tertiary institutions, requires that these marks be put on one scale, it is necessary to align the subjects in terms of difficulty. This process can be done in part during the design of the courses and the design of school assessment, but the criteria of same difficulty level can be verified empirically post hoc using Rasch measurement.

In addition to the courses mentioned above, there are also Vocational Education and Training (VET) subjects, part of the senior secondary curriculum that are linked to the Australian Recognition Framework (ARF). These subjects can then be used in the calculation of tertiary entrance scores, (Griffin, Gillis & Taylor, 2002, p. 1).

The process of establishing a single tertiary entrance score from the complex array of both courses and assessment practices that is fair across all levels of the system, provides a challenge for the Curriculum Council. According to Andrich (2005), assessment has to be sufficiently rigorous and sufficiently fine grained so that the assessment results can be used for equitable selection into tertiary programmes. According to measurement principles, different measures, for example, two different assessments, have to be converted to the same scale before they can be added (Andrich, 2005). This can be illustrated by an example from physical measurement: Before comparing a centigrade temperature and a Fahrenheit temperature, the temperatures have to be converted to the same scale. In order to add the school mark to the examination mark, to establish a single score, the two sets of data have to have the same mean and the same standard deviation. Only once the two scores are scaled to the same mean and same standard deviation does it make sense to add the two scores. This is an example of linear equating (Section 2.4.1).

### 4.2.2 Assessment bodies

The Curriculum Council[9] of Western Australia has been mandated by the Western Australian government to design the curriculum from K-12 and to establish assessment and certification pro-cedures. The Curriculum Council is the forum for all stakeholders who have an interest in the direction of education.

[9] The name Curriculum Council is soon to be changed.

### 4.2.3 Standardisation, moderation and scaling

The Curriculum Council for Western Australia uses a standardisation, moderation and scaling procedure (Curriculum Council, pamphlet to parents, n.d.) to convert the school marks and the examination marks into a tertiary entrance score and then a tertiary entrance ranking (TER). The TER is based on the sum of the scaled marks achieved by each student in a particular combination of subjects. Subject marks are scaled prior to aggregation to ensure students are not disadvantaged by choosing a difficult subject. Subject marks are made up of school marks and external examination marks (TEE).

#### *4.2.3.1 Rationale for the standardisation process*

The rationale of the standardisation, moderation and scaling process is that this responds to the needs of three educational stakeholders: the students themselves, the teachers, and the tertiary institutions (Curriculum Council, pamphlet to parents, n.d.). The process enables:

- students to study different courses;
- teachers to develop teaching and assessment programmes suited to their student needs and acknowledge the work done throughout the year; and
- tertiary institutions to compare students who undertake different programmes.

Prior to the moderation process, students whose performance on TEE does not reflect their performance on school-based assessments and who might otherwise bias the results, are identified and taken out of the group that is used to calculate the parameters (Certification & Examination documentation, 2001). This requirement is necessary for scaling procedure, as anomalies are likely to affect the overall picture.

#### *4.2.3.2 School marks and examination marks*

In Western Australia, the school marks are scaled to the examination marks because all students write the examination. The two sets of marks for each course, the school mark and the examination mark are combined to give a score that is located on a common scale. Before this process can take place, both the examination mark and the school mark are put through a process. The process is illustrated in **Figure 3.**

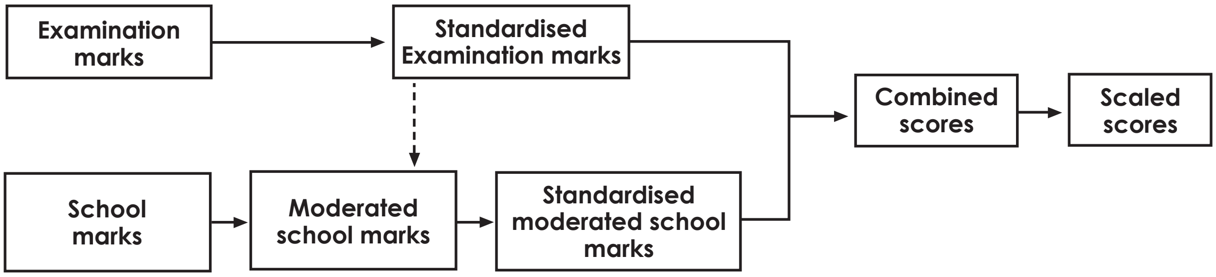

*Figure: 3 Marks adjustment process (Curriculum Council pamphlet to parents n.d.)*

#### *4.2.3.3 External Examinations*

A final external examination in all subjects is conducted by the Curriculum Council. Written papers are set by an examining panel of experienced teachers and are independently reviewed. This pro¬cess seeks to ensure that the examinations properly reflect the syllabus and are a fair test of student achievement. The examination scripts are separately marked by two qualified markers[10]. If they disagree, the markers decide together which mark is correct, or get a third marker to re-mark. This mark is called the **raw examination mark**.

The external examinations themselves are transformed to ensure that the unit is the same at different points on the scale and to fix the examination to the "conventional distribution, which is

[10] This process relies on professional judgment (see section 2.4.3)

the same from year to year" (Andrich 2005, p. 13). The outcomes of standardisation are:

- The same distribution of standardised marks from year to year and subject to subject.
- The top student is given a standardised mark of 100.
- Specified percentiles are matched to pre-determined scores in the standardised distribution. These are control points.

**Table 7** presents the transformation from raw scores to standardised scores.

**Table 7: Standardisation of raw examination scores**

| Raw score of: | Standardised Score |
|---|---:|
| 0 | 0 |
| Lowest score (non-zero)* | 20 |
| 10th percentile | 48 |
| 30th percentile | 60 |
| 70th percentile | 72 |
| 90th percentile | 80 |
| Highest score | 100 |

**The lowest non-zero raw score is assigned a standardised score by linear interpolation between the point (0.20) and the 10th percentile*

*(Source: Curriculum Council, 2005)*

### *4.2.3.4 Standardisation Procedure*

The standardisation function *(depicted in* **Figure 4***)* goes through the following process:

1. The percentile rank of each score is determined.
2. Raw scores, which fall on specified control points, are assigned predetermined standardisation scores.
3. Raw scores, which fall between control points, are assigned standardised scores by linear interpolation between the two nearest control points.

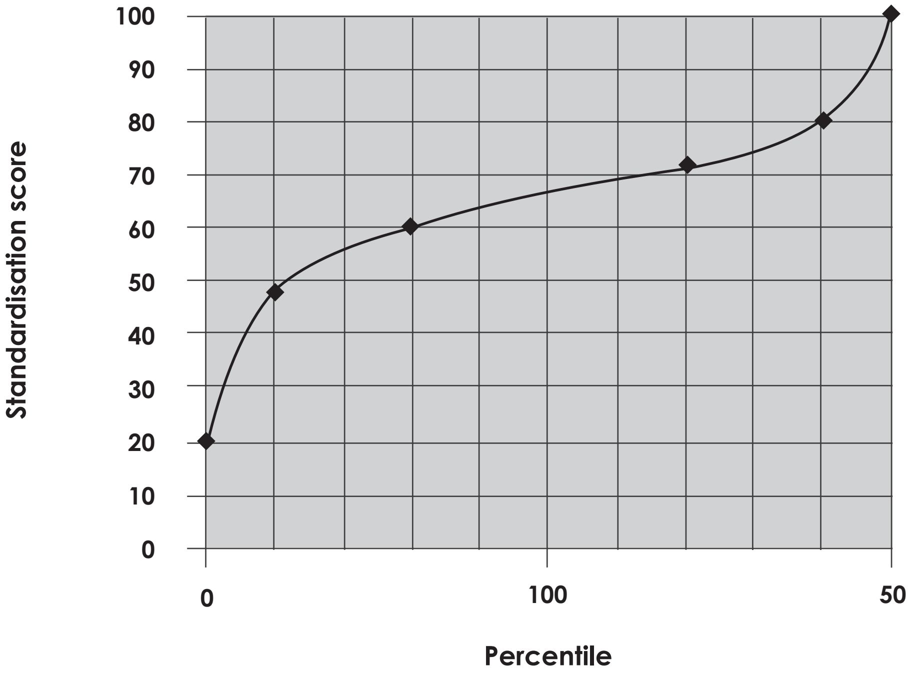

*Figure 4: The standardisation function*

*(Source: Curriculum Council, 2005)*

TEE examiners set papers so that the average raw mark is between 55 and 60. Standardised TEE marks tend to have an average above 66. If an examination yields exceptionally high raw TEE marks with an average above 66, standardised scores may be lower. When marks are standardised, they have an approximate linear relationship to achievement.

#### *4.2.3.5 School marks*

The school mark (made up of semester examinations, classroom tests, class work, research assignments and practical work) is submitted in the form of a symbol (A, B, C, D, or E), and a numerical mark, for each subject and course unit studied in the final year of senior secondary schooling. The symbols are recorded on the statement of results.

Moderated school marks are calculated from the numerical mark. These marks are adjusted to the same scale as the standardised examination marks. The procedure converts moderated school marks from all parts of the state to a common scale. This process is conducted using statistical modelling (probably Rasch measurement). Standardisation gives the raw TEE marks and moderated school marks similarly shaped distributions before they can be combined.
The process, according to the Curriculum Council (2003), is as follows

1. Each provisional moderation population is to consist of a school/subject group.
2. The numerical School Assessment of each student is adjusted so that the set of provisionally moderated school assessments has the same median and mean absolute deviation as the standardised TEE marks.
3. For each student the mark difference between the standardised TEE and the provisionally moderated school assessment is calculated.
4. The mean and standard deviations of the mark differences are calculated.

This monitoring process involves a statistic describing the scale of assessment of the school marks of each school/subject group relative the group's performance on TEE.

#### *4.2.3.6 Combining the examination and the school mark*

The standardised school mark and standardised examination mark are averaged to arrive at a combined mark. In order to add the two scores together, both school assessments and examinations need to be scaled to the same mean and standard deviation (Andrich, 2005). This scaling process is based on the notion of a "linking construct which has been referred to as 'academic ability' or 'potential'" (Partis, as cited in Newton 2005, p. 108). Score distributions, for each subject, are scaled to reflect the ability distributions of the students who studied them. The "ability is estimated through an iterative statistical process which models scores across a number of subjects studied" (Newton 2005, p. 121).

The differences in difficulty level are adjusted by applying the average marks scaling (AMS) statistical method (see section 2.4.2). For example, if the geography students as a whole group perform better across all their subjects relative to accountancy students, then the geography marks will be raised. The basis of the approach is to use a student's average mark across all subjects as the anchor variable for scaling each subject.

1. Scaling preserves the order of the students and the shape of the distribution in each subject.
2. The mean scaled score in EACH subject is equal to the mean scaled score across ALL subjects taken by all the students in that subject.
3. The mean scaled score across all subjects and all students will be 58.
4. The standard deviation of the scaled marks in each subject is equal to the standard deviation of the unscaled marks across all subjects taken by all the students in that subject.

### 4.2.4 Implications for South Africa

The Western Australian assessment system has been influenced by the principles of measurement that underpin the Rasch model. The universities in Western Australia have a long tradition of Rasch modelling and are used as consultants to the Department of Education (see Andrich, 2005).

Equating studies are used each year by the Western Australian Curriculum Council to enable system level comparisons. Guidance and technical expertise are provided by the universities in Western Australia, in particular Murdoch and the University of Western Australia, for the education system in general and, in particular, the assessment units. In addition, Rasch measurement is used to make comparisons between different grades and between different cohorts of students in the same grade over time (Humphry, 2005).

Using the Rasch model to statistically analyse the data in order to create measures is used in the WA system to diagnose and locate problems at various stages of the assessment process. This can inform the finer levels of precision that are required in the design of the assessment for a further cycle. The requirement for school-based assessment is that the marking be sufficiently fine grained and consistent so that the school based assessment can be scaled to the examination (Andrich, 2005).

While Western Australia has embarked on an outcomes-based curriculum, the policies for tertiary selection remain the same (Andrich, 2005, p. 49). The expectation of an exit level examination, tertiary entrance score and tertiary entrance ranking has been retained by the tertiary institutions. Andrich (2005, p. 49) recommends that the TES and TER be disaggregated and that tertiary institutions may want to consider “a school-based component, or a performance based component for particular programs”. Andrich (2005) perceives a need for the disaggregation of the TER, as for some tertiary programmes the requirement of a performance component or a school-based component may be more suitable than a single score, as in the TER. This process would, however, have to take place earlier in the process.

Another important component of the Western Australian system is the transparency with which the examination process is conducted. The Curriculum Council website provides information on the scaling, moderation and standardisation process. Teachers and parents are thereby informed on the process of obtaining a Tertiary Entrance Ranking (TER).

## 4.3 Indonesia

Indonesia has a surface area of 181 040 km² and has a population of 223 million people (World Development Indicators, 2007). The Gross Domestic Product (GDP) in 2005 was US$ 1 302 (Human Development Report, 2007). Of the population in Indonesia, approximately 36.1% lies below the national poverty line, with approximately 9.6% (1995-2005) of adults aged 15 and older being illiterate (Human Development Report, 2007). In 2002, the net enrolment rate for primary schools was 92.7% and for junior secondary education 61.7% (see **Table 9**). By 2005 the combined gross enrolment ratio for primary, secondary and tertiary education was 68.2%. Public expenditure on education as a percentage of the total government expenditure for 2002–2005 was 9% (Human Development Report, 2007).

### 4.3.1 Schooling in Indonesia

The structure of the Indonesian system consists of six years elementary or primary school, three years of junior secondary or middle school, three years of senior secondary or high school and then tertiary education (see **Figure 5** and **Figure 6**) (Mohandas, Wei & Keeves, n.d.). Furthermore, Indonesia also has a vocational track that is a sub-system of the Directorate of General and Secondary Education with the Ministry of Education and Culture (MOEC). The vocational track consists of junior technical and vocational schools (three years) and senior technical and vocational schools for three or four years (UNESCO, 1995).

The academic year runs from mid-July to mid-June. Primary school is free and compulsory and comprises three semesters. Secondary school, on the other hand, only comprises two semesters (World Education Services, 2008). Basic education is considered to be nine years of schooling, which includes six years of primary school and three years of junior secondary school (Quality Education for All, 2004).

At the end of each cycle in primary and secondary school, final examinations are administered (Mohandas et al., n.d.). The target grades for the examinations are 6, 9 and 12. The results are issued typically four weeks after the examination is written (World Bank, 2000). The pass rates are depicted in **Table 9**.

### Table 8: Enrolment rates (%) in Indonesia

| | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Net enrolment ratio in primary education** | | | 88.7 | | 92.1 | 91.5 | 91.5 | 92.3 | 92.1 | 92.7 | 92.3 | 92.9 | 92.7 |
| **Net enrolment rate in junior secondary education** | | | 41.9 | | 50 | 51 | 54.5 | 57.8 | 57 | 59.2 | 60.3 | 60.5 | 61.7 |
| **Proportion of pupils starting grade 1 who reach grade 5** | 75.6 | 74.7 | 74.3 | 75.6 | 77.5 | 80.2 | 81 | 80.9 | 82.2 | 81.8 | 82.6 | 81.9 | 82.2 |
| **Proportion of pupils starting grade 1 who complete primary school** | 62 | 62.6 | 63.4 | 64.4 | 66.1 | 68.1 | 70 | 71.3 | 71.9 | 73.3 | 74 | 75.1 | 74.4 |
| **Proportion of pupils starting grade 1 who complete 9 years of compulsory education** | 32.1 | 30.7 | 20.6 | 32.3 | 33.6 | 32.3 | 36.6 | 40.2 | 45.3 | 44.4 | 45.7 | 46.8 | |
| **Ratio of boys to girls in junior secondary school** | | | 101 | | 100 | 100 | 103 | 102 | 103 | 103 | 104 | 105 | 103 |

### Table 9: Candidate numbers and pass rates

| Year | Candidates | Pass rate (%) |
|---|---|---|
| 1997 | 9 100 000 | 97.76 |
| 1998 | 9 150 000 | 97.51 |
| 1999 | 8 560 000 | 97.85 |
| 2000 | 8 270 000 | 98.88 |

*(Source: World Bank, 2000)*

| Age | Year | Stages | Level | In-School | | | | | Out of School |
|---|---|---|---|---|---|---|---|---|---|
| 27 | 21 | Doctorate | Higher Education | Institute | S3 | | | | OPEN UNIVERSITY |
| 26 | 20 | | | | | | | | |
| 25 | 19 | | | | | | | | |
| 24 | 18 | Post Graduate | | | S2 | | | | |
| 23 | 17 | | | | | | | | |
| 22 | 16 | Under Graduate | | | | POLy | | | |
| 21 | 15 | | | | S1 | | ACA | D3 | |
| 20 | 14 | | | | | | | D2 | |
| 19 | 13 | | | | | | | D1 | |
| 18 | 12 | 3 | Secondary Education | GSS | VSS | RSS | OSS | SSS | |
| 17 | 11 | 2 | | | | | | | |
| 16 | 10 | 1 | | | | | | | |
| 15 | 9 | 3 | Basic Education | Junior Secondary School | | | | | Package B Programme |
| 14 | 8 | 2 | | | | | | | |
| 13 | 7 | 1 | | | | | | | |
| 12 | 6 | 6 | | Primary School | | | | | Package A Programme |
| 11 | 5 | 5 | | | | | | | |
| 10 | 4 | 4 | | | | | | | |
| 9 | 3 | 3 | | | | | | | |
| 8 | 2 | 2 | | | | | | | |
| 7 | 1 | 1 | | | | | | | |
| 6 | | OA | Kindergarten | Pre-School | | | | | |
| 5 | | OB | | | | | | | |

| | | |
|---|---|---|
| SMU | = | Sekolah Menengah Umum |
| SMK | = | Sekolah Menengah Kejuruan |
| SMKd | = | Sekolah Menengah Kedinasan |
| SLB | = | Sekolah Luar Biasa |
| MA | = | Madrasah Aliyah |

*Figure 5: The schooling system in Indonesia including age and grade levels (Source: SAMEO, 2006)*

A startling change, recorded by (Syahril, 2007, p. 4-5), was that 30% or 400 000 out of 1.9 million senior and vocational school students failed the national examination. Some high schools even had 0% passing rates. This obviously requires further investigation. The change could be attributed to changes in the education system noted above.

## 4.3.2 The examination system

The examination systems have undergone a number of changes in the past few decades (Syahril, 2007). Until the early 1970s the examination system was referred to as the state examination (Mohandas et al., n.d.). The state examination was undertaken for most of the subjects at primary school, junior secondary school and senior secondary (elementary, middle and high school) (Syahril, 2007). In this system a national committee at a central level prepared the examination papers for all subject-matter areas. From the early 1970s to the early 1980s the examination changed to school examinations. In this system, the school was given the authority to construct its own examinations, to score the examinations, as well as to decide on the passing grade for the students who wrote the examination (Mohandas et al., n.d.). However, Indonesia reverted to a more centralised examination system known as Evaluasi Belajar Tahap Akhir Nasional or Ebtanas for short. The purposes for Ebtanas were:

1. to determine the path of learners;
2. to filter students to the next education level; and
3. to inform quality improvement (Syahril, 2007).

In 1998, more reforms took place as a decentralised education system was instituted (Syahril, 2007). In 2002, as a result of strong considerations to abolish standardised testing, Ebtanas, at the primary school level, was abolished. Junior and senior secondary school examinations remained in place although only three subjects now form part of the national examinations as opposed to five and seven subjects respectively. Three subjects are examined, namely Indonesian, English and mathematics. The new standardised examination was called Unjian Akhir Natsional or UAN for short. In 2005 the name changed again, this time to Unjian Natsional or UN for short (Syahril, 2007). The grading system in Indonesia was initially on a scale of 0 to 10 with 3.01 indicating a pass. However, the minimum threshold has subsequently been revised to 4.51.

### 4.3.3 Assessment Bodies

Essentially the Ministry of Education and Culture (MOEC) was responsible for the examinations at the various levels of the education system as described above, as part of the centralised government system which prevailed at the time (Bjork, 2004) (see Figure 5 for the organisation of MOEC). A drive from the 1980s to the present to raise the standards of education across the country resulted in the institution of the National Final Examination of Student Achievement (Mohandas et al., n.d.). In this examination system, the administration of the teaching-learning processes as well as the examinations was the responsibility of the Directorate of Primary and Secondary Education (see **Figure 6**).

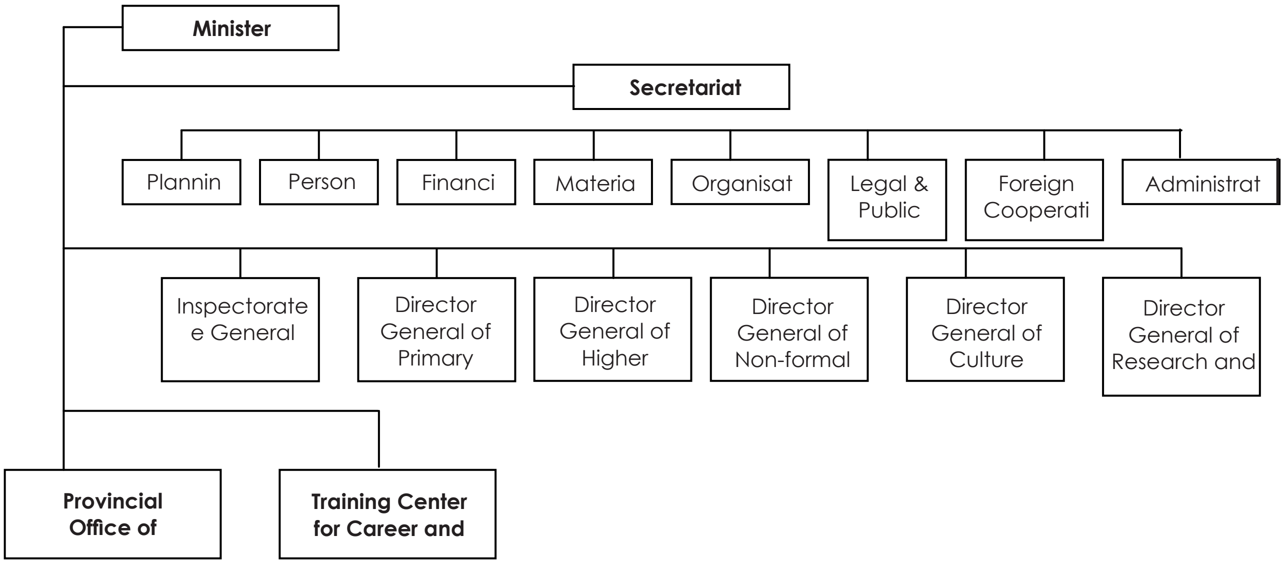
Minister
Secretariat
Plannin
Person
Financi
Materia
Organisat
Legal & Public
Foreign Cooperati
Administrat
Inspectorate e General
Director General of Primary
Director General of Higher
Director General of Non-formal
Director General of Culture
Director General of Research and
Provincial Office of
Training Center for Career and

*Figure 6: The organisation of the Ministry of Education and Culture*

*(Source: UNESCO, 1995)*

The Examination Development Centre (a research and development institute within the Office of Educational and Culture Research and Development) has been responsible for the design and development of examinations. The tasks included construction of a national item bank as well as technical guidance to the item writers (Mohandas et al., n. d.).

In previous test administrations, several forms are constructed for each subject tested. The forms are constructed according to test specifications in terms of similar content areas as well as level of difficulties (Mohandas, n.d.), specifically for the primary and junior secondary schools (Mohandas et al., n.d.). The provincial offices were given the authority to:

- Review the forms
- Finalise the examinations
- Administer the examinations
- Score the papers.

The Directorate General of Primary and Secondary Education centrally determined the passing score for the examinations. For senior secondary education, the situation was slightly different. For senior secondary education, the provincial committees develop one set of items for each subject area. The items are then sent to Jakarta to be reviewed by the national team that compiled the test specifications. The national team is then responsible for the preparation of seven sets of items out of the 26 sets reviewed from the various provinces within Indonesia. The final versions are then sent to the provinces for duplication and administration. The provinces are also given the authority to decide on the results of the examination (Mohandas et al., n.d.).

In 2000/2001 the Minister of National Education gave full authority to the Examination Development Centre to administer the examinations. As a result, the examination at the end of primary school was abandoned. The Centre is responsible for the construction and administration of the examinations at junior and senior secondary schools. The construction of test items at both levels is carried out by choosing the relevant subject matter area and appropriate difficultly level from an item bank that is developed and managed by the Examination Centre (Mohandas et al., n.d.).

In the past, learners had to pass public examinations in seven subjects. The examinations were multiple choice questions drawn from national items and marked according to national procedures (as mentioned earlier). The final examinations are set at a national level as well as partly at a school level. The national portion of the examination takes place in three subjects, namely mathematics, English and Indonesian Language. The examinations are multiple-choice drawn from an item bank, as discussed previously. The school portion comprises seven subjects and is based on theory and practical work (Quality Education for All, 2004). The core content for basic education curriculum includes Pancasila (which is state ideology), religion, civic education, Indonesian language, reading and writing, mathematics (this includes an arithmetic component), introduction to sciences and technology, geography, history (national and world), handicraft and art, physical and health education, drawing, English, and local content (SEAMEO, 2006).

### 4.3.4 The rationale for implementing the system

The purpose of the examination is for certification of the student's level of educational achievement, selection to the next level of education, evaluation of performance of the school and teacher and to provide feedback to schools and teachers alike (World Bank, 2000). More specifically for elementary and junior secondary schools, the final examination serves for certification and selection to a higher level of education. For senior secondary school, the examination serves as certification only, as universities administer their own entrance examinations (Mohandas et al., n.d.).

### 4.3.5 The use of IRT/Rasch

In 1994, the Examination Development Centre introduced a common set of items in five different test forms (Mohandas, 1996). The use of the common set of items is to equate the different test forms in each of the subjects. The common items are used for linking the different forms; Rasch equating is used to obtain comparable scores across the different test forms (Mohandas et al., n.d.).

Rasch analysis is used by the Examination Development Centre. Common items in each subject are used to link Forms A and B, B and C, C and D and finally, D and E. This linking of assessment forms means that the different test forms can be equated (Mohandas, 1996).

Thus, Test A or Form A and Test B or Form B is linked with a set of items common to both. This can be referred to as AB. Test B and Test C are link using the same procedure which is referred to BC. This process is repeated using each Form or Test (see **Figure 7**).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Test A** | **AB** | | | | | | | |
| | **AB** | **Test B** | **BC** | | | | | |
| | | | **BC** | **Test C** | **CD** | | | |
| | | | | | **CD** | **Test D** | **DE** | |
| | | | | | | | **DE** | **Test E** |

*Figure 7: A matrix of common set of items used in one year (Mohandas, n.d., Mohandas, 1996)*

In order to undertake this type of linking and equating, it is of importance that the assumptions of the Rasch model as described earlier, not be violated. Thus any person or item that does not fit in the Rasch sense is checked and possibly removed from the exercise (Mohandas, n.d.). By using a design as described above, the test forms can be equated by means of anchor item equating or concurrent equating, as described earlier in the document. Mohandas (1994) recommends that concurrent equating should be used as the estimates seem to more robust and of greater strength than with anchor item equating.

**The requirements of the system**

- Item banks
- Test specifications
- Different forms
- Scoring specifications
- Pass requirements
- Technical expertise in Rasch and all the statistical background needed to explore the assumptions of Rasch, such as unidimensionality, statistically.

### 4.3.6 Benefits of the system

Perhaps the most important aspect of equating is that learner scores from different test forms can be compared. This is all important not only for the maintaining of standards within a year but also when looking at comparisons across years. In the words of Mohandas et al. (n.d., p. 15) equating "serves the purpose of monitoring the achievement levels of students over the years and the different locations, so that this information can also be used in an effort to improve the quality of education provided by the schools."

### 4.3.7 Implications for South Africa

For South Africa, the implications of adopting such a system would be that an item bank be developed over a period. The item banks for each learning area should be divided into the various learning outcomes and aligned with the assessment standards and, for this phase, subject specialists will be needed from the various provinces within South Africa. Furthermore, the conceptual as well as empirical item difficulty of each item should be thoroughly explored and included in the item bank. In order to explore the conceptual and empirical difficulty of items both subject specialists and psychometricians, specialising in IRT, will be needed. The item bank should include items with a range of difficulty levels and will need to be large enough to facilitate the extraction of a number of forms for each learning area and learning outcome. The items in the item bank will also have to be revised regularly and updated to facilitate the releasing of items for teaching and learning purposes. A detailed log of which items are released should also be kept. These items should be exemplar items that will be able to assist in teaching and learning. Furthermore, subject specialists, psychometricians and administrators will be needed in order to make such a system feasible in the context of South Africa.

## 4.4 Summary

The three case studies were chosen not only because of their location but also to reflect a developing and developed world context. Within each case study, the country context is provided in addition to the examination system used. The examination system in all three cases is administered by the various Departments of Education in addition to centres of excellence in the field of test development. Where applicable the standardisation process and the use of Rasch have been elaborated.

In Western Australia and The Netherlands, assessment experts have broadened the assessment methods applied within schooling. The system accommodates a variety of assessments but simultaneously retain a focus on student selection for university. They have retained public confidence and that of the universities and avoided the latter striving to develop their own entrance examinations. This is despite the fact that both countries combine both school-based marks with examination marks to compile the final overall marks for the pupils. This situation is not explicitly stated in the Indonesian literature.

There are a number of interim examinations in all systems with Indonesia offering examinations at the end of junior secondary for entrance into senior secondary and again at the end of senior secondary for entrance into university. The Netherlands imposes a national examination at the end of primary school to enable pupils, parents and teachers to advise the pupils which of their education tracks to undertake in secondary school. The next examination is that in the final year of secondary schooling.

For all three countries examinations are driven centrally by the government. In Western Australia, this is the Ministry of Education, for The Netherlands two separate parastatal organisations responsible for organising and then designing and developing the examinations and, in Indonesia, an examination unit within the Ministry of Education is responsible. The support systems, however, are substantial in two out of the three cases. Western Australia is home to one of the world's authorities on IRT, Prof. David Andrich at the University of Western Australia. He and his team of researchers are intimately involved in giving technical support to the Western Australian government in the design, development and implementation of the examination system. In The Netherlands, two national organisations established by the government have considerable expertise and resources. Furthermore, The Netherlands is renowned for its technical expertise in the field of psychometrics and the two main organisations are well supported by experts at the Universities of Groningen, Twente and Nijmegen, to mention but three. The Indonesian examination unit is an island containing most of the countries' experts who, as mentioned previously, have been trained internationally in the USA, England, the Netherlands and Australia. Their external technical support has come largely from international expertise; they have now had to become self-sufficient.

In terms of IRT itself, each system utilises IRT in different ways. Indonesia has established item banks that apply IRT. These are used to produce different examination forms which, after administration, are equated using IRT. In contrast, The Netherlands only uses IRT in the analytical phase when they apply Rasch modelling as a standard procedure to analyse the performance of the pupils.

**Table 10: Summary of information present in the case studies**

| | The Netherlands | Western Australia | Indonesia |
|---|---|---|---|
| **Structure of schooling** | VMBO - preparatory middle-level and vocational education<br>HAVO – higher general education<br>VWO – preparatory scientific education | Compulsory schooling from years K to 10<br>Primary, junior and senior secondary education<br>A number of options are available for years 11 and 12, leading to different forms of tertiary education | Six years elementary or primary school, 3 years junior secondary, 3 senior secondary, University |
| **Examination system** | School exit level<br>Selection to tertiary | School exit level Selection for tertiary | School exit level<br>Selection to next level |
| **function** | education | education | Monitoring |
| **Assessment Bodies** | Central Committee for Ratification of Examination (CEVO)<br>National Institute for Educational Measurement (Cito) | Curriculum Council of Western Australia | Ministry of Education and Culture (MOEC)<br>Examination Development Centre |
| **Use of IRT/Rasch** | Standards compared before examination<br>Reference examination | Research, comparisons, equating studies undertaken after examinations. | Test forms, based on test specifications, are equated by means of common item linking |
| **Implications** | Thorough research into the best possible system is undertaken before action is taken | Equating studies necessary to enable comparisons.<br>Guidance and technical expertise is needed.<br>Rasch can be used to locate and diagnose problems at various stages.<br>Transparency of the examination process. | Item bank in order to generate various forms of assessments.<br>Empirical and conceptual difficulty has to be ascertained.<br>Guidance and technical expertise is needed. |

In Western Australia, the Rasch model is applied in their post hoc analyses to provide feedback for the following cycle. Interestingly The Netherlands and Indonesia both apply a 10-point scale (1-10) and, in the case of the former, 5.5 points on this scale represents a pass mark in contrast to Indonesia where 4.51 is used as the pass mark.

In summary, although the case studies presented and compared in the preceding sections provide the depth and substance for the reader, a crude summary is possible (see Table 10). The three systems have various "sectors" within the system with exit levels where examinations are key and fulfil a certain function (as described in Section 3) whether certification, monitoring or selection to the next level of education. In the case of The Netherlands, Western Australia and Indonesia, the prominent functions are selection and certification. All three case studies make use of Rasch at various stages of the process, whether linking various test forms to obtain comparable scores or post hoc. South Africa can learn from these three case studies. This is discussed in more detail in the section to follow (Section 5).

# Section 5:

## Conclusions and Implications

The introduction of a new national set of examinations in 2008 in South Africa offers a unique opportunity to review the types of procedures and techniques utilised to date to support the maintenance of standards and moderation of examinations in South Africa. Traditionally in South Africa, examinations have served the functions of accreditation and selection into further/higher education as well as the job market. The matriculation examination is a high stakes examination with a great deal of public interest. The present methods of moderation, though little understood by the general public, are regarded as valid. It is essential to maintain the confidence of the school, public and tertiary systems in the matriculation system, for a number of reasons, one of which is that a good assessment and examination system does a great deal to hold an education system together. Even when other aspects of the system break down, such as poor teaching, access by students to a clearly defined curriculum, together with the exemplar papers, which provide information about the depth of knowledge required, enable a reasonable measure of success for some learners.

Very clearly, Greaney and Kellaghan's (1996) warning that "a clear definition of standards needs to be maintained when introducing a newly constructed examination is very pertinent to the South African situation, otherwise it is not possible to make meaningful comparisons about performance from one examination to another "(Greaney and Kellaghan, 1996, p. 35). With this warning in mind, this report provides an overview of approaches used internationally to link assessment results of high-stakes examinations, as well as examination systems in general, associated techniques, with a specific focus on IRT and the Rasch model and three case studies of examination systems in selected countries.

Arising from Section 2 and the technical background needed for this report, the information from the case studies suggests the need for the combination of "subjective" judgement and empirical measurement. In addition to the more traditionally and widespread practices of involving judgement as the means to equate tests or examinations, there is a wider call for the inclusion of a model such as Rasch to provide scientific measurement. Arguments in favour of including Rasch are that more rigour is required in examinations processes in general. The Rasch model to date is a stringent tool and is the only model to provide tools for approximating objective reproducible additive measures in human sciences. The fact that the Rasch model can deal with missing data and can therefore link tests through common examinees or sets of examinees through common test items, confirms the appropriateness of this model. A requirement for the Rasch model, and indeed all psychometric enterprises, is that of conceptual clarity concerning the underlying construct. The review of this model and the measurement possibilities seems to suggest that it can provide a means to introduce more rigour into the measurement aspect of examinations in South Africa.

The three case studies reviewed are from very different contexts: two well-resourced countries with state of the art test "technology" and the third working with similar state-of-the-art test technology but within a poorer environment and with considerable under-funding. In Western Australia and The Netherlands, there is a well established and strong assessment culture. In Indonesia, one might describe it as emerging thanks to the efforts of a few and the leadership of a remarkable man who has systematically sent 10 to15 staff members to undertake their master's and PhD degrees in the USA, England and Australia in psychometrics in order to develop the examination unit within the Ministry of Education.

Whilst no calculations of actual costs from any one of the three systems could be included in this report, it should be noted that applications of IRT are generally more expensive than similar applications of classical test theory and many applications of IRT require the appropriate software. It is also evident from analyses of the case studies that the different educational, historical and current contexts, the type of examination system, the time factor and the amount of financial and human resources determine the sophistication of the methodology that is used.

IRT was found to be more prevalent across developed and developing countries in their national assessments than in their public examinations. One of the possible reasons for this is the time pressure associated with the turnaround from writing examinations to publishing the results in time for the application and entry into universities. However, in systems where it is used, such as the three case studies, one of the overriding reasons is to maintain standards and the ability to link one year's results with the previous and following years. Furthermore, there is a sense of social equity as the data is empirically generated for the purpose of fairness in ascertaining the difficulty of examination items and in the preservation of standards. The information derived from the application of IRT to examination items is also used for feedback to the test constructors to validate the decisions with regard to difficulty levels of items. In the case of The Netherlands, the further analysis of 2 000 scripts provides justification for moderation on individual items. In particular, the Rasch model was found to be widely used and recommended because of its speed and simplicity.

In summary, IRT provides for many powerful applications to measurement problems (Stocking, 1999, p. 59-62). The following are examples of these:

- Test construction – tests with pre-specified measurement properties can be constructed from a pool of calibrated items.
- Redesigning an existing test – relative efficiency functions provide a convenient way of investigating various design changes in a test and comparing them with the original test.
- Equating – the process of finding corresponding scores on different forms of a test.
- Item bias – items should function similarly for different subgroups of the population. Where items function differently for different subgroups, they need to be investigated for item bias.
- Mastery testing – to determine if a person has reached a specified level of achievement.
- Tailored or adaptive testing – every person is administered items that target the person's ability best.

The recommendation is therefore made that consideration be given to the incorporation of Rasch measurement in the South African examination system at selected nodes in the overall examination process. Various examples of the use of Rasch have been given in this report with the benefits of these highlighted and the limitations of the approach also described. The major limitation of using the Rasch model is the fact that it is little known in South Africa. Even in countries such as the UK the Rasch model is met with a measure of scepticism underpinned by the wariness of change. The Rasch model and IRT models are used in different ways in the Australian states and territories. To date only Tasmania has used the Rasch model in relation to their exit level examinations. South Africa should aim for a system whose processes are scientifically rigorous and can be defended, and therefore made available to stakeholders.

The first step in this process is for the existing moderation, scaling and standardisation process of the South African system to be captured and made explicit to the advisory group, so that nodes can be identified at which immediate improvement in the system can take place. In order to do this, the cooperation between the National Department of Education and Umalusi and their complementary roles need to be clearly defined. There needs to be agreement as to processes to be instated. An advisory body (as recommended by WA in response to the Andrich report 2005) for the purpose of instating alignment of the processes needs to be in place. This may be the role of the existing Umalusi Research Group, the Umalusi Statistical committee, or another group specifically assigned with this task. In any case, the specific roles of the Umalusi advisory committees could also be made clear.

Maintaining confidence in the examination system means thorough review and piloting of aspects of the system before implementation. The success of The Netherlands system is attributed to the stability of the system, the fact that they have not rushed into making changes before thorough review of the processes has been in place. Maintaining confidence means educating all sectors from the education officials, school principals and teachers, to parents and students about the process of moderation and standardisation.

The systems presently in place in the South African system such as the 1) moderation of school marks and 2) the combining of school marks and examination marks can be improved through targeting the specific nodes in the system. The systems of moderation may at this point include IRT (or Rasch modelling) for retrospective confirmation or for the need for radical improvement. These could be instituted at specific nodes in the process.

The moderation of school marks is achieved through what is termed a social moderation or professional ratification process in the literature. This process has been perfectly acceptable up until now. According to Newton (2005; 2007) this is the system in place in most examination systems in the UK. We have, however, now additional techniques for supporting ratification by professionals. Both The Netherlands and Western Australia support the ratification process through Rasch modelling. According to recent improvements in psychometrics such as IRT (and Rasch modelling), the exam-ination marks and the school marks need to be on the same scale, that is, have the same mean and standard deviation, before they can be combined (see Rasch and linking section of this report). Anomalies, such as students whose school marks and examination marks are not in alignment, need to be excluded from this process. Their results can then be estimated at a later stage in the process.

The implications for South Africa in adopting such reforms are significant in terms of both the financial investment into human resource development, and the infrastructure and the equipment needed. These would include the need to establish a unit, within Umalusi or closely associated with Umalusi and the National Department of Education, capable of designing and developing a system where IRT or the Rasch model is used at various stages of the examination process. The possibility is to follow the Indonesian example and to develop item banks of examination questions that have been piloted and subjected to a Rasch Analysis. These items could be piloted under controlled conditions and administered by officials from the National Department of Education.

The South African educational context is defined by OBE. The assessment practices required for a qualification such as the matriculation examination, the results of which enable tertiary entrance, are required no matter what the educational principles are that guide the learning and teaching. It is assumed with Andrich (2005) that "OBE is a much wider set of educational principles than a set of assessment practices". The need is that the assessment practices be "compatible with OBE practices in meeting the requirements of rigour and precision required for tertiary selection" (2005, p. 3).

## 5.1 Recommendations

Firstly, the National Department of Education and Umalusi, the institutions responsible for the South African examination system, have structures in place that are the starting points for reform. Many of the processes are also in place, for example, the school moderation of marks, and the Statistical Committee's moderation of matriculation results using Subject Pairs Analysis (SPA), a process commonly used by examination bodies in the UK. Against this background, the following is recommended:

- From the existing Umalusi structures, i.e. committees, an advisory body is set up to design and monitor the reform of the system. Information on how the Western Australian system works is available in publications on the Curriculum Council website. This group needs to include pivotal people from the department who make policy decisions, statisticians and researchers who can take on new modelling techniques, subject experts and experienced examiners.

- Professionals with experience in this field, such as David Andrich's group at the University of Western Australia could be called in to assist with the reform. In addition, other professionals with experience in the field of monitoring and standardising examinations such as Robert Coe, deputy director of the Curriculum, Evaluation and Management (CEM) Centre, who has conducted research into the existing methodologies for evaluating difficulty levels, could provide advice on the reform.
- The reform strategy at this point could include:
  - The moderation of school marks to include both professional ratification and, in addition, scaling to conform to standards, as is done in Western Australia.
  - The training and informing relevant stakeholders of the processes that are involved.
  - The capturing of examination data by item and part of item. This round of examinations could form the pilot for the capturing and use of Rasch modelling, both the dichotomous and the partial credit model. This would provide feedback for test constructors. This analysis would inform the next cycle, which presumably has already started.
  - The analysis of a sample of scripts from a sample of schools of the preliminary examinations in September 2008. This could inform the subsequent analysis of the final scripts.
- Much of the reform process rests on existing structures. The importance, therefore, of the National Department of Education and Umalusi making those structures and processes explicit to an advisory body is essential.

In conclusion, this desktop study encountered a number of challenges on accessing detailed information on examination systems. As already mentioned, much literature was implicit rather than explicit, requiring the research team to deduce the methodology in several cases. The late arrival of key resources, required for the completion of the case studies, resulted in changes to the report up to the current date. No financial information was available from any of the sources found, making a cost analysis impossible to compute given the parameters of this study. It is, therefore, recommended that given the rather limited information freely available, it is essential that both an in-depth analysis of the technical requirements for applying Rasch analysis at pivotal points in the study be investigated, and an in-depth study of the critical literature, much of which is referenced in this study, be undertaken.

# References

Afemikhe, O. A. (2005). Reflections on the quality of assessment in large class in Nigeria. Paper presented at the International Association for Educational Assessment. Retrieved 20 March, 2008, from http://www.iaea.info/abstract_files/paper_051218094115.doc.

Afemikhe, O. A. (2007). Assessment and educational standard improvement: Reflections from Nigeria. Paper presented at the International Association for Educational Assessment. Retrieved 20 March, 2008, from http://iaea2007.tqdk.gov.az/cp/assessment%20and%20educational%20standard%20improvement%20reflections%20from%20nigeria.pdf.

Alberts, R. V. J. (2001). Equating exams as a Prerequisite for Maintaining standards: experience with Dutch centralised secondary examinations, Assessment in Education: Principles, Policy and Practice 8(3), 353-367.

Alberts, R. V. J. (2007). Verslag van de examencampagne 2007 voortgezet onderwijs. Arnhem: Stichting Cito Instituut voor Toetsontwikkeling.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In Keats, J. A. et al. (Eds.), Mathematical and Theoretical Systems, pp. 7-16. North Holland: Elsevier Publishers BV.

Andrich, D. (2005) A report to the Curriculum Council of Western Australia regarding assessment for tertiary selection.

Andrich, D. & Marais, I. (2006). EDU435/635 Instrument Design with Rasch IRT and Data Analysis I, Unit Materials. Perth: Murdoch University.

Beaton, A., & Johnson, E. G. (1992). Overview of the scaling methodology used in the national assessment. Journal of Educational Measurement, 29(2), 163-175.

Beaton, A., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. Journal Educational Statistics, 17(2), 95-109.

Beaton, A., & Johnson, E. G. (1992). Overview of the scaling methodology used in the national assessment. Journal of Educational Measurement, 29(2), 163-175.

Béguin, A. A. (2000). Robustness of Equating High-Stakes Tests. PhD thesis. FEBODRUK B. V., Enschede.

Béguin, A., Alberts, R. & Kremers, E. (2008). Normhandhaving bij examens. Retrieved 22 May, 2008, from http://toetswijzer.kennisnet.nl/html/normering/home.htm.

Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. The Journal of Economic Education, 29 (2), 171-182.

Bjork, C. (2004). Decentralisation in education, institutional culture and teacher autonomy in Indonesia. International review of Education, 50, 245-262.

Bond, T. & Fox, C. (2007). Applying the Rasch model: Fundamental Measurement in the Human Sciences. New Jersey: Lawrence Erlbaum Associates.

Boone, W. J. & Rogan, J. M. (2005) Rigour in quantitative analysis: The promise of Rasch analysis techniques. African Journal of Research in SMT Education, 9(1), 25-38.

Bramley, T. (2006). Equating methods used in KS3 English and Science. Retrieved 20 March, 2008, from http://www.cambridgeassessment.org.uk/ca/digitalAssets/114007 Equating methods KS3 TB.pdf.

Certification and Examinations Documentation (2001) Procedures for identifying anomalous performers in the TEE. Retrieved 5 May, 2008, from www.curriculum.wa.edu.au.

Cody, R. P. & Smith, J. K. (1997) Applied Statistics and the SAS Programming Language, (4th ed.) New Jersey: Prentice Hall.

Coe, R, (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch model, Oxford Review of Education, 34, 1-28.

Coe, R., Searle, J., Barmby, P., Jones, K. & Higgens, S. (2008) Relative difficulty of examinations in different subjects. Durham University.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. United Kingdom: Wadsworth Thomson Learning.

Curriculum Council (2008). Government of Western Australia, Standardisation of marks. Retrieved 2 May, 2008, from www.curriculum.wa.edu.au.

Curriculum Council, (2003) Monitoring small group partnerships. Retrieved 2 May, 2008, from www.curriculum.wa.edu.au.

Crown. (2008a). About KS3. Retrieved 20 March, 2008, from

http://www.standards.dfes.gov.uk/secondary/keystage3/aboutks3/.

Crown. (2008b). Key Stage 3 national strategy guide: Overview. Retrieved 20 March, 2008, from http://www.standards.dfes.gov.uk/secondary/keystage3/aboutks3/strategyguide/

? view= Standard.

De Wit, D., Van de Ven, A., & Van der Mijl, J. P. (2007). Key figures 2002-2006, Education, Culture and Science in the Netherlands, Kelpen-Oler: Hub. Tonnaer.

Downing, S.M & Halaydna, T.M. (2006). Handbook on test development. London: Routledge.

Eckstein, M.A., & Noah, H. J. (1989). Forms and functions of secondary-school-leaving examinations. Comparative Education Review, 33 (3), 295-316.

Field, A. (2005). Discovering Statistics Using SPSS. London: SAGE Publications.

Glass, G. V. & Stanley, J. C. (1970). Measurement, scales and statistics. Statistical methods in education and psychology, New Jersey: Prentice Hall.

Greaney, V. and Kellaghan,T. (1996). Monitoring the learning outcomes of Education systems: Directions in development. Washington. D.C: World Bank.

Griffin, P., Gillis, S. & Taylor, M. (2002). Scored Assessment for Senior Secondary Certificates. Retrieved 2 May, 2008, from www.aare.edu.au/02pap/gri02636.htm.

Hasan, S., Bagayoko, D. and Kelley, E. L. (1999). Misconceptions and the Certainty of Response Index (CRI). Physical Education, 34(5), 294-299.

Henson, R. K. (1999). Understanding the one-parameter Rasch model of item response theory. Paper presented at the Annual Meeting of the Southwest Educational Research Association, United States of America.

Howie, S.J. (2008). Standard setting: some issues and considerations. Presented at Umalusi workshop for setting standards, Pretoria, 23 April 2008.

Human Development Report. (2007). Indonesia. Retrieved 20 March, 2008, from http://hdrstats.undp.org/countries/data_sheets/cty_ds_IDN.html.

Humphry, S. (2005). Maintaining a common arbitrary unit in social measurement. PhD thesis. Perth: Murdoch University.

Humphry, S. & Andrich, D. Humphry, S. & Andrich, D. (2008), The role of the unit in the Rasch Measurement Model. Presented at 3rd International Rasch Conference, 22nd-24th January, 2008.

Huysamen, G. K. (1983). Introductory Statistics and Research Design for the Behavioural Sciences. Bloemfontein: UOFS.

Keeves, J.P. (1994). Examinations: Public. In T. Husen and T. N. Postlethwaite, (Eds.), The international Encyclopedia of Education, Oxford: Pergamon.

Linn, R.L. (1993). Linking results of distinct assessments. Applied Measurement in Education, 6(1), 83-102.

Linn, R.L. (2000). Assessment and Accountability. Educational Researcher, 29(2), 4-16.

Lord F.M. (1980) Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. Journal of Educational Statistics, 9 (1), 25-44.

Madaus, G. F. & Greaney, V. (1985). "The Irish Experience in Competency Testing: Implications for American Education." American journal of Education, 93, 268-94.

Madaus, G. F. & Kellaghan, T. (1992). "Curriculum Evaluation and Assessment." In P. W. Jackson, ed., Handbook of Research on Curriculum. New York: Macmillan.

Masters, G.N. & Keeves, J.P. (Eds). Advances in Measurement in Educational Research and Assessment. Oxford: Pergamon.

McCamey, R. (2002). A primer for the one-parameter Rasch model. Paper presented at the Annual Meeting of the Southwest Educational Research Association, United States of America.

Mislevy, R. J., Johnson, E. G. & Muraki, E. (1992). Scaling Procedures in NAEP. Journal of Educational Statistics, 17(2), 131-154.

Mohandas, M., Wei M.H., & Keeves, J. P. (n.d.). Evaluation and accountability in Asia and Pacific Countries. Retrieved 20 March, 2008, from

http://info.worldbank.org/etools/docs/library/117783/Eval_and_Accountability_in_Asian_Pacific_ramon.pdf.

Mohandas, R. (n.d.) Test equating. Retrieved 20 March, 2008, from http://info.worldbank.org/etools/docs/library/117785/handout_equating.pdf.

Mohandas, R. (1996). Test equating, problems and solutions: Equating English test forms for the Indonesian Junior Secondary School Final Examination administered in 1994. Unpublished Master's Dissertation: University of Flinders, Australia.

National Assessment of Educational Progress. (2007). Technical report on the NAEP mathematics assessment in Puerto Rico. National Center for Education Statistics. U.S. Department of Education. Retrieved 20 March, 2008, from

http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007462rev.

Newton, P. E. (2005). Examination standards and the limits of linking, Assessment in Education, 12(2), pp 105 – 123.

Newton, P. (2007). Techniques for monitoring the comparability of examination standards. Paper presented at the International Association for Educational Assessment 33rd Annual Conference, 16-21 September 2007, Baku, Azerbaijan. Retrieved 20 May, 2008, from http://iaea2007.tqdk.gov.az.

Obioma, G., & Salau, M. (2007). The predictive validity of public examinations: A case study of Nigeria. Retrieved 20 March, 2008, from http://iaea2007.tqdk.gov.az/cp/THE%20PREDICTIVE%20VALIDITY%20OF%20PUBLIC%20EXAMINATIONS.pdf.

Partis, M. T. (1977). Scaling of Tertiary Entrance Marks in Western Australia, Curriculum Council. Retrieved 20 March, 2008, from www.curriculum.wa.edu.au.

Planinic, M., Boone W. J., Krsnik, R. and Beilfuss, M. L. (2006). Exploring Alternative Conceptions from Newtonian Dynamics and Simple DC Circuits: Links between Item Difficulty and Item Confidence. Journal of Research in Science Teaching, 43(2) 150-171.

Potgieter, M., Davidovitz, B. & Blom, B. (2004). Chemical Concepts Inventory of First Year Students at Two Tertiary Institutions in South Africa. Paper presented at the SAARMSTE Conference.

Power, C., & Wood, R. (1984). A review of programs in Australia, the United Kingdom, and the United States. Comparative Education Review, 28(3), 355-377.

Qualifications and Curriculum Authority. (2008). Subjects: Key Stage 3. Retrieved 1 June, 2008, from http://curriculum.qca.org.uk/key-stages-3-and-4/subjects/index.aspx.

Quality Education for All (2004). Quality education for all young people: Challenges, trends, and priorities in Indonesia. Paper presented at the 47th International Conference on Education. Geneva, Switzerland, 8–11 September 2004. Retrieved 20 March, 2008, from http://www.ibe.unesco.org/International/ICE47/English/Natreps/reports/indonesia.pdf.

Rasch, G. (1960/1980). Some probabilistic models for the measurement of attainment and Intelligence, Chicago: Mesa Press.

Rentz, R. & Bashaw, W.L. (1977). The National Reference Scale for Reading: An application of the Rasch Model. Journal of Educational Measurement, 14(2), 161-179.

Ryan, J. & Williams, J. (2005). Rasch modeling in Test Development, Evaluation, Equating and Item Banking. Paper presented at the Third International SweMaS conference, Umea, October 14-15, 2003.

SAEMEO. (2006). Indonesia: National education system. Retrieved 20 March, 2008, from http://www.seameo.org/index.php?option=com_content&task=view&id=62&Itemid=85.

Schulz, M. (2002) Standardization of mean-squares. Rasch Measurement Transactions, 16(2), 879, Retrieved on 10 December 2007, from http://www.rasch.org/rmt/rmt162g.htm.

Smith, E. V., & Smith, R. M. (2004). Introduction to Rasch Measurement. Maple Grave, Minnesota: JAM Press.

Stocking, M.L. (1999). Item Response Theory. In Masters, G.N. and Keeves, J.P. (Eds). Advances in Measurement in Educational Research and Assessment. Oxford: Pergamon. pp. 55-63.

Syahri, I. (2007). Standardized testing in Indonesian secondary education: An analysis on the impact of the national exit examination (2005-2007).

Thomas, N. (2000). Assessing model sensitivity of the imputation methods used in the National Assessment of Educational Progress. Journal of Educational and Behavioral Statistics, 25(4), 351-371.

The American Heritage® Dictionary of the English Language, (4th ed.). Retrieved on 31 May, 2007, from http://www.answers.com/topic/measure.

Tognolini, J. (2006). Meeting the challenge of assessing in a standards based education system, Curriculum Council: Western Australia.

Tognolini, J. & Andrich, D. (1996) Analysis of profiles of students applying for entrance into universities. Applied Measurement in Education, 9(4), 323-353.

UNESCO. (1995). National profiles in technical and vocational education in Asia and the Pacific: Indonesia. Retrieved 20 March, 2008, from http://unesdoc.unesco.org/images/0010/001049/104925E.pdf.

Van der Ree, R. (2007). The Education System in the Netherlands. Dutch Eurydice Unit, The Hague.

Venter, E. J. (2008). The implication of ignoring basic assumptions in a testing situation. Unpublished.

World Development Indicators. (2007). Indonesia. Retrieved 1 June, 2008, from http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20535285~menuPK:1192694~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html.

Wright, B. D. & Stone, M. H. (1979). Best Test Design: Rasch Measurement. Chicago: Mesa Press.

Willmott, Alan S. 1977. CSE and GCE Grading Standards: The 1973 Comparability Study. London: Macmillan.

World Bank. (2000). Examination profile report: Indonesia. Retrieved 20 March, 2008, from http://www1.worldbank.org/education/exams/database/ExamProfileRpt.asp?ExamCode=337&Statecode=54.

World Bank (2001a). The nature of public examination systems. Retrieved 20 March, 2008, from http://www1.worldbank.org/education/exams/nature.asp.

World Bank (2001b). Purposes and functions. Retrieved 20 March, 2008, from http://www1.worldbank.org/education/exams/purposes.asp.

World Bank. (2006). Project performance assessment report: Indonesia. Retrieved 20 March, 2008 http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB

/2006/08/10/000160016_20060810124747/Rendered/PDF/36427.pdf.

World Education Services. (2008). Indonesia's system of education. Retrieved 11 March, 2008, from http://www.wes.org/ewenr/00july/practical.htm.

# Appendix A: Overview of Rasch Measurement

This section provides a more detailed technical discussion about the Rasch model and the requirements for the model. Repetition of central features discussed in Section 2 may occur for the purposes of logical coherence. This introduction is an excerpt from Venter (2008).

## Measurement

Great care is taken in the physical world where fundamental measurement is concerned. Finely calibrated instruments are used to measure, for example, the volume of fuel in a vehicle, the temperature at which a cake should be baked or the levels of certain vitamins in the human body. The same rigour is absent however, when educational research or a psychological investigation is undertaken (Bond & Fox, 2007). Linacre (as cited in Bond & Fox, 2007) positions the current interaction of human sciences with measurement as at the same descriptive level of physical science that was prevalent prior to the publication of Isaac Newton's Philosophiae Naturalis Principia Mathematica in 1687. Bond and Fox (2007) are also of the opinion that:

> quantitative researchers in the human sciences are too narrowly focused on statistical analysis, and not concerned nearly enough about the quality of the measures on which they use these statistics.

Stevens (as cited in Bond & Fox, 2007) defined measurement as the assigning of numbers to objects or events in terms according to a rule and that some form of measurement exists at nominal, ordinal, interval and ratio level. Some other physical scales of measurement are, of course, not merely a sequence of concatenated units, for example, the density of two separate litres of water added together does not add up to the sum of the densities. The density scale is an example of a derived scale.

Luce and Tukey (as cited in Bond & Fox, 2007) argued for another type of fundamental measurement called simultaneous conjoint measurement "that subsumed the existing categories of fundamental and derived measurement", essentially opening the way to measuring inter alia, psychological constructs (Bond & Fox, 2007). The important aspect in terms of additivity in the measurement structure is found in relationships in and between the cells of a data matrix.

Originally, Ben Wright (1979) suggested that an ordinary ruler that measured linear units is a good analogy for a measuring instrument in the social sciences, but his argument is that human traits cannot be concatenated. Bond and Fox (2007) are of the opinion that the measurement of temperature presents a better analogy for a measurement instrument in the human sciences. For example, 0°C would be an arbitrary point on the temperature scale, but would not constitute the total absence of temperature. Likewise in the case of a human trait, for instance, the ability to interact socially with members of a different gender would, with a score of 0 on a certain scale, not constitute no ability on the latent trait.

### Linear Measures

A variable on an ordinal measurement scale would have the characteristics of classification into different distinct and ordered categories in terms of a certain attribute on the one hand. On the other hand, these categories can possess more of that attribute in an ascending fashion (Huysamen, 1983). Although scores on such a variable could be added and subtracted, careful consideration must be given to the meaning of the total scores. If careful thought is given to raw scores, it becomes evident that they also only act as a device to order persons in ascending or descending order, because there is no evidence that the difference (or distance) between two points; for instance, on the lower part of the scale would be exactly the same as the difference between two points higher up on the scale. In other words, a person scoring 60 on a test has double the marks that a person scoring only 30 on the same test has, but it does not necessarily mean that the person has double the attribute of the other person.

The question arises if raw scores per se, can be realistically viewed as measures. According to The American Heritage® Dictionary of the English Language (2004) a measure can be loosely defined as: “Dimensions, quantity, or capacity as ascertained by comparison with a standard”. Wright and Linacre (1989) stated, “a ‘measure’ is a number with which arithmetic (and linear statistics) can be done, … yet with results that maintain their numerical meaning”. Measurement on an interval scale, on the other hand, would be able to provide a distinction between more or less of an attribute, but also provide for equal distances or differences between two points on the scale. A zero point on this scale does not indicate a total absence of an attribute (Glass & Stanley, 1970). Bond and Fox (2007) argue strongly for the same rigour in measurement in the physical sciences to be applied in the field of psychology. This proposed rigour in measurement should be extended also to the field of education in South Africa. The Rasch model provides an avenue to attain this goal.

## The Rasch model

Georg Rasch, a Danish mathematician, developed the Rasch model in the 1950s. It is a probabilistic model by which measures are created to be used in subsequent parametric tests. Through the years the Rasch model has been developed to include a family of models, not only addressing dichotomies, but also inter alia rating scale and partial credit models. The Rasch model has only item difficulty as a parameter. The Rasch model provides a “stringent modelling tool”, which is useful when the data fit the model and which will provide information about test inequity or differential item functioning when misfit is identified (Ryan & Williams, 2005). The Rasch model is appropriate for constructing a scale because of “the capacity to deal with missing information”, and therefore link “tests through common examinees, or sets of examinees through common test items” (Ryan & Williams, 2005).

### Assumptions

One of the basic assumptions of the Rasch model is that a relatively stable latent trait underlies test results (Boone & Rogan, 2005). For this reason the model is also sometimes called ‘the latent trait model’. A latent trait or construct is an underlying, unobservable characteristic of an individual (Hambleton & Swaninamathan, 1999) that cannot be directly measured, but will explain scores attained on a specific test pertaining to that attribute (Ryan, 1983). Unidimensionality is the term used for the focus on one attribute or dimension at a time (Bond & Fox, 2007). An example of a latent trait is, for instance, a student’s attitude towards taking an examination.

Through the application of this model, raw scores undergo log transformations that render an interval scale where the intervals are equal, expressed as log odds units or logits. The Rasch model may also be the only model whereby a scale can be constructed that is separable or invariant to the abilities of the persons tested (Bond & Fox, 2007). They also mention that the Rasch model is “the only model to date that provides tools for approximating objective reproducible additive measures in the human sciences”.

### Dichotomous Rasch model [11]

The dichotomous Rasch model applies to items where a correct response is awarded a score of 1 and an incorrect response a score of 0. An example would be in the case of a multiple choice item, where a person n provides an answer to an item i and attains a score of $x_{ni}$, with the person’s ability $\beta_n$ and the item difficulty level of $\delta_i$.

Considering the analogy of a high jumper (as cited in Smith & Smith, 2004) and transposing the idea to that of item difficulty and person ability, one would expect a situation where, if a person’s ability (to jump) is higher than an item’s difficulty (the height of the crossbar), the person is expected to get such an item correct (high jumper succeeds to clear the crossbar) most of the time. In this case, the probability of a person getting the item correct can be expressed as follows:

[11] The whole logic and explanation of the derivation of formulas is based on Andrich and Marais (2006).

$$\textbf{If } (\beta_n - \delta_i) > 0 \ \textbf{ then } P\{x_{ni}=1\} > 0.5$$

If the person's ability (to jump) is much lower than the item difficulty (the height of the crossbar), one would expect that person to almost never get the answer correct (succeed in clearing the crossbar) and can be represented as follows:

$$\textbf{If } (\beta_n - \delta_i) < 0 \ \textbf{ then } P\{x_{ni}=1\} < 0.5$$

Where the person's ability (to jump) and the item's difficulty (height of the crossbar) coincide, the person is expected to answer the item correctly (clear the crossbar) only half of the time.

$$\textbf{If } (\beta_n - \delta_i) = 0 \ \textbf{ then } P\{x_{ni}=1\} = 0.5$$

Probability can range from 0 to 1:

$$0 \leq P\{x_i = 1\} \leq 1$$

The difference between a person's ability and an item's difficulty can range between - ∞ and + ∞:

$$-\infty \leq (\beta_n - \delta_i) \leq +\infty$$

The difference between ability and difficulty can be used as an exponent of base *e* and this expression will have limits of 0 and infinity, that is

$$0 \leq e^{(\beta_n - \delta_i)} \leq +\infty$$

An expression can be obtained through an adjustment whereby the limits are 0 and 1. The following expression could serve as the basis for a formula for a probability of a correct response:

$$0 \leq \left\{ \frac{e^{(\beta_n - \delta_i)}}{1 + e\,(\beta_n - \delta_i)} \right\} \leq 1$$

This formula now serves as an estimate of the probability of a correct response for person *n* on item *i* and the relationship is as follows:

$$P\{x_{ni}=1 \mid \beta_n, \delta_i\} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

The above formula is the one that Georg Rasch chose when he developed the latent trait test theory. It is a simple logistic function and the units are called "logits". The formula in a simpler form is used for the dichotomous Rasch model:

$$P_{ni} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

As an example, if a person with an ability of $\beta_n$=5 interacts with an item of difficulty $\delta_i$=2, the probability of the person answering the item correctly will be:

$$P\,\{\,X_{ni} = \,|\beta_n, \delta_i\} = \frac{e^{(5\text{-}2)}}{1 + e^{(5\text{-}2)}}$$

$$= \frac{e^{3}}{1 + e^{3}}$$

$$= \frac{20.086}{21.086}$$

$$= \mathbf{0.95}$$

**Table 11** is a table of more examples of the probabilities generated from differences between ability and difficulty.

**Table 11: Probabilities of correct responses for persons on items of different relative difficulties**

| $\beta_n - \delta_i$ | Probability |
|---:|---:|
| **3** | 0.95 |
| **2** | 0.88 |
| **1** | 0.73 |
| **0** | 0.50 |
| **-1** | 0.27 |
| **-2** | 0.12 |
| **3** | 0.05 |

One can generate many more probabilities from many such differences and then represent the resulting function graphically. This graph is also known as the item characteristic curve. **Figure 8** displays the function graphically.

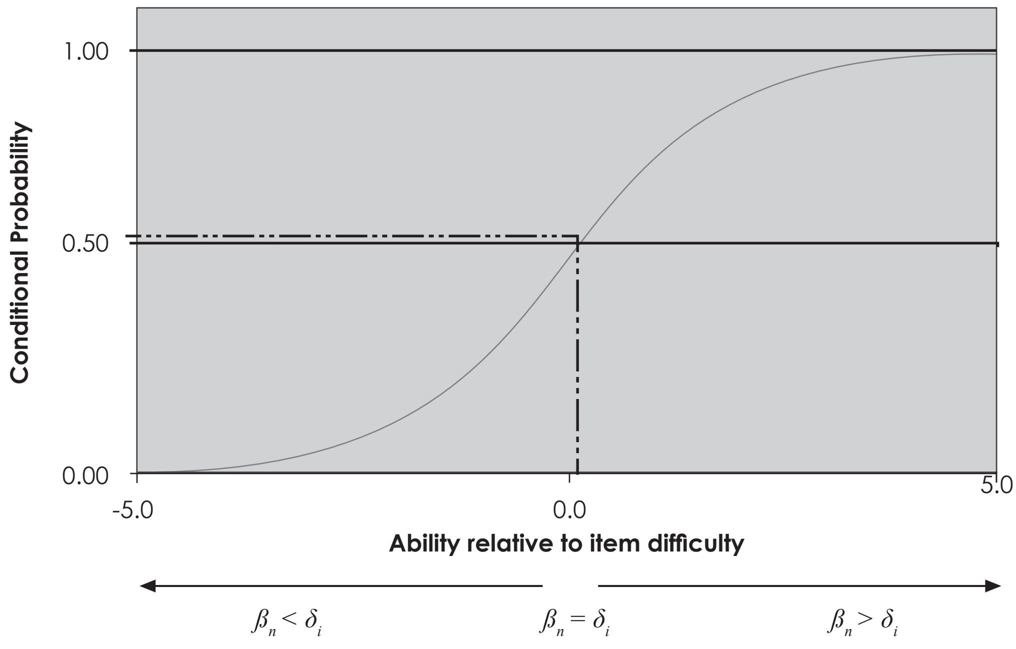

*Figure 8: Function of the dichotomous Rasch model*

The item characteristic curve provides the opportunity to directly establish the probability of a person of ability $\beta$ answering an item of difficulty $\delta$ correctly. For example, if in Figure 8 a person with ability $\beta=0.0$ interacts with an item of difficulty $\delta = 0.0$ the probability is 50% that the answer will be correct (see dotted line on graph).

## Estimating abilities and difficulties

Andrich and Marais (2006) refer to the sufficient statistics for the estimates of person ability and item difficulty. All the item scores for a person are added together to create a total person score: $r_n = \sum_{i=1}^{k} x_{ni}$ with k the number of items, is the sufficient statistic for the person ability estimate. The total item score, $s_i = \sum_{n=1}^{N} x_{ni}$ is the sufficient statistic for the item difficulty estimate, in other words, all the information about $\beta_n$ and $\delta_i$ is contained in the respective total scores as indicated above. The estimation of person's ability will be explained with an example in the following paragraph.

Louis Guttman devised the Guttman scale (also called a 'scalogram') in 1944 (Bond & Fox, 2007). Essentially this is a data matrix where the items are ranked from easy to difficult and the persons likewise are ranked from lowest achiever on the test to highest achiever on the test. Table 11 is an example of a scalogram adapted from Bond and Fox (2007). The data in the scalogram is a subset of nine items and 12 students from the Chemistry data set still to be discussed in the Data section. Student S4 is the best performing student and Student S12 performed the worst on this test. Items 2, 9 and 1 are the easiest items and Item 3 is the most difficult item.

**Table 12: Ordered data matrix**

| | | | | | Items | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Students** | **2** | **9** | **1** | **8** | **5** | **7** | **4** | **6** | **3** | **Ability (most)** | $\frac{n}{N}\%$ |
| **S4** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 8 | 89 |
| **S6** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 7 | 78 |
| **S7** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | ***1*** | 7 | 78 |
| **S3** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 6 | 67 |
| **S1** | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 5 | 56 |
| **S2** | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 5 | 56 |
| **S5** | 1 | 1 | 1 | 0 | 0 | 0 | 0 | ***1*** | 0 | 4 | 44 |
| **S8** | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 4 | 44 |
| **S9** | 1 | ***0*** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 44 |
| **S10** | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 33 |
| **S11** | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 33 |
| **S12** | 1 | 1 | 0 | 0 | ***1*** | 0 | 0 | 0 | 0 | 3 | 33 |
| | | | | | | | | | | (Least) | |
| **Difficulty** | 11 | 11 | 11 | 7 | 6 | 6 | 3 | 3 | 1 | | |
| | *(Least)* | | | | | | | | *(Most)* | | |
| | 92 | 92 | 92 | 58 | 50 | 50 | 25 | 25 | 8 | | |
| $\frac{n}{N}\%$ | | | | | | | | | | | |

Let us consider the responses of Student S3 in **Table 12**.

**Table 13: Probabilities of responses of Student S3 to 9 items**

| **Random variable** | $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ | $x_{n5}$ | $x_{n6}$ | $x_{n7}$ | $x_{n8}$ | $x_{n9}$ | **Total Score** $(r_n)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Observed value** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 6 |
| **Average** $\overline{X}_{ni}$ | 0.92 | 0.88 | 0.88 | 0.82 | 0.73 | 0.62 | 0.50 | 0.38 | 0.27 | 6.00 |
| **Probability** $P_{ni}$ | 0.92 | 0.88 | 0.88 | 0.82 | 0.73 | 0.62 | 0.50 | 0.38 | 0.27 | 6.00 |

In **Table 13** the correct/incorrect responses by Student S3 to each item is given in the row labelled 'Observed value'. It is for the sake of this argument assumed that the ability of Student S3 is $\beta_n = 0.5$ and the difficulties of the items are $\delta_1 = -2.00; \delta_2 = -1.50; \delta_3 = -1.45; \delta_4 = -1.00; \delta_5 = -0.50; \delta_6 = 0.00;$ $\delta_7 = 0.5; \delta_8 = 1.0; \delta_9 = 1.5.$ We now imagine the average score of each item as if the person interacted many times with the same item. If one assumes that memory about the item does not play a role, the probability of the person answering each item correctly would be the average score of many repetitions of the item by the person. In reality this is not possible, but the reasoning is used to build an equation to estimate each person's ability and each item's difficulty from single responses of many people on many items. The probabilities over the nine items add up to 6. Algebraically this can be written as:

$$r_n = \sum_{i=1}^{9} x_{ni} = \sum_{i=1}^{9} P_{ni} \qquad \textbf{(1)}$$

Because the fundamental Rasch equation, $P_{ni} = \frac{e^{(\beta_n - \delta_i)}}{1+e^{(\beta_n - \delta_i)}}$, expresses the probability that a person obtains a correct score in terms of the person's ability and items difficulties, equation **(1)** can be written as:

$$r_n = \sum_{i=1}^{9} \frac{e^{\left(\beta_n - \delta_i\right)}}{1+e^{\left(\beta_n - \delta_i\right)}} \qquad \textbf{(2)}$$

$$= \frac{e^{\beta_n - \delta_1}}{1+e^{\beta_n - \delta_1}} + \frac{e^{\beta_n - \delta_2}}{1+e^{\beta_n - \delta_2}} + \ldots + \frac{e^{\beta_n - \delta_9}}{1+e^{\beta_n - \delta_9}}$$

If the difficulty values of the items are known, one could solve for $\beta_n$ (or $\beta_{S3}$ in this case) from the above-mentioned equation. An iterative process is now entered into to solve this equation[12]. Based on experience, an initial value of $\beta_n^{(0)} = 0.25$ is chosen, substituted into the equation and if the value is less than $r_n$, it stands to reason that the initial $\beta_n$ -value is too small and should be increased. If the value is far greater than the initial $\beta_n$ -value, this value should be increased. The iteration procedure stops when a predetermined criterion is reached, for instance if the value differs with less than 0.001 from $r_n$.

As an example, the initial value of $\beta_{S3}^{(0)} = 0.25$ is used as well as the difficulty values mentioned above. Substituting these values into equation (2) produce the following value:

[12] The development of computers and appropriate software has greatly enhanced the application of the Rasch model since calculations can now be done with speed and accuracy.

$$\sum_{i=1}^{9}\frac{e^{\left(\beta_{S3}-\delta_i\right)}}{1+e^{\left(\beta_{S3}-\delta_i\right)}}=\frac{e^{\beta_{S3}^{(0)}-\delta_1}}{1+e^{\beta_{S3}^{(0)}-\delta_1}}+\frac{e^{\beta_{S3}^{(0)}-\delta_2}}{1+e^{\beta_{S3}^{(0)}-\delta_2}}+\ldots+\frac{e^{\beta_{S3}^{(0)}-\delta_9}}{1+e^{\beta_{S3}^{(0)}-\delta_9}}$$

$$=\frac{e^{\beta_{S3}^{(0)}-\delta_1}}{1+e^{\beta_{S3}^{(0)}-\delta_1}}+\frac{e^{\beta_{S3}^{(0)}-\delta_2}}{1+e^{\beta_{S3}^{(0)}-\delta_2}}+\ldots+\frac{e^{\beta_{S3}^{(0)}-\delta_9}}{1+e^{\beta_{S3}^{(0)}-\delta_9}}$$

$$=\frac{e^{0.25+2.00}}{1+e^{0.25+2.00}}+\frac{e^{0.25+1.50}}{1+e^{0.25+1.50}}+\ldots+\frac{e^{0.25-1.50}}{1+e^{0.25-1.50}}$$

$$=\mathbf{0.90+0.85+0.85+0.78+0.68+0.56+0.44+0.32+0.22=5.60}$$

Student S3, with an ability of $\beta_{S3}^{(0)} = 0.25$, is expected to obtain a score of 5.60, but actually this person has a score of 6.00. A higher ability estimate, for instance $\beta_{S3}^{(1)} = 0.35$, could be substituted into equation **(2)** and the value should be compared to 6.00. This iteration process will continue until an acceptable $\hat{\beta}_{S3}$ - value is reached.

The argument for obtaining difficulty estimates for the items is along the same lines as that of the person ability estimates.

## Polytomous Rasch model

The Greek meaning of the word ‘polytomous’ is literally ‘many cuts’ and is used to indicate the rating scale and partial credit models in Rasch.

## Rasch-Andrich rating scale model

David Andrich (1978) in Linacre (2007) in a conceptual breakthrough, comprehended that a rating scale, for example, a Likert-type scale, could be considered as a series of Rasch dichotomies. According to Linacre (2007), the Rasch-Andrich Rating Scale Model specifies the probability, $P_{nix}$ , that person n of ability $\beta_n$ is observed in category $x$ of a rating scale applied to item $i$ with difficulty level $\delta_i$ as opposed to the probability $P_{ni\,(x-1)}$ of being observed in category $(x-1)$. In a Likert scale, $x$ could represent "Strongly Agree’ and $(x-1)$ would then be the previous category ‘Agree’. Mathematically the function is depicted as follows:

$$\ln\left(\frac{P_{nix}}{P_{ni(x-1)}}\right)=\beta_n-\delta_i-\tau_x$$

Linacre (2007) makes the point that similar to the Rasch original dichotomous model, a person’s ability or attitude is represented by $\beta_n$ whereas $\delta_i$ is the item difficulty or the “difficulty to endorse”. The difficulty or endorsability value is the “balance point” of the item according to Bond and Fox (2007) and is situated at the point where the probability of observing the highest category is equal to the probability of observing the lowest category (Linacre, 2007).

In the Rasch-Andrich rating scale, a Rasch-Andrich threshold, $\tau_x$ , is also located on the latent variable. This ‘threshold’ or ‘step’ is according to Linacre (2007) “the point on the latent variable (relative to the item difficulty) where the probability of being observed in category, $x$, equals the probability of being observed in, the previous category, $x-1$”. A threshold is in other words the ‘transition’ between two categories. Wright and Mok (as cited in Smith & Smith, 2004) are of the opinion that if Likert scale items have the same response categories, that it is quite reasonable to assume that the thresholds would be the same for all items.

## Partial credit model

The partial credit model applies for instance to achievement items where marks are allocated for partially correct answers or where a sequence of tasks has to be completed. Essentially, the partial credit model is the same as the rating scale model, with the only difference being that in the partial credit model, each item has its own threshold parameters. The threshold parameter, $\tau_x$ , in the partial credit model becomes $\tau_{ix}$ and mathematically the model in Smith and Smith (2004) changes to:

$$\ln\left(\frac{P_{nix}}{P_{ni(x-1)}}\right) = \beta_n - \delta_i - \tau_x$$

## Model fit

In order to discuss model fit, **Table 12** with observed values will be considered again. As mentioned previously, a certain pattern of responses is to be expected. The probability is high that a low performing person (see Student S12) would only answer the easy questions correctly and also that the higher performing persons (see Students S4 and S6) will be able to answer the more difficult questions correctly. In the 'middle' area, the probability is that a person answers items correctly only half of the time (see Student S1). The Rasch model assumes this pattern of responses. The area where the correct answers meet the incorrect answers (shaded) provides for some unpredictability (Bond & Fox, 2007). Although one could have argued that Student S1 should have answered Item 5 correctly, it is not realistic to expect a perfect transition zone from 1s to 0s and therefore is not of much concern.

The bold italicised responses in **Table 12** highlight unexpected (observed) responses for students and/or items. According to Bond and Fox (2007), the number of unexpected responses and their position in the scalogram will determine the seriousness of the concern about them. Student S7, for instance, probably guessed the answer to the more difficult Item 3.

The Guttman scalogram provides the basis for determining the fit of the data to the model. From the ordered data matrix in **Table 12**, a similar data matrix can be constructed by calculating the expected response value for each person on each item. This is done by substituting each corresponding pair of ability and difficulty value into the Rasch model equation. Subsequently, the response residual for each cell is calculated by determining the difference between the observed score and the expected score. A matrix of response residuals can now be constructed. A problem frequently experienced in statistics is that residuals can be either negative or positive and mere addition of the residuals will be 0.0. The residuals are therefore squared to render a positive result. A standardised residual *($Z_{ni}$)* is calculated by dividing the raw residual by its standard deviation and fit statistics in the Rasch model are based on these (Bond & Fox, 2007). It becomes less likely for the item or person to fit the model as the standardised residual gets greater. The standardised residual can be algebraically expressed as follows:

$$z_{ni} = \frac{x_{ni} - E_{ni}}{\sigma_{ni}}$$

An approximate $\chi$-distribution is obtained when the standardised residuals are squared and summed. This value can be compared to critical values of a $\chi$- distribution with the specific degrees of freedom (Andrich & Marais, 2006).

The extent to which the data fit the model is expressed through either the outfit mean square "outlier sensitive mean square residual goodness of fit statistic" or the infit mean square "information weighted mean square residual goodness of fit statistic".

The outfit mean square statistic is an unweighted version of the fit statistic (Wright & Master, 1982 in Smith & Smith, 2007); in other words, it is simply an average of the standardised residual and not multiplied or influenced by other information. The formula is:

$$\text{outfit} = \frac{\sum Z_{ni}^2}{N}$$

This estimate is more sensitive to outliers or unexpected responses far from a person's ability or an item's difficulty measures.

The infit mean square statistic gives relatively more impact to unexpected responses close to a person's ability level or an item's difficulty level, because it is weighted by its variance $(W_{ni})$. This is done to counter the influence of unexpected responses far from the specific measure. The formula for the infit mean square statistic is:

$$\text{infit} = \frac{\sum Z_{ni}^2 W_{ni}}{\sum W_{ni}}$$

The expected value of the infit and outfit mean square (MnSq) statistics is 1.0. If, for example, a person would guess the answer to a difficult item correctly (one that the person should really get wrong) the outfit statistic would be much larger than 1.0 because it is sensitive to outliers.

Items and persons in this study were deemed misfitting when the outfit mean square statistic fell outside the range of 0.5 to 1.5, because in this range the fit is productive for measurement (Linacre, 2007). Linacre (2007) also suggests that stricter criteria can be used and that the range for multiple choice items should be between 0.7 and 1.3 and for rating scale items between 0.6 and 1.4. In Bond and Fox (2007) it is suggested that the sample size should also play a role in determining unaccep-table departures from the model by evaluating the misfit statistics.

Where the values are less than 0.5, too much predictability or overfit is experienced and when the value exceeds 1.5, too much noise was present in the data or a situation of underfit existed. Mean-square statistics indicate the size of the misfit, but the "significance" of the improbability of the misfit is also important, and this is indicated by the standardised residual as explained in a previous paragraph. Correspondingly therefore to each mean-square a $z_{ni}$- statistic shows the probability of the mean-square as a unit-normal deviate whereas absolute values of 2 or more indicate statistically significant model misfit (Andrich & Marais, 2006; Linacre, 2007).

## Sample invariance

In a physical measurement environment Bond and Fox (2007) state that: "the values attributed to variables by any measurement system should be independent of the particular measurement instrument used".

Intuitively, in a social science testing situation, one would expect that the difficulty value of an item should be intrinsic to the item and not dependent on the persons writing the test, if the test was properly targeted at the group; similarly, a person's ability on the underlying construct should not be dependent on or influenced by the specific test that is written. Andersen (1977, in Smith & Smith, 2004) points out that the Rasch family of models are the only latent trait models that provide sufficient statistics (in this case cumulative total raw scores) to estimate item and person parameters. He calls it the separability of item and person estimates. According to Wright and Masters (as cited in Smith & Smith, 2004), because of this separability, the person ability estimates are "freed from the distributional properties of the specific items" and vice versa.

37 General Van Ryneveld Street, Persequor Technopark, Pretoria
Telephone: +27 12 349 1510 • Fax: +27 12 349 1511
E-mail: Info@umalusi.org.za • Web: www.umalusi.org.za

UMALUSI
Council for Quality Assurance in
General and Further Education and Training