CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

INTRODUCTION

In this chapter, I describe how I went about investigating my research questions (posed in section 3.2). I explain how I moved from an informal position, based on my observations and interpretation over many years as a mathematics lecturer of undergraduate students, to a formal research-oriented position. By speaking of 'how' I moved, I am referring to my methods of doing formal research and collecting 'relevant' data, and to my justification for the appropriateness of these methods. These methods, together with their motivations and characterisations, constitute the methodology of my research.

Initially, in section 3.1 the research design is described. This is followed by my research questions formulated in section 3.2. Section 3.3 outlines the qualitative research methodology of the study in which the interviews with the sample of undergraduate students are described. In section 3.4, the quantitative research methodology is discussed. In this section the Rasch model, the particular statistical method employed, is described. Lastly, issues related to reliability, validity, bias and ethics are discussed in section 3.5.

3.1 RESEARCH DESIGN

According to Burns and Grove (2003), the purpose of research design is to achieve greater control of the study and to improve the validity of the study by examining the research problem. In deciding which research design to use, the researcher has to consider a number of factors. These include the focus of the research (orientation of action), the unit of analysis (the person or object of data collection) and the time dimension (Bless & Higson-Smith, 1995).

Research designs can be classified as either *non experimental* or *experimental*. In non experimental designs the researcher studies phenomena as they exist. In contrast, the various experimental designs all involve researcher intervention (Gall, Gall & Borg, 2003). This research study is non experimental in design, and as the purpose of this study is prediction, a *correlational research design* is used. Correlational research refers to studies in which the purpose is to discover relationships between variables through the use of correlational statistics. The basic design in correlational research is very simple, involving collecting data on two or more variables for each individual in a sample and computing a correlation coefficient.

Many studies in education have been done with this design. As in most research, the quality of correlational studies is determined not by the complexity of the design or the sophistication of analytical techniques, but by the depth of the rationale and theoretical constructs that guide the research design. The likelihood of obtaining an important research finding is greater if the researcher uses theory and the results of previous research to select variables to be correlated with one another (Gall, Gall & Borg, 2003).

Correlational research designs are highly useful for studying problems in education and in the other social sciences. Their principal advantage over causal-comparative or experimental designs is that they enable researchers to analyse the relationships among a large number of variables in a single study. In education and social sciences, we frequently confront situations in which several variables influence a particular pattern of behaviour. Correlational designs allow us to analyse how these variables, either singly or in combination affect the pattern of behaviour.

In this study, first year Mathematics Major students from the University of the Witwatersrand were selected from the MATH109 course and their performance on assessment in the PRQ format was compared to their performance on assessment in the CRQ format. In addition, students were asked to indicate a confidence of response corresponding to each test item, in both the CRQ and

PRQ assessment formats. Further data was collected from experts who indicated their opinions of the difficulty of the test items, both PRQs and CRQs, independent of the students' performance in each question. Further discussion on the research methodology is presented in section 3.4.

3.2 RESEARCH QUESTIONS

The objective of this research study is to design a model to measure how good a mathematics question is and to use the proposed model to determine which of the mathematics assessment components can be successfully assessed with respect to the PRQ format, and which can be successfully assessed with respect to the CRQ format.

To meet the objective of the study described above, the study will be designed according to the following steps:

[1] Three measuring criteria are used to develop a model for determining the quality of a mathematics question (the QI model).

[2] The quality of all PRQs and CRQs are determined by means of the QI model.

[3] A comparison is made within each assessment component between PRQ and CRQ assessment.

Based on these design steps and having defined the concept of a *good mathematics question*, the research question is formulated as follows:

Research question:

Can we successfully use PRQs as an assessment format in undergraduate mathematics?

In order to answer the research question, the following subquestions are formulated:

Subquestion 1:

How do we measure the quality of a good mathematics question?

Subquestion 2:

Which of the mathematics assessment components can be successfully assessed using the PRQ assessment format and which of the mathematics assessment components can be successfully assessed using the CRQ assessment format?

Subquestion 3:

What are student preferences regarding different assessment formats?

3.3 QUALITATIVE RESEARCH METHODOLOGY

Qualitative research in education has roots in many academic disciplines (Cresswell, 2002). Some qualitative researchers also have been influenced by the postmodern approach to inquiry that has emerged in recent years (Angrosino & Mays de Pérez, 2000; Merriam, 1998).

Cresswell (1998, p150) lists the advantages of using qualitative research methodology as follows:

- Qualitative research is value laden
- The researcher has firsthand experience of the participant during observation
- Unusual aspects can be noted during observation
- Information can be recorded as it occurs during observation
- It saves the researcher transcription time
- The researcher can control the line of questioning in an interview
- The participants can provide historical information.

3.3.1 Qualitative data collection

Purpose of the interviews

The purpose of the interviews was to probe MATH109 students' beliefs, attitudes and inner experiences about the different assessment formats they had been exposed to in their tests and examinations. The task in the interviews was designed with a research purpose; my responses (as interviewer) were more geared to finding out what the student was thinking (the research role) rather than assisting (the teacher role). The very fact that I was present at the interviews must also have affected the thinking and responses of the students that were being interviewed.

The qualitative data will be used to address the third research subquestion of what student preferences are regarding different assessments formats.

Interviews

The interviews were structured along certain dimensions, and semi-structured along others. It was structured in that all students were asked exactly the same set of predetermined questions (see page 88 for the questions); it was semi-structured in that my responses and prompts, as interviewer, depended to a large extent on the responses of the interviewee and on my relationship with that particular student. As the interviewer, I strove for consistency on certain dimensions in all interviews. Each interview was framed by the same set of questions and timeframe which provided a type of structure to the interview.

Despite these commitments to a measure of consistency, the clinical interviews in this study (as in other educational research type studies) are necessarily not neutral. This is because clinical interviews, just like any other learner-teacher engagement, are social productions. In this regard, Minick, Stone and Forman (1993) assert:

> Educationally significant human interactions do not involve abstract bearers of cognitive structures but real people who develop a variety of interpersonal relationships with one another in the course of their shared activity in a given institutional context. … For example, appropriating the speech or actions of another person requires a degree of identification with that person and cultural community he or she represents (p6).

I was able to engage far more effectively with some students rather than others in the interview situations (in the sense of being able to generate more penetrative probes). For example, with certain students whose home language is not English, much of my time was spent on interpreting what they said.

Format of the interviews

Nine MATH109 students with various gradings (weak/average/good) based on their June class record marks, from different racial backgrounds and different gender classes were interviewed, one at a time over a period of about two weeks in October 2004. Each interview took place in my office and was tape recorded and later transcribed. The maximum duration of each interview was 30 minutes. Table 3.1 lists the MATH109 student interviewees and their academic backgrounds.

[A: ≥75%; B: 70-74%; C: 60-69%; D: 50-59%; Fail: <50%]

Table 3.1: MATH109 student interviewees and their academic backgrounds.

INTERVIEWEE	October Class record [%]	Exam (%)	Final (%)	Symbol
[1]	70.05	32.77	51.41	D
[2]	80.67	85	82.84	A
[3]	81.26	81	81	A
[4]	58.11	29.16	43.64	Fail
[5]	59.43	53.33	56.38	D
[6]	42.92	26.28	34.65	Fail
[7]	68.28	44.44	56.36	D
[8]	74.48	82.22	78.35	A
[9]	36.57	31.11	33.84	Fail

At the commencement of the interview, I reminded each student that I was doing research to probe their beliefs, attitudes and inner experiences about the different assessment formats they had been exposed to in their tests and examinations. My opening questions were to find out about the background of each student i.e. why they registered for Mathematics I Major; career choice etc. This seemed to put the student at ease and they found the situation less threatening. I then moved on to the ten interview questions.

Interview questions:

[1] I'm interested in your feelings about the different ways in which we asked questions in your maths tests, a percentage being multiple choice provided response questions and the other the more traditional open-ended constructed response questions. Do you like the different formats of assessment?

[2] Why / Why not?

[3] Which type of question do you prefer in maths?

[4] Why do you prefer type A to type B?

[5] Which type of questions did you perform better in? Why?

[6] Do you feel that the mark you got for the MCQ sections is representative of your knowledge? What about the mark you got for the traditional long questions? Do you feel this is representative of your knowledge?

[7] Do you have confidence in answering questions in maths tests which are different to the traditional types of questions? Elaborate.

[8] What percentage of the maths tests do you recommend should be multiple-choice questions, and what percentage should be open-ended long questions?

[9] How would you ask questions in maths tests if you were responsible for the course?

[10] Is there opportunity for cheating in these different formats of assessment? Please tell me about them.

After asking these ten questions, I concluded the interview by asking each student if they had anything else to add or if they had any questions for me.

Examples of responses will be given and discussed in greater detail in the qualitative data analysis presented in section 4.1.

3.4 QUANTITATIVE RESEARCH METHODOLOGY

According to McMillan and Schumacher (2001), quantitative research involves the following:

- Explicit description of data collection and analysis procedures
- Scientific measurement and statistics used
- Deductive reasoning applied to numerical data
- Statements of statistical relevance and probability.

The *Rasch model* was used as the quantitative research methodology in this study. It is a probabilistic model that estimates person ability and item difficulty (Rasch, 1960). Although it is common practice in the South African educational setting to use raw scores in tests and examinations as a measure of a student's ability, research has shown that misleading and even incorrect results can stem from an erroneous assumption that raw scores are in fact linear measures (Planinic, Boone, Krsnik & Beilfuss, 2006). Linear measures, as used in the Rasch model, on the other hand, are on an interval scale, where arithmetic and statistical techniques can be applied and useful inferences can be made about the results (Rasch, 1980).

3.4.1 The Rasch model

In the following poem written by Tang (1996), each verse highlights a different characteristic of the Rasch model: A model of probability; uniformity; sufficiency; invariance property; diagnosticity and ubiquity.

Poem: What is Rasch?

Rasch is a model of probability
that estimates person ability,
that estimates item difficulty,
that predicts response probability
nothing but a function of ability and difficulty.

Rasch is a model of uniformity
that places the values of person ability
and the values of item difficulty
on the same scale with no diversity.

Rasch is a model of sufficiency
that uses number right for estimating person ability
and count of correct responses for item difficulty;
that relates raw score to person ability
and response distribution to item difficulty
-- with no ambiguity.

Rasch is a model with invariance property
that fosters person-free estimation of item difficulty
and test-free estimation of person ability;
that frees difficulty estimates from sample peculiarity
and ability estimates from difference in test difficulty.

Rasch is a model with diagnosticity
that flags item away from unidimensionality,
or items with local dependency;
that identifies persons with response inconsistency,
or person or groups measured with inappropriacy;
that maintains construct fidelity and enhances test validity.

Rash is a model of ubiquity;
from educational assessment to sociology,
from medical research to psychology,
from item analysis to item banking technology,
from test construction to test equity….
-- nothing beats its utility and popularity.

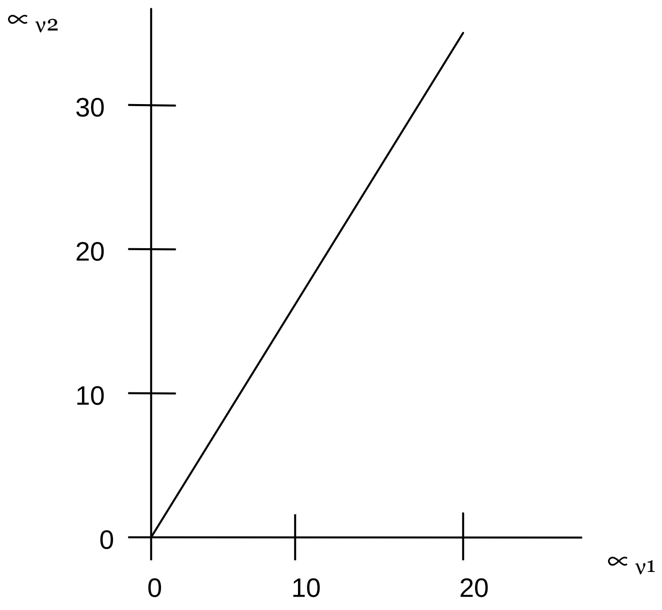
(Huixing Tang, 1996, p507)

3.4.1.1 Historical background

The Rasch model was developed during the years 1952 to 1960 by the Danish mathematician and statistician Georg Rasch (1901-1980). The development of the Rasch model took its beginning with the analysis of slow readers in 1952. The data in question were from children who had trouble reading during their time in school and for that reason were given supplementary education. There were several problems in the analysis of the slow readers. One was that the data had not been systematically collected. The children had for example not been tested with the same reading tests, and no effort had been made to standardise the difficulty of the tests. Another problem was that World War II had taken place between the two testings. This made it almost impossible to reconstruct the circumstances of the tests. It was therefore not possible to evaluate the slow readers by standardisation as was the usual method at the time (Andersen & Olsen, 1982).

Accordingly, it was necessary for Rasch to develop a new method where the *individual* could be measured independent of which particular reading test had been used for testing the child. The method was as follows: two of the tests that had been used to test the slow readers were given to a sample of school children in January 1952. Rasch graphically compared the number of misreadings in the two tests by plotting the number of misreadings in test 1 against the number of misreadings in test 2 for all persons. This is illustrated in Figure 3.1.

Figure 3.1: Number of misreadings of nine subjects in two tests.



(Source: Rasch ,1980)

The graphical analysis showed that, apart from random variations, the number of misreadings in the two tests was proportional for all persons. Further, this relationship held, no matter which pair of reading tests he considered.

To describe the random variation Rasch chose a Poisson model. The probability that person number v had misread α_{vi} words in test number i he accordingly modelled as

$$P(\alpha_{vi}) = e^{-\lambda_{vi}} \frac{(\lambda_{vi})^{\alpha_{vi}}}{\alpha_{vi}!} \qquad (1.1) \quad ; \quad \text{where}$$

λ_{vi} is the expected number of misread words.

Rasch then interpreted the proportional relationship between the number of misreadings in the two tests as a corresponding relationship between the parameters of the model, i.e.

$$\frac{\lambda_{v1}}{\lambda_{vi}} = \frac{\lambda_{01}}{\lambda_{0i}} \Leftrightarrow \lambda_{vi} = \frac{\lambda_{v1}}{\lambda_{01}} \lambda_{0i} = \theta_v \delta_i \qquad (1.2)$$

Thus the parameter of the model factorised into a product of two parameters, a *person parameter* θ_v and an *item parameter* δ_i. Inserting factorisation (1.2) in model (1.1), Rasch obtained the *multiplicative Poisson model*

$$P(\alpha_{vi}) = e^{-\theta_v \delta_i} \frac{(\theta_v \delta_i)^{\alpha_{vi}}}{\alpha_{vi}!} \qquad (1.3)$$

The way Rasch arrived at the multiplicative Poisson model was characteristic for his methods. He used graphical methods to understand the nature of a data set and then transferred his findings to a mathematical and a statistical formulation of the model.

The graphical analysis, however, was not Rasch's only reason to choose the multiplicative Poisson model. Rasch (1977) wrote:

> Obviously it is not a small step from Figure 1 [our Figure 3.1] to the Poisson distribution (1.1) with the parameter decomposition (1.2). I readily admit that I introduced this model with some mathematical hindsight: I realized that if the model thus defined was proven adequate, the statistical analysis of the experimental data and thus the assessment of the reading progress of the weak readers, would rest on a solid – and furthermore mathematically rather elegant – foundation.
>
> Fortunately the experimental result turned out to correspond satisfactorily to the model which became known as the multiplicative Poisson model (p63).

Rasch later developed the "elegant foundation" of the multiplicative Poisson model into a concept. Though in the beginning of the 1950s Rasch merely used it as a tool to estimate the ability of the slow readers by a method he called *bridge-building*. The point in using the bridge-building is that one can estimate the attainment of the individual regardless of which particular item the individual has been tested with. Bridge-building can be exemplified by the multiplicative Poisson model as follows:

Rasch writes that the main point of bridge-building is that it should be possible to assign to each item a degree of difficulty *that is independent of the persons the item has been applied to* (Rasch, 1960, pp20-22). This is possible in the multiplicative Poisson model, because the distribution of a person's responses to two different items conditioning on the sum of his responses only depends on the item parameters: $P(\alpha_{vi},\alpha_{vj}\left|\alpha_{vi}+\alpha_{vj};\theta_v,\delta_i,\delta_j\right.)=g(\delta_i,\delta_j)$. The person parameter, θ_v, is thus eliminated. Having estimated the item parameters in a distribution only depending on the item parameters, this estimate, $\hat{S}_i$, may be inserted in the distribution (1.3) giving

$$P(\alpha_{vi})=e^{-\theta_v\hat{S}_i}\,\frac{(\theta_v\hat{S}_i)^{\alpha_{vi}}}{\alpha_{vi}\,!} \qquad (1.4)$$

which only depends on the person parameter. Hence it is possible to estimate the parameter θ_v of the individual person even if only one item has been responded to. This is done by using a person's frequency of misreadings as an estimate of i and solving the equation (1.4) with regard to θ_v.

The way Rasch solved the problem of parameter separation for the slow readers was not the method he used later. But it represents the first trace of the idea of separating the estimation of item parameters from the estimation of person parameters.

In comparison to traditional analysis techniques, the Rasch model can be used (i) to analyse and improve a test instrument; and (ii) to generate linear (interval strength) learner scores, thus meeting the assumptions of parametric statistical tests such as t-tests and ANOVA (Birnbaum, 1968).

Rasch analysis has been the method of choice for moderate size data sets since 1965. Now the theoretical advantages and directly meaningful results of Rasch analysis can be easily obtained for large data sets, as follows:

- Scores and analyses dichotomous items, or sets of items with the same or different rating scale, partial credit, rank or count structures for up to 254 ordered categories per structure, with useful estimation of perfect scores.

- Missing responses or non-administered items are no problem.
- Analyse several partially linked forms in one analysis.
- Analyse responses from computer-adaptive tests.
- Item reports and graphical output include calibrations, standard errors, fit statistics, detailed reports of the particular improbable person responses which cause item misfit, distracter counts, and complete DOS files for additional analysis of item statistics.
- Person reports and graphical output include measures, standard errors, fit statistics, detailed reports of the particular improbable item responses which cause person misfit, a table of measures for all possible complete scores, and complete DOS files for additional analysis of person statistics
- Rating scale, partial credit, rank and count structures reported numerically and graphically.
- Complete output files of observations, residuals and their errors for additional analyses of differential item function and other residual analyses.
- Observations listed in conjoint estimate order to display extent of stochastic Guttman order. The Guttman scale (also called 'scalogram') 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.
- Option to pre-set and/or delete some or all person measures and/or item calibrations for anchoring, equating and banking, and also to pre-set rating scale step calibrations (Rasch, 1980).

The advantages of the Rasch model above other statistical procedures, used as the quantitative research methodology in this study, will be clarified further in section 3.4.1.4.

3.4.1.2 Latent trait

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'*.

Latent trait models focus on the interaction of a person with an item, rather than upon total test score (Wright & Stone, 1979). They use total test scores, but the mathematical model commences with a modelling of a person's response to an item. They are concerned with how likely a person v of an ability β_v on the 'latent trait' is to answer correctly, or partially correctly, an item i of difficulty δ_i. The latent trait or theoretical construct of concern to the tester is an underlying, unobservable characteristic of an individual which cannot be directly measured, but will explain scores attained on a specific test pertaining to that attribute (Andrich & Marais, 2006). For instance, in this study, the latent trait is the mathematical performance of first year tertiary students.

When items are conceived of as located, according to difficulty level, along a latent trait, the number of items a person answers correctly can vary according to the difficulties of the particular items included in the test. The relationship between person ability and total score is not linear. The non-linearity in this relationship means that test scores are not on an interval scale unless the items are evenly spaced in terms of difficulty. With a test designed according to the strategic of traditional test theory this would be unlikely to be the case because of the tendency to pick items clustered in the middle difficulty with only a few out towards the 0.8 and 0.2 levels of difficulty.

In latent trait models, the construct or latent trait is conceived as a single dimension along which items can be located in terms of their difficulty (δ_i) and persons can be located in terms of their ability (β_v).

If the person's ability β_v is above the item's difficulty δ_i we would expect the probability of the person observed in category x of a rating scale applied to item i being correct to be greater than 0.5, i.e.

if $(\beta_v - \delta_i) > 0,$ then $P\{\chi_{vi} = 1\} > 0.5$

If the person's ability is below the item's difficulty, we would expect the probability of a correct response to be less than 0.5, i.e.

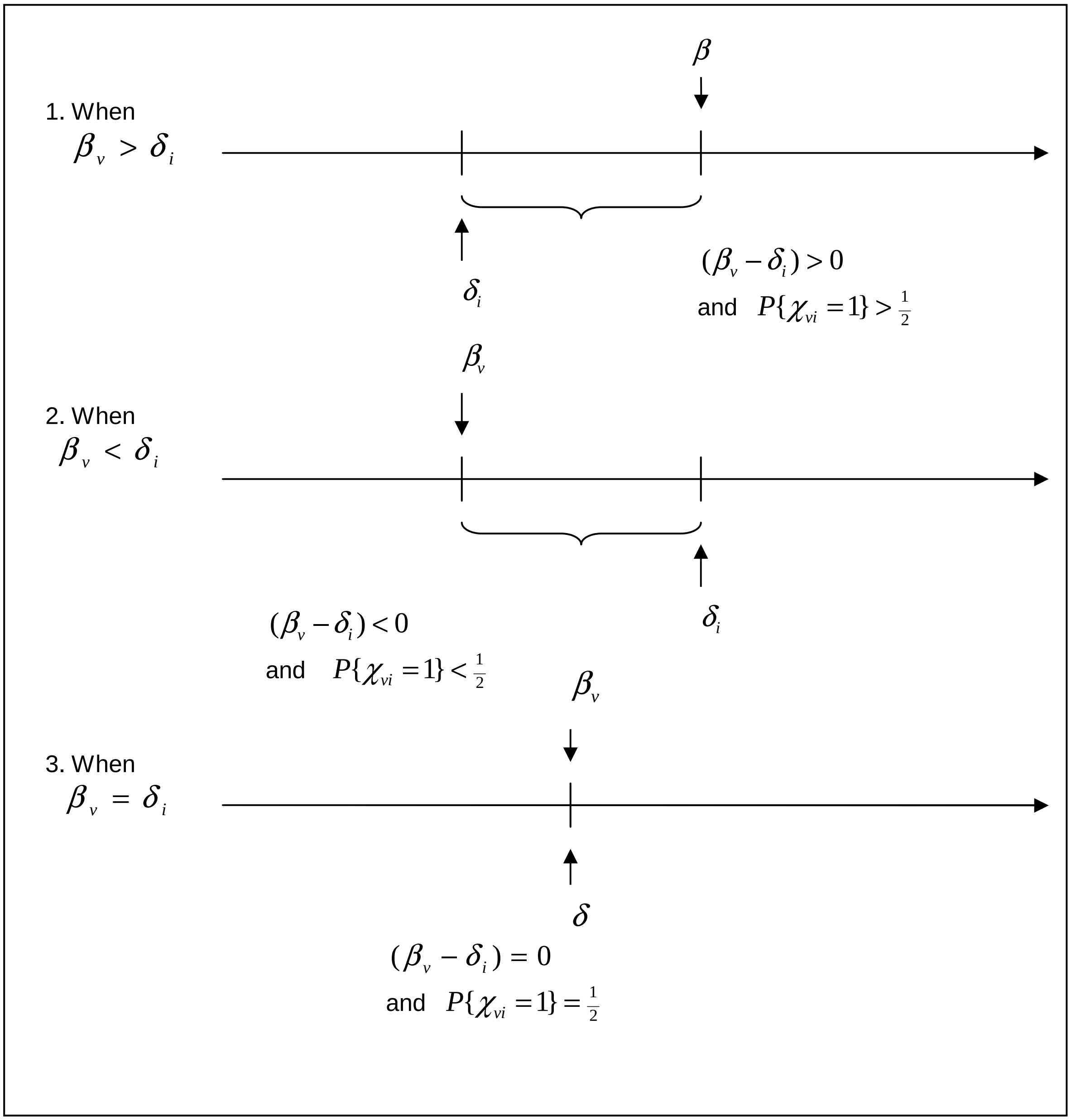
if $(\beta_v - \delta_i) < 0,$ then $P\{\chi_{vi} = 1\} < 0.5$

In the intermediate case where the person's ability and the item's difficulty are at the same point on the scale, the probability of a successful response would be 0.5 i.e.

if $(\beta_v - \delta_i) = 0,$ then $P\{\chi_{vi} = 1\} = 0.5$

Figure 3.2 illustrates how differences between person ability and item difficulty ought to affect the probability of a correct response.

Figure 3.2: How differences between person ability and item difficulty ought to affect the probability of a correct response.

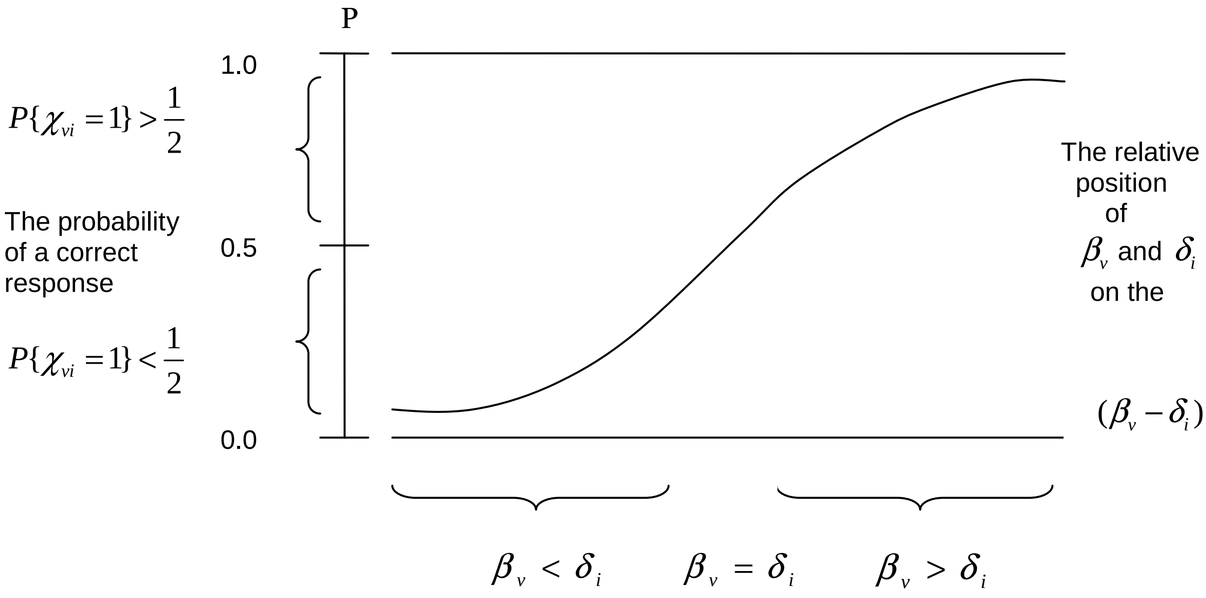

β
1. When
$\beta_v > \delta_i$
δ_i
$(\beta_v - \delta_i) > 0$
and $P\{\chi_{vi} = 1\} > \frac{1}{2}$
β_v
2. When
$\beta_v < \delta_i$
δ_i
$(\beta_v - \delta_i) < 0$
and $P\{\chi_{vi} = 1\} < \frac{1}{2}$
β_v
3. When
$\beta_v = \delta_i$
δ
$(\beta_v - \delta_i) = 0$
and $P\{\chi_{vi} = 1\} = \frac{1}{2}$

(Source: Andrich & Marais (2006), Lecture 5, p60).

The curve in Figure 3.3 summarises the implications of Figure 3.2 for all reasonable relationships between probabilities of correct responses and differences between person ability and item difficulty. This curve specifies the conditions a response model must fulfill. The difference $(\beta_v - \delta_i)$ could arise in 2 ways. It could arise from a variety of person abilities reacting to a single item, or it could arise from a variety of item difficulties testing the ability of one person.

When the curve is drawn with ability β as its variable so that it describes an item i, it is called an *item characteristic curve*, because it shows the way the item elicits responses from persons of every ability.

Figure 3.3: The item characteristic curve.


P
1.0
0.5
0.0
$P\{\chi_{vi}=1\}>\frac{1}{2}$
The probability of a correct response
$P\{\chi_{vi}=1\}<\frac{1}{2}$
The relative position of β_v and δ_i on the
$(\beta_v-\delta_i)$
$\beta_v < \delta_i$
$\beta_v = \delta_i$
$\beta_v > \delta_i$

$$P\{\chi_{vi}=1|\beta_v,\delta_i\} = f(\beta_v-\delta_i)$$

(Source: Andrich & Marais (2006), Lecture 5, p65).

In Figure 3.3 if we thought of the horizontal axis as the latent trait, the item characteristic curve would show the probability of persons of varying abilities responding correctly to a particular item. The point on the latent trait at which this probability is 0.50 would be the point at which the item should be located.

In order to construct a workable mathematical formula for the item characteristic curve in Figure 3.3, we begin by combining the parameters, β_v for person ability, and δ_i for item difficulty through their difference $(\beta_v-\delta_i)$. We want this difference to govern the probability of what is supposed to happen when person v uses their ability β_v against the difficulty δ_i of item i. But the difference $(\beta_v-\delta_i)$ can vary from minus infinity to plus infinity, while the probability of a successful response must remain between zero and one. That is

$$0 \leq P\{\chi_{vi} = 1\} \leq 1 \qquad (1)$$

$$-\infty \leq \beta_v - \delta_i \leq +\infty \qquad (2)$$

If we use the difference between ability and difficulty as an exponent of the base e, the expression will have the limits of zero and infinity. That is

$$0 \leq e^{(\beta v - \delta i)} \leq +\infty \qquad (3)$$

With a further adjustment we can obtain an expression which has the limits zero and one and therefore could perhaps be a formula for the probability of a correct response. The expression and its limits are:

$$0 \leq \frac{e^{(\beta v - \delta i)}}{1 + e^{(\beta v - \delta i)}} \leq 1 \qquad (4)$$

If we take this formula to be an estimate of the probability of a correct response for person ν on item i, the relationship can be written as:

$$P\{\chi_{vi} = 1 / \beta_{v,}\delta_i\} = \frac{e^{(\beta v - \delta i)}}{1 + e^{(\beta v - \delta i)}} \qquad (5)$$

The left hand side of (5) represents the probability of person v being correct on item i (or of the response of person v to item i being scored 1), given the person's ability β_v and the item's difficulty δ_i.

The function (5) which gives us the probability of a correct response is a simple logistic function. It provides a simple, useful response model that makes both linearity of scale and generality of measure possible. It is the formula Rasch chose when he developed the latent trait test theory. It is a simple logistic function. Rasch calls the special characteristic of the simple logistic function which makes generality in measurement possible *specific objectivity* (Rasch, 1960). He and others have shown that there is no alternative mathematical formula for the ogive curve in Figure 3.3 that allows estimation of the person measures β_v and the item calibrations δ_i independently of one another (Andersen, 1973, 1977; Birnbaum, 1968; Rasch, 1960, 1980).

3.4.1.3 Family of Rasch models

The responses of individual persons to individual items provide the raw data. Through the application of the Rasch model, raw scores undergo logarithmic transformations that render an interval scale where the intervals are equal, expressed as a ratio or log odd units or *logits* (Linacre, 1994). The Rasch model takes the raw data and makes from them item calibrations and person measures resulting in the following:

- valid items which can be demonstrated to define a variable
- valid response patterns which can be used to locate persons on the variable
- test-free measures that can be used to characterise persons in a general way
- linear measures that can be used to study growth and to compare groups (Bond & Fox, 2007).

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.

1. Dichotomous Rasch model

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 (PRQ), where a person v provides an answer to an item i and attains a score of χ_{vi}, with the person's ability β_v and the item difficulty level of δ_i. Formula (5) in a simpler form is used for the dichotomous Rasch model:

$$P_{vi} = \frac{e^{(\beta v - \delta i)}}{1 + e^{(\beta v - \delta i)}}$$

As discussed before, this formula is a simple logistic function and the units are called 'logits'.

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

$$P\{\chi_{vi} = 1 | \beta_v, \delta_i\} = \frac{e^{(5-2)}}{1 + e^{(5-2)}}$$

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

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

$$= 0.95$$

Table 3.2 is a table of more examples of the probabilities generated from differences between ability and difficulty.

Table 3.2: Probabilities of correct responses for persons on items of different relative difficulties.

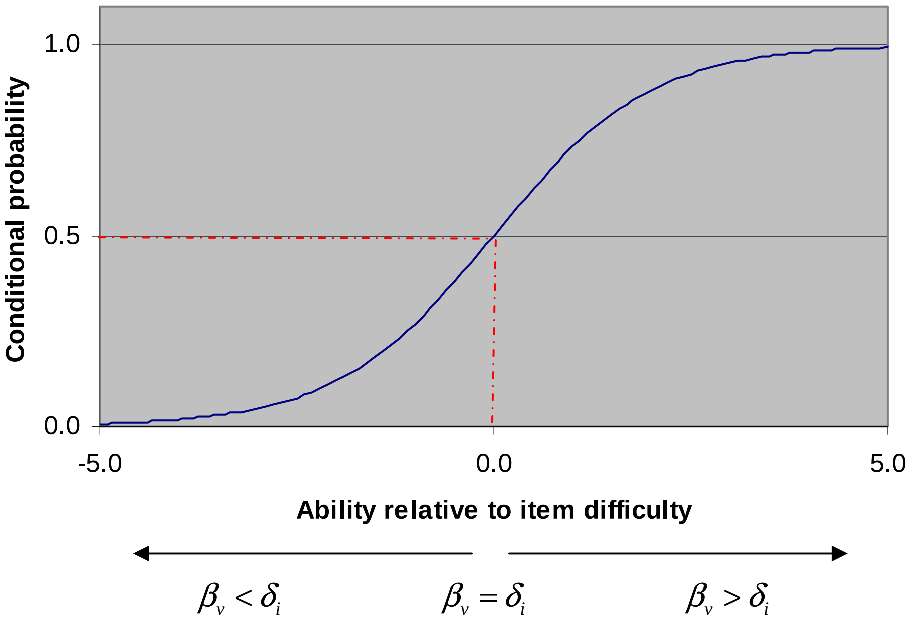
$\beta_v - \delta_i$	Probability
3	0.95
2	0.88
1	0.73
0	0.50
-1	0.27
-2	0.12
-3	0.05

The explanation of the dichotomous Rasch model is based on Andrich and Marais (2006).

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

Figure 3.4 displays the function of the dichotomous Rasch model graphically.

Figure 3.4: Item characteristic curve of the dichotomous Rasch model.



The item characteristic curve provides the opportunity to directly establish the probability of a person of ability β_v answering an item of difficulty δ_i correctly. For example, if in Figure 3.4 a person with ability $\beta_v = 0.0$ interacts with an item of difficulty $\delta_i = 0.0$ the probability is 50% that the answer will be correct (see dotted line on graph).

2. Polytomous Rasch models

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

Rasch-Andrich rating scale model

Andrich (as cited in Linacre, 2007, p7) in a conceptual breakthrough, comprehended that a rating scale, for example a Likert-type scale, could be considered as a series of Rasch dichotomies. Linacre (2007) makes the point that similar to the Rasch original dichotomous model, a person's ability or attitude is represented by β_v, whereas δ_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, p8), 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, F_x, is also located on the latent variable. This 'threshold' or 'step' is, according to Linacre (2005), 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, in other words, is the transition between two categories. Wright and Mok (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.

According to Linacre (2005), the Rasch-Andrich rating scale model specifies the probability, P_{vix}, that person v of ability β_v is observed in category x of a rating scale applied to item i with difficulty level δ_i as opposed to the probability $P_{vi(x-1)}$ 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:

$$\ln\left(\frac{P_{vix}}{P_{vi(x-1)}}\right)=\beta_v-\delta_i-F_x$$

In this research study, the categories for the Rasch-Andrich rating scale were:

1: Complete guess

2: Partial guess

3: Almost certain

4: Certain

A high raw score on an item would indicate a lot of confidence. When this figure is transformed to a log odds or logit, as it is done in the Rasch model, a low Rasch measure of endorsability is obtained. According to Planinic and Boone (2006), it is better to invert the scale for easier interpretation, since a high logit would then correspond to high confidence. This is the strategy adopted in this study.

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, F_x, in the partial credit model becomes F_{ix} and mathematically the Rasch-Andrich rating scale model changes to:

$$\ln\left(\frac{P_{vix}}{P_{vi(x-1)}}\right) = \beta_v - \delta_i - F_{ix}$$

These models will be re-visited in Chapter 6 in the data analysis methodology, to show how they were applied in this study.

3.4.1.4 Traditional test theory versus Rasch latent trait theory

In both traditional test theory and in the Rasch latent trait theory, total scores play a special role. In traditional test theory, test scores are test-bound and test scores do not mark locations on their variable in a linear way. In traditional test theory, the observed measure used for a person's performance would be the total score on the test. A higher total score on the test would be taken to reflect a higher level of understanding than would a lower total score on the test. The advice about item difficulties which develops from a traditional theory framework is that all items should be at a difficulty level of 0.5. Just how difficult an item needs to be for it to have a difficulty of 0.5 depends on how able the persons are who will take it. How able the persons are, is in turn judged from their performance on a set of items. There is no way within traditional test theory of breaking out of this reciprocal relationship other than through the performance of some carefully sampled normative reference group. The performance of individuals on subsequent uses of the test can be judged against the spread of performances in the normative group.

The Rasch model focuses on the interaction of a person with an item rather than upon the total test score. Total test scores are used, but the model commences with a modelling of a person's response to an item. The total score emerges as the key statistic with information about the ability β_v. A feature of traditional test theory is that its various properties depend on the distribution of the abilities of the persons. Many of the statistics depends on the assumption that the true scores of people are normally distributed (Andrich, 1988). An important advantage of the Rasch latent trait model is that no assumptions need to be made about this distribution, and indeed, the distribution of abilities may be studied empirically. It was for this reason that the Rasch model was chosen above other traditional statistical procedures for the quantitative research methodology of this study.

If we intend to use test results to study growth and to compare groups, then we must make use of the Rasch model for making measures from test scores that marks locations along the variable in an *equal interval* or *linear* way.

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 one has double the attribute that the other person has.

The question arises if raw scores per se can be realistically viewed as measures. Wright and Linacre (1989, p56) state ‘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.

3.4.1.5 Reliability and validity

Reliability and validity are approached differently in traditional test theory from the way they are approached in latent trait theory. The process of mapping the amount of a trait on a line necessarily involves numbers. The use of numbers in this way gives precision to certain kinds of work. However, there is always a trade-off in the use of such numbers – in particular, they can be readily over interpreted because they appear to be so precise, hence affecting the reliability of the data. In addition, the instrument may not measure what we really want to measure and this affects the validity of the research.

In the latent trait model, the use of a total score from a set of items implies an assumption of a single, unidimensional underlying trait which the items, and therefore the test, measure. Those reliability indices which reflect internal consistency provide a direct indication of whether a clear single dimension is present. If the reliability is low, there may be only a single dimension but one measured by items with considerable error. Alternatively, there may be other dimensions which the items tap to varying degrees.

The calculation of a reliability index is not very common in latent trait theory. However, it is possible to calculate such an index, and in a simple way, once the ability estimates and the standard error of the persons is known. Instead of using the raw scores for the reliability index formula, the ability estimates are used, where the ability estimate β_v for each person v can be expressed as the sum of the true latent ability and the error ε, i.e.

$$\beta_v = \beta_v + \Sigma\varepsilon\beta_v$$

The key feature of reliability in traditional test theory is that it indicates the degree to which there is systematic variance among the persons relative to the error variance i.e. it is the ratio of the estimated true variance relative to the true variance plus the error variance. In traditional test theory, the reliability index gives the impression that it is a property of the test, when it is actually a property of the persons as identified by the test. The same test administered to people of the same class or population but with a smaller true variance, would be shown to have a lower reliability.

Having the facility to capture the most well known and commonly used discrimination index of traditional test theory; to provide evidence of the degree of conformity of a set of responses to a Guttman or 'scalogram' scale in a probabilistic sense and to provide these from a latent trait formulation, indicates that Rasch's simple logistic model provides an extremely economical and reliable perspective from which to evaluate test data (Andrich, 1982).

3.4.2 Quantitative data collection

As discussed in Chapter 1, this study is set within the context of the Mathematics 1 Major Course at the University of the Witwatersrand. In Chapter 1, I indicated that the course has a mixed and heterogeneous student population; students coming from both the economically and culturally advanced sector of the population (for example, both parents may be university graduates) as well as from the economically and culturally disadvantaged sector (for example, one or more parents may be illiterate or innumerate).

In the years of this study, July 2004 to July 2006, student numbers registering for MATH109 were high with 483 in 2004, 414 in 2005 and 376 in 2006. The reduction in numbers in 2006 coincided with the increase in the entrance requirements to the Faculty of Science at the University of the Witwatersrand. In each of these years, the students were allocated, subject to timetable constraints, to one of two parallel courses presented by different lecturers. The lectures took place six times a week (45 minutes per lecture) in a large lecture theatre. MATH109 consists of a Calculus and an Algebra component. In Semester 1, Algebra constituted one-third and Calculus two-thirds of each assessment task, corresponding to the same ratio of lectures. In Semester 2, Algebra and Calculus were weighted equally with students receiving 3 lectures of Algebra and 3 lectures of Calculus per week. I lectured one set of Calculus and one set of Algebra classes while my colleagues lectured the other parallel courses. All the students from the MATH109 classes constituted the group from which data was collected for this study. As course co-ordinator for the duration of the study, I had more contact with these students than my colleagues. I was personally involved, either as examiner or as moderator, for all the tests and projects which contributed to the assessment programme. I was also directly responsible for the invigilation duties of this group and hence administered all the tests at which the data was collected.

The collection of data for this study was directly related to the Mathematics I Major assessment programme as illustrated in Figure 3.5.

Figure 3.5: Mathematics 1 Major (MATH109) assessment programme.

Diagnostic and Formative (Continuous)	**Summative**
● to get more information about the progress of learning and teaching.	● aimed at the results of the whole teaching process.
● from known to unknown	● from synthesis to consolidation.
● from corrective feedback to reinforcement	
Method of Assessment:	**Method of Assessment:**
Student's Portfolio	Final exam (3 hrs) November
● 2 MCQ tutorial tests ● Poster ● Groupwork tutorial tasks ● 2 Semester assignments: Calculus / Algebra ● Self-study tasks ● 3 class tests (1 hr) March/May/August ● 1 mid-year test (1.5 hrs) June	
50% - 60% of overall grade	40% - 50% of overall grade

Test instruments

Data was collected from the 2 MCQ Tutorial tests, the 3 class tests (CRQs and PRQs) (1 hour) in March/May/August, the mid-year test (CRQs and PRQs) (1.5 hrs) in June and the final examination (CRQs and PRQs)(3 hrs) in November, in each of the years 2004, 2005 and 2006 respectively.

Tutorial tests

Two tutorial MCQ tests were written during the course of the year in March and August respectively. Each test, of duration 20 minutes, consisted of 8 multiple-choice questions (total = 16 marks), 4 MCQs on Algebra content and 4 MCQs on Calculus content. Each of these MCQs was followed by a confidence of response question in which a student was asked to indicate their confidence about the correctness of their answer, where A implies no knowledge (complete guess), B a partial guess, C almost certain and D indicates complete confidence or certainty in the knowledge of the principles and laws required to arrive at the selected answer. Each of the MCQs had 3 distracters and 1 key, indicated by the letters A, B, C, or D.

Sample MCQ calculus question

If f is continuous and $\int_0^4 f(x)dx = 10$, find $\int_0^2 f(2x)dx$.

A. 5
B. 10
C. 15
D. 20

A COMPLETE GUESS	B PARTIAL GUESS	C ALMOST CERTAIN	D CERTAIN

(Adapted from MATH109 Tutorial Test, August 2005)

Tutorial tests were written during the last 20 minutes of one of the 45 minutes compulsory tutorial periods, in the first semester and the second semester. The tests were administered by the tutor who handed out the question papers together with a blank computer card. The instruction to each student was to shade the correct answers on the computer card to questions 1-8 in the first column. In these questions there was only one possible answer. There was no negative marking. In addition, the students had to shade their confidence of response answers on the computer card corresponding to Questions 1-8 in the second column, i.e. Questions [26] – [33]. Students were reminded that there is no correct answer in the confidence of responses. Students were also informed that marks were not awarded for the confidence of response answers, as these were purely for educational research purposes.

Once the tests had been written, the tutor collected both the question paper and the computer cards. The question papers were kept for reference only should any queries arise, and not returned to the students. The computer cards were marked by the Computer and Networking Services (CNS) division of the University of the Witwatersrand. On completion, CNS provided a print out of the quantitative statistical analysis of data, including the performance index, discrimination index and easiness factor per question. CNS also captured the students' confidence of responses.

Class tests and examinations

Three 1-hour class tests were written during the year in March, May and August. A 1.5 hour mid-year test was written in June and the final 3-hour examination took place in November. The final examination constituted 40% - 50% of the overall assessment grade. Each of these tests and exams followed the same format, with Section A following the PRQ format, in particular MCQs; Sections B and C followed the CRQ format with Section B testing the Algebra component of the course and Section C testing the Calculus component of the course.

In 2005, confidence of response questions were not included in Section B and Section C. This data was only collected for the MCQs in Section A. From 2006 onwards, the confidence of response questions were included in all 3 sections, for both the CRQ and PRQ formats. In the CRQ sections, a confidence of response question followed each subquestion of the main question.

Sample CRQ question:

Question 4.

a. Give the condition that is required to ensure continuity of a function $f(x)$ at the point $x = \alpha$.

A	B	C	D
COMPLETE GUESS	PARTIAL GUESS	ALMOST CERTAIN	CERTAIN

b. Let $\llbracket x \rrbracket$ be the greatest integer less than or equal to x.

(i) Show that $\lim\limits_{x \to 2} f(x)$ exists if $f(x) = \llbracket x \rrbracket + \llbracket -x \rrbracket$.

A	B	C	D
COMPLETE GUESS	PARTIAL GUESS	ALMOST CERTAIN	CERTAIN

(ii) Is $f(x) = \llbracket x \rrbracket + \llbracket -x \rrbracket$ continuous at $x = 2$? Give reasons.

A	B	C	D
COMPLETE GUESS	PARTIAL GUESS	ALMOST CERTAIN	CERTAIN

(Adapted from MATH109, Calculus, March 2006, Section C)

For Section A, students were provided with blank computer cards to indicate their choice of answers and the corresponding confidence of responses. As in the tutorial tests, students were informed that no marks were awarded for the confidence of responses. In Sections B and C, students were provided with space on the question papers to complete their solutions. The computer cards were used only to indicate the corresponding confidence of responses. On completion of the tests, all three sections, together with the filled in computer card, were collected. CNS provided a print out of all the results for Section A, together with confidence of responses for Sections A, B and C.

Expert opinions

In this study, the term *expert* refers to *content experts*. In this case the content experts were my colleagues who taught the MATH109 course, either Algebra or Calculus or both, as well as my supervisors from the University of Pretoria who were familiar with the content. In total, the opinions of eight experts on the level of difficulty of the questions were obtained, independent of each other. Five of the experts gave their opinions on Calculus, and six of the experts gave their opinions on Algebra. Each expert was given a full set of the following tests: MATH109 August Tutorial Test (2005); March Tutorial Test 1A (2006); March Tutorial Test 1B (2006); March Section A (2005); May Section A (2005); June Section A (2005); August Section A (2005); November Section A (2005); March Section A (2006); May Section A (2006); June Section A (2006); March Sections B & C (2005); May Sections B & C (2005); June Sections B & C (2005); August Sections B & C (2005); November Sections B & C (2005); March Sections B & C (2006); May Sections B & C (2006) and June Sections B & C (2006). The reader is to note that the August Tutorial Test was the same in both 2005 and 2006. Also the March Tutorial Test 1A which was written during a tutorial period on a Tuesday and March Tutorial Test 1B written during a tutorial period on the Wednesday of the same week, although testing the same content, were different. These tests were the same for 2005 and 2006. The experts chose to give their opinions on either the Calculus or Algebra questions, depending on which courses they taught. Hence for Calculus, Section C was appropriate and for Algebra, Section B was appropriate. In the MCQ Section A, there was a mixture of both Calculus and Algebra questions. Experts were asked for their opinions on the level of difficulty of both the PRQs and CRQs, and were asked to indicate their opinions as follows:

- Use a 1 if your opinion is that the students should find the question easy
- Use a 2 if your opinion is that the question is of average difficulty
- Use a 3 if your opinion is that the students would find the question difficult or challenging.

Experts were informed that their opinions were completely independent of how the students performed in the questions. Experts worked independently and did not collaborate with other experts. In the study, the students' performance is referred to as *novice performance*. Once all the expert opinions were collected, the data was captured separately for Calculus and Algebra on spreadsheets. An expert opinion on the level of difficulty of each question (PRQs and CRQs) was calculated as the average of the eight expert opinions per question.

3.5 RELIABILITY, VALIDITY, BIAS AND RESEARCH ETHICS

3.5.1 Reliability of the study

Reliability is the extent to which independent researchers could discover the same phenomena and to which there is agreement on the description of the phenomena between the researcher and participants (Schumacher & McMillan, 1993).

As this study consisted of both a qualitative and quantitative component, it is necessary to examine both the constraints on qualitative and quantitative reliability. According to Schumacher and McMillan (1993), reliability in quantitative research refers to the consistency of the test instrument and test administration in the study. Reliability in qualitative research refers to the consistency of the researcher's interactive style, data recording, data analysis and interpretation of participant meanings from the data.

Schumacher and McMillan (1993) have suggested the following reliability threats to research. These are:

- the researcher's role
- the informant selection of the sample
- the social context in which data is collected
- the data collection strategies
- the data analysis strategies
- the analytical premises i.e. the initial theoretical framework of the study.

In this study reliability was enhanced by means of the following:

- The importance of my social relationship with the students in my role as the co-ordinator and lecturer of the Mathematics 1 Major Course was carefully described.
- The selection of the population sample of this study and the decision process used in their selection was described in detail.
- The social context influencing the data collection was described physically, socially, interpersonally and functionally. Physical descriptions of the students, the time and the place of the assessment tasks, as well as of the interviews, assisted in data analysis.
- All data collection techniques were described. The interview method, how data was recorded and under what circumstances was noted.
- Data analysis strategies were identified.
- The theoretical framework which informs this study and from which findings from prior research could be integrated was made explicit.
- Stability was achieved by administering the same tutorial tests in March and August over the period 2004-2006.
- Equivalence was achieved over the period of study, by administering different tests to the same group of students.
- Internal consistency was achieved by correlating the items in each test to each other.
- A large number of data items were collected over the period of 2 years, and were all used in the data analysis.

3.5.2 Validity of the study

In the context of research design, the term *validity* means the degree to which scientific explanations of phenomena match the realities of the world (Schumacher & McMillan, 1993). *Test validity* is the extent to which inferences made on the basis of numerical scores are appropriate, meaningful and useful. Validity, in other words, is a situation-specific concept. Validity is assessed depending on the purpose, population and environmental characteristics in which measurement take place.

In quantitative research there are two type of design validity. *Internal validity* expresses the extent to which extraneous variables have been controlled or accounted for. *External validity* refers to the generalisability of the results i.e. the extent to which the results and conclusion can be generalised to other people and settings. In this study, internal validity was addressed as the population sample of first year mainstream mathematics students were always fully informed and aware that their confidence of responses, in both the CRQs and PRQs, were not for assessment purposes, but used purely for this research study. All students wrote the same test on the same day in a single venue. All the data collected was used, irrespective of whether the students completed all of the confidence of responses, or not.

According to Messick (1989), validity is articulated in terms of the following four ideas: *content validity*, *concurrent validity*, *predictive validity* and *construct validity*.

- Content validity would be established by experts judging whether the content was relevant
- Concurrent validity would be established by showing that the results on a particular test were related in the expected way with results on other relevant tests
- Predictive validity would be established by relating the results of a test with performance in the future on the same trait
- Construct validity would be established by demonstrating that the test was related to performances on other tests that were theoretically related.

Andrich and Marais (2006) point out that it is now considered standard that construct validity is the overarching concept, and that the other three so called forms of validity are pieces of evidence for construct validity. Construct validation is addressed to the identification of the dimension in a substantive sense. The test developer must have a clear idea of what the dimension is when the items are written.

In order to enhance the validity of this study, the following steps were taken:

- The literature was examined in order to identify and develop the seven mathematical assessment components.
- The test instrument was validated after implementation by a panel consisting of my 2 supervisors at the University of Pretoria and 6 mathematics lecturers from the University of the Witwatersrand.
- The questions used for data collection were all moderated by colleagues and were in line with the theoretical framework. Minor adjustments were made to a number of test items to avoid ambiguity and to strengthen weak distracters.
- Expert opinions obtained from colleagues were completely independent of student performance (novice performance).
- Three measuring criteria were identified in order to develop a model for addressing the research questions. These criteria were modified and adapted in collaboration with my supervisors to address the issue of what constitutes a good mathematical question and how to measure how good a mathematics question is.
- All marking of PRQs was done by computers using the Augmented marking scheme. This programme accommodates the fact that not all questions are equally weighted. There was no negative marking.
- Marking of CRQs was done by the MATH109 team of lecturers, using a detailed marking memorandum which had been discussed prior to each marking session. In addition, all marking was moderated by the researcher, except for the examinations which were moderated by an external examiner.

3.5.3 Bias of the study

Bias is defined by Gall, Gall and Borg (2003) as a set to perceive events in such a way that certain types of facts are habitually overlooked, distorted or falsified.

In this study, an attempt was made to decrease bias by the following:

- A representative sample of undergraduate students studying tertiary mathematics
- A comprehensive literature review
- Verified statistical methods and findings.

3.5.4 Ethics

Ethics generally are considered to deal with beliefs about what is right or wrong, proper or improper, good or bad (Schumacher & McMillan, 1993). Most relevant for educational research is the set of ethical principles published by the American Psychological Association in 1963.

The principles of most concern to educators are as follows:

- The primary investigator of a study is responsible for the ethical standards to which the study adheres.
- The investigator should inform the subjects of all aspects of the research that might influence willingness to participate.
- The investigator should be as open and honest with the subjects as possible.
- Subjects must be protected from physical and mental discomfort, harm and danger.
- The investigator should secure informed consent from the subjects before they participate in the research.

In view of these principles, I took the following steps:

- Permission to conduct research in the first year Mathematics I Major course was sought and granted by the Registrar of the University of the Witwatersrand. Permission was granted on the understanding that information furnished to me by the University of the Witwatersrand may not be used in a manner that would bring the University in disrepute. I further agreed that my research may be used by the University if it is so desired (Declaration letter can be found in the Appendix A1, p265).

- In the interview, all respondents were assured of confidentiality. Respondents were informed that they had been randomly selected, based on their June class record marks. Permission was obtained from each candidate to tape-record the interviews. Candidates were informed that they were free to withdraw from the interview or not to answer any question, if they wished. Candidates were assured of the confidentiality and anonymity of their responses and, in particular, that the information they provided for the research would not be divulged to the University or their lecturers at any time.
- The researcher assured all participants that all data collected from the confidence of responses would not affect their overall marks. No person, except the researcher, supervisors and the data analyst, would be able to access the raw data. All raw data was used, irrespective of whether the student indicated a confidence of response or not.
- The research report will be made available to the University of the Witwatersrand and to the University of Pretoria, should they so desire it.
- Informed consent was achieved by providing the subjects with an explanation of the research and an opportunity to terminate their participation at any time with no penalty. Since test data was collected over the research period to chart performance trends, the research was quite unobtrusive and had no risks to the subjects. The students were at no times inconvenienced in the data collection process, as all data was collected during the test times as set out in the assessment schedule for MATH109.
- In the data analysis, student names and student numbers were not used. Thus, confidentiality was ensured by making certain that the data cannot be linked to individual subjects by name. This was achieved by using the Rasch model.
- In my role as researcher, I will make every effort to communicate the results of my study so that misunderstanding and misuses of the research is minimised.
- To maximise both internal and external validity, research has shown it seems best if the subjects are unaware that they are being studied

(Schumacher & McMillan, 1993). In this regard, the research methodology was designed in order to collect data from the students during their normal tutorial times or formal test times. As a result, students did not feel threatened in any way and the resulting data was sufficiently objective.

- The methodology section of my study shows how the data was collected in sufficient detail to allow other researchers to extend the study.
- In my roles as co-ordinator, lecturer and researcher, I was very aware of ethical responsibilities that accompanied the gathering and reporting of data. The aims, objectives and methods of my research were described to all participants in this research study.