Comparing different assessment formats
in undergraduate mathematics

by

Belinda Huntley

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor

in the Department of Mathematics and Applied Mathematics
in the Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

April 2008

DECLARATION

I, the undersigned, hereby declare that the thesis submitted herewith for the degree Philosophiae Doctor to the University of Pretoria contains my own, independent work and has not been submitted for any degree at any other university.

Name:…………………………………
      Belinda Huntley

Date:………………………………..

ABSTRACT

In this study, I investigate how successful provided response questions, such as multiple choice questions, are as an assessment format compared to the conventional constructed response questions. Based on the literature on mathematics assessment, I firstly identify an assessment taxonomy, consisting of seven mathematics assessment components, ordered by cognitive levels of difficulty and cognitive skills. I then develop a theoretical framework, for determining the quality of a question, with respect to three measuring criteria: discrimination index, confidence index and expert opinion. The theoretical framework forms the foundation against which I construct the Quality Index (QI) model for measuring how good a mathematics question is. The QI model gives a quantitative value to the quality of a question. I also give a visual representation of the quality of a question in terms of a radar plot. I illustrate the use of the QI model for quantifying the quality of mathematics questions in a particular undergraduate mathematics course, in both of the two assessment formats – provided response questions (PRQs) and constructed response questions (CRQs). I then determine which of the seven assessment components can best be assessed in the PRQ format and which can best be assessed in the CRQ format. In addition I also investigate student preferences between the two assessment formats.

Keywords: Mathematics assessment, Quality Index, good mathematics questions, assessment components, assessment taxonomies, provided response questions, constructed response questions, multiple choice questions.

DEDICATION

"Yea, if thou criest after knowledge, and liftest up thy
voice for understanding; if thou seekest her as silver,
and searchest for her as for hidden treasures; then shalt
thou understand the fear of the Lord, and find the
knowledge of God. For the Lord giveth wisdom; out of
His mouth cometh knowledge and understanding".

PROVERBS 2: 3 - 6

ACKNOWLEDGEMENTS

The author would hereby like to thank all people and organisations whose assistance and co-operation contributed to the completion of this thesis, and in particular:

My supervisor, Professor Johann Engelbrecht, for setting high professional standards which provided the much-needed challenge and motivation, and for his interest and moral support.

My co-supervisor, Professor Ansie Harding, for her invaluable guidance and expert assistance throughout the period of this research.

Elsie Venter, a senior lecturer from the Centre for Evaluation and Assessment, School of Education, University of Pretoria, for introducing me to the Rasch method of data analysis and for her assistance in analysing my research data.

Marie Oberholzer, for editing and type-setting the final draft of my thesis with great care and diligence.

My parents, Roland and Daisy Hill, for their prayers of upliftment and loving support.

My husband, Brian and children, Byron, Christopher and Cayla, for their total devotion and patience and on-going faith in my abilities.

INDEX OF TABLES

INDEX OF FIGURES

TABLE OF CONTENTS

CHAPTER 7:   DISCUSSION AND CONCLUSIONS

REFERENCES

APPENDIX

CHAPTER 1:   INTRODUCTION

1.1   PURPOSE OF STUDY

> The quickest way to change student learning is to change the assessment system (Biggs, 1994, p5).

The purpose of this research study is to investigate to what extent alternative assessment formats, such as *provided response questions* (PRQs) format, in particular multiple choice questions (MCQs), can successfully be used to assess undergraduate mathematics.   For this purpose I firstly develop a model to measure how good a mathematics question is.   To my knowledge, no such model currently exists and such a measure of the quality of a question is original. The objective is then to use the proposed model to determine whether all undergraduate mathematics can be successfully assessed.   For this purpose a taxonomy of *assessment components* of mathematics is developed to enable us to identify those components of mathematics that can be successfully assessed using alternative assessment formats.   Where this is not the case, the proposed model is used to determine whether the conventional *constructed response questions* (CRQs) format is more suitable for assessment purposes. By using the proposed model to compare the PRQ assessment format with the more conventional, open-ended CRQ assessment format applied in tertiary first year level mathematics courses, I attempt to address the research question of whether we can successfully use PRQs as an assessment format in undergraduate mathematics.

One of the aims of tertiary education in mathematics should be to develop proficiency within all components of mathematics.   A greater knowledge of the suitability of question formats within different components can assist educators and assessors to improve their assessment programmes, enhancing problem-solving abilities, reducing misconceptions, restricting surface learning and simultaneously improving the efficacy of marking and maintaining standards in a

first year tertiary mathematics course with large student numbers, as described in this study. This research study aims to assist mathematics educators and assessors in reducing their large marking loads associated with continuous assessment practices in first year undergraduate mathematics courses, by determining in which of the assessment components the PRQ assessment format can be used successfully, without undermining the value of assessment of undergraduate mathematics courses.

1.2 STATEMENT OF PROBLEM

In South Africa, as in the rest of the world, higher education has been forced to respond to the demands placed on the sector by two late modern imperatives, globalisation and massification of education (Luckett & Sutherland, 2000). In Southern Africa, and in particular South Africa, the accessibility of higher education to the masses has a particularly moral dimension, as it implies the need to respond to the historical inequalities of the past apartheid era, by making the higher education sector accessible to previously disadvantaged black and working class communities. The apartheid government in South Africa attempted to limit access by black students by excluding them from most higher education institutions, imposing a quota system and by establishing institutions that are now regarded to be 'historically disadvantaged' universities (Makoni, 2000). With the consolidation of democracy, economic and political changes are taking place at the same time as the radical rethinking of the educational philosophies underlying higher education. Higher education needs to be more open, flexible, transparent and responsive to the needs of underprepared, lifelong and part-time learners (Luckett & Sutherland, 2000). This statement has implications for appropriate assessment practices in higher education.

My interest in different forms of assessment at the first year level in undergraduate mathematics grew out of my role as a lecturer and coordinator of the Mathematics I Major course at the University of the Witwatersrand. In South Africa, the socio-economic and policy contexts emerging from the post-colonial

and post-apartheid reconstruction, pose enormous challenges for assessment practices in higher education. With more and more students being drawn to higher education, the numbers of first year undergraduate students studying tertiary mathematics are increasing rapidly. The growth in numbers of students enrolling for first year mathematics courses is not unique to the School of Mathematics at Wits University, in which the study was based. In a study conducted by Engelbrecht and Harding (2002), it was observed that this increase in first year enrolment numbers in mathematics is a national trend over the past decade in South African universities. At first year level Mathematics is regarded as a pre-requisite for many courses and is considered essential for students who venture into engineering and many other fields of technology.

With this increase in student numbers, one of the challenges facing academics is that the more conventional open-ended constructed response questions (CRQ) assessment format is placing increased pressure on academic staff time. The assessment load created by increasing numbers of students and the shift in thinking towards competency frameworks are among the most prominent of many pressures. Improving student learning, encouraging deep rather than surface learning and nurturing critical abilities and skills all require time. However, in an expanding higher education system with increased student numbers and large classes, the conscientious educator is faced with a problem. Larger classes lead to more marking and, if properly done, takes more time. While lecturers can usually handle many more students in a lecture, the corresponding increase in their marking loads is another matter entirely. Continuous assessment of large undergraduate mathematics classes, which is generally considered as essential, can no longer be afforded because of the corresponding huge marking load. Alternatives have to be found.

As the sizes of first year mathematics classes increase, so does the teaching load and especially the marking load. Decreasing the amount of feedback to each student in order to complete the task in the limited time available is clearly undesirable, given the great potential of feedback in assessment (Boud, 1995). The notion of 'working smarter, not harder' (Brown & Knight, 1994) should be

pursued. If assessment is to be a useful part of the learning experience of students, it is beneficial to employ a fairly diverse variety of assessment types and formats. The implementation of alternative assessment formats such as provided response questions (PRQ), including multiple choice items, matching and the single-response item assessment format, amongst others, is gathering support. Firstly, their simplicity is such that implementation for marking by computer, either through optically marked response sheets, or directly online is straightforward. Processing through optically marked recorders is fast, easy and is amenable to a variety of analysis. Secondly, scoring is immediate and efficient. PRQs can be very useful for diagnostic purposes for helping students to see their strengths and weaknesses. Thirdly, as this study aims to show, PRQs can be constructed to evaluate higher order levels of thinking and learning, such as integrating material from several sources, critically evaluating data and contrasting and comparing information.

1.3   SIGNIFICANCE OF THE STUDY

In South Africa, as in the rest of the world, the changes in society and technology have imposed pressures on academics to review current assessment approaches. In these years of post-colonial and post-apartheid reconstruction in South Africa, academics are tasked with ensuring that graduates are able to apply their knowledge outside of the tertiary environment and to communicate and apply that expertise in a wide range of contexts (Makoni, 2000).

Changes in educational assessment are currently being called for, both within the fields of measurement and evaluation as well as in specific academic disciplines such as mathematics. Geyser (2004, p90) summarises the paradigm shift that is currently under way in tertiary education as follows:

> The main shift in focus can be summarized as a shift away from assessment as an add-on experience at the end of learning, to assessment that encourages and supports deep learning. It is now important to distinguish between learning

for assessment and learning from assessment as two complementary purposes of assessment….

Assessment should be seen as an integral and vital part of teaching and learning. An emerging vision of assessment is that of a dynamic process that continuously yields information about student progress toward the achievement of learning goals (NCTM, 1995). This vision of assessment acknowledges that when the information gathered is consistent with learning goals and is used appropriately to inform instruction, it can enhance student learning as well as document it (NCTM, 2000). Rather than being an activity separate from instruction, assessment is now being viewed as an integral part of teaching and learning, and not just the culmination of instruction (MSEB, 1993). Assessment drives what students learn (Hubbard, 1997). Every act of assessment gives a message to students about what they should be learning and how they should go about it. It controls their approach to learning by directing them to take either a surface approach or a deep approach to learning (Smith & Wood, 2000). Students gear their learning processes to be effective for the type of assessment they will undergo. They will seek and request teaching methods that will best fulfil their ability to respond to the assessment.

Because assessment is often viewed as driving the curriculum and students learn to value what they know they will be tested on, we should assess what we value. The type of questions we set show students what we value and how we expect them to direct their time (Hubbard, 1995).

This study attempts to define the concept of a 'good' or successful question which can be used to successfully assess mathematics in both the PRQ and CRQ formats. Assessment must be linked to and be evidence of the levels of learning and in particular the learning outcomes and competencies required.

> Assessment defines for students what is important, what counts, how they will spend their time and how they will see themselves as learners. If you want to change student learning, then change the methods of assessment (Brown, Bull & Pendlebury, 1997, p6).

5

The more data one has about learning, the more accurate the assessment of a student's learning. Assessment forms a critical part of a student's learning.

> Student assessment is at the heart of an integrated approach to student learning (Harvey, 1992, p139).

Mathematics at tertiary level remains conservative in its use of alternative formats of assessment. As goals for mathematics education change to broader and more ambitious objectives (NCTM, 1989), such as developing mathematical thinkers who can apply their knowledge to solving real problems, a mismatch is revealed between traditional assessment and the desired student outcomes. It is no longer appropriate to assess student knowledge by having students compute answers and apply formulas, because these methods do not reveal the current goals of solving real problems and using mathematical reasoning.

During the period of this study (2004-2006) enrolment numbers for the first year mainstream mathematics course were large, with numbers between 400 to 500 students in each year. These large numbers placed increased pressures on academic staff time. In particular, the more conventional open-ended CRQ assessment format, which was the predominant method of assessment, resulted in very large marking loads. Recent expansions in student numbers have tended to result in an increase in teaching class sizes accompanied by a reduction in small group tutorial provisions. The wider access to higher education together with increased recruitment of tertiary students, have added to the burden of making provision both for larger groups and for individuals. This challenge led me to re-evaluate current assessment practices and to explore alternative assessment approaches.

I hope that, based on the research findings, more support will be gained for assessment using the provided response (PRQ) format in undergraduate mathematics. Perhaps it is time for those involved in course co-ordination and curriculum design of large undergraduate mathematics courses to examine the learning benefits and experiment with changes in assessment. Computer

assisted multiple choice testing can provide a means of preserving formative assessment within the curriculum at a fraction of the time-cost involved with written work. Furthermore, developing a model by which to measure the quality of a question (PRQ or CRQ) is of great benefit to the successful assessment of such large undergraduate mathematics courses, improving the efficacy of the marking with respect to both time and quality. No such measure currently exists and such a model can be used to measure the quality of questions, either in PRQ or CRQ format. A greater knowledge of the quality of questions within the assessment components can assist mathematics educators and assessors to improve their assessment programmes and enhance student learning in mathematics.

1.4   CONTEXT OF THIS STUDY

In this study, I firstly investigate how we can measure whether a mathematics question is of a good quality or not. Three measuring criteria are used to develop a model for determining the quality of a question. Secondly, using this model, the quality of all PRQs and CRQs are determined. Thirdly, a comparison is made within each mathematics assessment component, between the PRQ assessment format and the CRQ assessment format. Furthermore, I investigate student preferences regarding the different assessment formats, both PRQ and CRQ, in a first year mainstream mathematics course at the University of the Witwatersrand in Johannesburg, South Africa.

University of the Witwatersrand

The study is set within the milieu of a first year mathematics course (Mathematics I Major) at the University of the Witwatersrand over the period July 2004 to July 2006. The University of the Witwatersrand is a major research-orientated South African institution that draws its students from diverse socio-economic backgrounds and a wide range of high schools (Adler, 2001). For example, some students come from schools which for the last several years have had close to 100% matriculation (Grade 12) pass rate; others come from

schools where the overall pass rate at the matriculation level over the last few years is less than 60%.

School of Mathematics

The School of Mathematics at the University of the Witwatersrand offered a three-year mathematics major course in the BSc, BA and BCom degrees between 2000 and 2004. From 2005 onwards, two majors were offered, Mathematics and Mathematics Techniques, a minor academic development that recognises the de facto distinction between the two essentially distinct suites of topics and their outcomes, aimed at students wishing to pursue careers in mathematics teaching. Student registrations in the School of Mathematics have increased by 73% since 2000, in line with an increase in registrations at the University of the Witwatersrand. In 2004, over 3400 students registered in the School of Mathematics and mathematics student numbers accounted for about 18.5% of the Faculty of Science. The average pass rate in the School of Mathematics was at the 70% level over the period of this study. A summary of course registration figures is given in Table 1.1.

Table 1.1: Student numbers and pass rates for undergraduate mathematics courses, 2000-2004.

Year	2000	2001	2002	2003	2004
Actual student course numbers	1998	2666	3203	3383	3447
Course Pass	1439	2053	2338	2402	2413
Course Fail	550	594	832	948	1017
Course Pass Rate	72	77	73	71	70
Course Cancelled	236	382	241	272	263

(Source: Executive Information System, School of Mathematics, Academic Review, University of the Witwatersrand)

First year Mathematics Major (MATH109)

The first year Mathematics Major course (MATH109) has a minimum entry level of a Higher Grade C Symbol in Grade 12 mathematics. MATH109 has two compulsory components, Calculus and Algebra, both taught and tested throughout the year with a final examination in November.

The Mathematics I Major course, MATH109, is intended both for students who wish to become professional mathematicians or high school mathematics teachers and for students who need to complete the course as a co-requisite to other courses in the Science Faculty such as Physics or Computer Science. Students who are studying the Biological Sciences do not generally take the Mathematics I Major course. They do a less theoretical, more skill-oriented first year Ancillary Mathematics course and they cannot proceed to a second year of mathematics.

The MATH109 course is compulsory for students entering degree courses in mathematics, computing, actuarial science, economics, statistics, but also attracts students from the biological sciences, humanities, education and business. This course thus attracts the kind of diversity now commonly found in undergraduate tertiary mathematics. Students' interests, levels of motivation and mathematical needs are very varied in the group. Although all students in the course have studied Grade 12 Higher Grade mathematics, the students emanate from a range of schools and thus have a range of mathematical backgrounds. For example, many students have taken Additional Mathematics as an extra subject at school and hence have covered most of the Calculus and Algebra material taught in the first semester. At the other end of the spectrum, students have achieved the minimum entrance requirements, and due to disadvantaged educational backgrounds, demonstrate weaknesses in some areas of school mathematics such as fundamental algebra, trigonometry, functions and graphing.

With the large number of students involved, the teaching in the first year is predominantly in large groups (up to 150 students per class) and each group comprises students from more than one faculty. It is also inevitable that an initial level of attainment and competence in a range of mathematical skills and knowledge is assumed of the class. Teaching in large classes is staff-efficient, but little direct provision can be made in lectures or classes to accommodate possible initial deficiencies of individual students where precise and detailed

feedback would be valuable. Supplementary assistance through tutorials are used to help students on a more individual basis. The tutorial classes are weekly 45-minute periods during which about 25 students come together in a class with a lecturer or student assistant. The tutorial classes are primarily periods in which the student can consult the lecturer or student assistant on particular tutorial problems or mathematical concepts. The tutorial problems are mathematical exercises which have been set, prior to the tutorial period, by the course co-ordinator (myself, in this instance), and are usually from the prescribed textbook.

An important aspect of the MATH109 course is the prescribed Calculus textbook (Stewart, 2000). The textbook has many features advocated by the Calculus Reform Movement: for example, multiple representations of mathematical objects are presented in the textbook as are real-life applications of many mathematical concepts. Unfortunately, the textbook is still used in a traditional and conservative way: inter alia, students are not allowed to use technology such as graphics calculators or computers in problem-solving or in examinations, and group projects are not considered acceptable components of the assessment programme. However, in 2004, a technology component in MATH109 was introduced in which students learned the rudiments of 'Mathematica'. This teaching innovation, using technology as a tool, had an impact on the assessment programme of MATH109. During the period of my study, the MATH109 assessment programme consisted of 4 class tests, a mid-year exam and a final examination. The October class record is the cumulative of all tests and assignments written before the final exam (continuous assessment). In order to pass MATH109, the students' final year mark must be ≥50%. Prior to the period of my study, assessment of the course had been very traditional with the CRQ assessment format being the predominant method of assessment. The implementation of alternative assessment formats such as PRQs, including MCQs, matching and single item-response questions for mathematics assessment was initially met with some resistance by the academic staff of the School of Mathematics at the University of the Witwatersrand. However, with the numbers of first year undergraduate students

studying tertiary mathematics increasing and the problems surrounding large-scale traditional CRQ format examinations, such as quick and efficient marking of these, becoming more and more acute, the use of alternative PRQ assessment format gathered support.

Conformity with qualification specifications

The interim registration of the BSc degree under the South African National Qualifications Framework (NQF) requires that graduates have certain skills and abilities. The NQF may briefly be described as a flexible structure for articulating the various levels of the educational enterprise, at a national level. Its main purpose is to provide a degree of standardisation and interchangeability of educational qualifications across the country (Dison & Pinto, 2000). The MATH109 course confirms to the NQF requirements. Graduates' skills and abilities are specified in Exit Level Outcomes (ELOs) in Table 1.2, found in Appendix A2. How these ELOs are assessed constitutes a series of Associated Assessment Criteria (AAC) in Table 1.3, found in Appendix A3. The ELOs and the AAC incorporate the Critical Cross-Field Outcomes (CCFOs) listed in Table 1.4, found in Appendix A4.

1.5 OUTLINE OF STUDY

In the purpose of this study outlined in Chapter 1, I indicated that my primary research focus is to develop a model to measure how good a mathematics question is and to use this model to determine to what extent provided response questions (PRQs) and constructed response questions (CRQs) can be used to successfully assess mathematics at undergraduate level.

In order to develop this research focus, I discuss and compare different purposes of assessment such as diagnostic, formative and summative. These will be reviewed in the literature review in Chapter 2. Terminology relevant to this study, as well as mathematics assessment components (Niss, 1993) will also be reviewed. Important issues in assessment practices for university

undergraduates will be identified (Biggs, 2000). Certain interesting alternative methods of assessment and question types in undergraduate mathematics will be explored (Cretchley, 1999; Anguelov, Engelbrecht, & Harding, 2001; Hubbard, 2001; Wood & Smith, 1999, 2001). In addition, various assessment taxonomies will also be discussed (Biggs & Collis, 1982; Bloom, 1956; Crooks, 1988; De Lange, 1994; Freeman & Lewis, 1998; Hubbard, 1995; Smith, Wood, Crawford, Coupland, Ball & Stephenson, 1996). What the literature on assessment reveals about good assessment practices and the qualities of a "good" question will be presented (Fuhrman, 1996; Haladyna, 1999; Webb & Romberg, 1992). This will become relevant when considering when a question in the assessment of mathematics is considered to be successful. Literature on the issue of confidence will also be presented. Other non-mathematical studies (Hasan, Bagayoko & Kelley, 1999; Potgieter, Rogan & Howie, 2005), where a respondent is requested to provide the degree of confidence he has in his own ability to select and utilise well-established knowledge, concepts or laws to arrive at an answer, will be elaborated upon in the literature review.

Having defined the necessary theoretical background in Chapter 2, I introduce new concepts pertinent to my research study in Chapter 3. In this chapter on research design and methodology, I state my research question and subquestions in a more focused way. I describe how I went about investigating my research question and subquestions. The population sample and sampling procedures are described. The organisation of the study discusses both the qualitative and quantitative research methodologies. In particular, an in-depth discussion of the Rasch model (Rasch, 1960) is presented as this is the method of quantitative data analysis used in this research study. Issues of reliability validity, bias and ethics are also discussed.

Chapter 4 presents the qualitative investigation which forms part of the qualitative research methodology. The qualitative investigation is in the form of interviews conducted with a representative sample of the target population of the study. These interviews were conducted to establish student preferences regarding different assessment formats that they had been exposed to in their

undergraduate mathematics course. Qualitative data in the form of student opinions will be summarised.

In Chapter 5, a set of seven mathematics assessment components, based on Niss's (Niss, 1993) mathematics assessment components discussed in Chapter 2, will be proposed. Further background will be given on the confidence index, together with a description of other statistical parameters pertinent to this study. In this chapter, I attempt to develop a theoretical framework to form a way of measuring the qualities of a *good mathematics question*. In particular, three measuring criteria: *discrimination index*, *confidence index* and *expert opinion*, will be described. These three parameters are used for measuring the quality of a test item. A Quality Index (QI) model, based on the measuring criteria, is developed to measure the quality of a good mathematics question. The QI model will be used both to quantify and visualise the quality of a mathematics question. The theoretical framework forms the foundation against which we address the research question and subquestions of how we can measure how good a mathematics question is and which of the mathematics assessment components can be successfully assessed in the PRQ format, and which can be better assessed in the CRQ assessment format.

Chapter 6 presents the quantitative research findings and results. In the quantitative data analysis methodology, an overview of the statistical procedures followed will be given. Both the traditional statistical analysis of the quantitative data and the Rasch (Rasch, 1960) method of data analysis is discussed under the methodology section. A description of the data follows in which details of the tests written, the number of PRQs per test, the number of CRQs per test and the number of students per test are summarised. A component analysis is presented within the different assessment components. In this analysis, examples of items, both PRQs and CRQs, together with a radar plot and a table summarising the quality parameters of each item, is presented. Finally an analysis of good quality items and poor quality items in each of the PRQ and CRQ assessment formats, in terms of the quality index developed in section 5.3.2, within each of the seven assessment components will be presented.

In Chapter 7, I set about discussing my research results. The discussion in this chapter will include the interpretation of the results and the implications for future research. I also discuss how the research results could have implications for assessment practices in undergraduate mathematics. Furthermore, I draw conclusions from my research about which of the mathematics assessment components, as defined in section 5.1, can be successfully assessed with respect to each of the two assessment formats, PRQ and CRQ. The Quality Index model will be used both to quantify and visualise the quality of a mathematics question. In this way, I endeavour to probe and clarify my research question and subquestions as stated in section 3.2. I will signal some limitations of my research study, as well as some pedagogical implications for further research.

CHAPTER 2:   LITERATURE REVIEW

In order to set the background for furthering research knowledge in the area of assessment in tertiary undergraduate mathematics, various documents on what other researchers have produced are reviewed. These will include preliminary sources i.e. hard-copy or electronic indices to the literature; primary sources i.e. reports of research studies written by those who conducted them; and secondary sources i.e. published reviews of particular bodies of literature.

2.1   TERMINOLOGY

Some technical clarification is necessary, as in this study the terms *assessment, evaluation, tests* and *examinations* shall be used frequently. According to Niss (1993) 'assessment in mathematics education is taken to concern the judging of the mathematical capability, performance and achievement of students whether as individuals or in groups' (p3). Assessment has been described as the heart of the student experience, the barometer of an educational system and the quality of teaching it provides (Luckett & Sutherland, 2000). Rowntree (1987) offers another definition, which emphasises the intimacy, subjectivity and professional judgement involved:

> Assessment in education can be thought of as occurring whenever one person, in some kind of interaction, direct or indirect, with another, is conscious of obtaining and interpreting information about the other person. To some extent or other it is an attempt to know that person. In this light, assessment can be seen as human encounter (p4).

The following two definitions by the South African Qualifications Authority (SAQA) for the registration of South African qualifications reflect only one aspect of assessment, namely the process:

Assessment is about collecting evidence of learners' work so that judgements about learners' achievements, or non-achievements, can be made and decisions arrived at.

Assessment is a structured process for gathering evidence and making judgements about an individual's performance in relation to registered national standards and qualifications (SAQA, 2001, pp15, 16).

Brown, Bull and Pendlebury (1997) provide a useful, working definition of assessment: 'Assessment consists, essentially, of taking a sample of what students do, making inferences and estimating the worth of their actions' (p8). *Assessment* is thus concerned with the outcomes of mathematics teaching at the student level. In its narrowest form, assessment seeks to measure the degree to which learning objectives have been met. In a broader context, it seeks to measure the achievement of graduate attributes (Groen, 2006).

Evaluation in mathematics education on the other hand, is taken to be the judging of educational systems or instructional systems as far as mathematics teaching is concerned. These systems include curricula, programmes, teachers, teacher training, schools or school districts. Thus, evaluation addresses mathematics education at the systems level. According to Scriven (1991), evaluation refers to both the methods of gathering information from students and the use of that information to make a variety of judgements (p139). Romberg (1992, p10) describes evaluation as 'a coat of many colours'. He emphasises that to assess student performance in mathematics, one should consider the kinds of judgements or evaluations that need to be made and consequently develop assessment procedures to address those judgements.

We need to view tests as 'assessments of enablement' (Glaser, 1988, p40). In other words, rather than merely judging whether students have learned what was taught, we should 'assess knowledge in terms of its constructive use for further learning' (Wiggins, 1989, p706).

The word *test* originated from a *testum*, which was a porous cup determining the purity of metal. Later it came to stand for any procedures for determining the worth of a person's effort. The root of the word *assessment* reminds us that an assessor (from *ad* + *sedere*) should *sit with* a learner in some sense to be sure that the student's answer really means what it seems to mean. The implication of this is that assessment is primarily concerned with providing guidance and feedback to the learner. This is ultimately still the most important function of assessment. Tests and exams should be central experiences in learning, not just something to be done as quickly as possible after teaching has ended in order to produce a final grade (Steen, 1999). To let students show what they know and are able to do is a very different business from the all too conventional practice of counting students' errors on questions. Such assessment practices do not welcome student input and feedback. Wiggins (1989) suggests that we think of students as apprentices who are required to produce quality work and are therefore assessed on their real performance and use of knowledge.

For the purpose of this study, the term *assessment* will be used to refer to any procedure used to measure student learning. When tests and examinations are considered to be ways of judging student performance, they are forms of assessment. On the other hand, when the outcomes of tests and examinations are used as indicators of the quality of an educational system, then examinations and tests belong to the realm of evaluation.

2.2 THE CHANGING NATURE OF UNIVERSITY ASSESSMENT IN THE SOUTH AFRICAN CONTEXT

In recent years, assessment has attracted increased attention from the international mathematics education community (MSEB, 1993; CMC and EQUALS, 1989). There are numerous reasons for this increase in attention, of which one seems to predominate. During the last couple of decades, the field of mathematics education has developed considerably in the area of outcomes and objectives, theory and practice (Hiebert & Carpenter, 1992; Niss, 1993;

Romberg, 1992; Schoenfeld, 2002; Stenmark, 1991). These developments have not, however, been matched by parallel developments in assessment. Consequently, an increasing mismatch and tension between the state of mathematics education and current assessment practices are materialising. Changing teaching without due attention to assessment is not sufficient (Brown, Bull & Pendlebury, 1997).

Changes in educational assessment in universities are currently being called for - in its intent and in its methods. While much assessment still focuses on ranking students according to the knowledge that they gained in a subject or course, pressure for change has come in at least three forms (Nightingale, Te Wiata, Toohey, Ryan, Hughes & Magin, 1996). The first is a growing need to broaden university education and to develop – and consequently assess – a much broader range of student abilities. The second is the desire to harness the full power of assessment and feedback in support of learning. The third area arises from the belief that education should lead to a capacity for independent judgement and an ability to evaluate one's own performance – and that these abilities can only be developed through involvement in the assessment process (Luckett & Sutherland, 2000).

> Assessment which requires the student only to regurgitate material obtained through lectures and required reading virtually forces the student to use a surface approach to learning that material. On the other hand, assessment which requires the student to apply knowledge gained on the course to the solution of novel problems, not previously seen by the student,… cannot be tackled without a deeper understanding (Entwistle, 1992, p39).

If one adopts an outcomes-based approach to assessment (as is required by SAQA), then one is obliged to state quite explicitly to all stakeholders concerned what knowledge and skills or learning outcomes one is assessing i.e. the assessment criteria. Students' performances are then assessed against these criteria. SAQA requires all qualifications to include *critical outcomes*, which consist of a list of general transferable skills that requires the learner to integrate knowledge, skills and attitudes while carrying out a task in a context of

application. This type of *criterion-referenced* assessment encourages links with teaching and learning. In contrast, in *norm-referenced* assessment, the criteria against which a student's performance is compared with that of his or her peers remain implicit. Criterion-referencing tends to be more transparent because of its explicit statement of criteria. Currently, the trend in assessment is to move towards criterion-referencing. In criterion-referenced education, more time would be spent teaching and testing the student's ability to understand and internalise the criteria of genuine competence (Wiggins, 1989). Criterion-referencing can help establish agreement amongst different assessors, which improves the reliability of the assessment. In order to implement criterion-referenced or outcomes-based assessment, it needs to be clear what the criteria are against which judgements will be made and what will count as evidence for meeting those criteria.

The socio-economic and policy contexts in South Africa have posed enormous challenges for assessment practice in higher education. Contextual criteria have led to the introduction of new assessment policies relating to education and the accreditation of qualifications through a National Qualifications Framework (NQF) (see Chapter 1, p11). Below is an extract from the document entitled "Revisions to the Senate Policy on the assessment of student learning", approved by the Senate of the University of the Witwatersrand, 2006, reflecting the changing nature of university assessment in the South African context.

> Assessment should be unbiased, fair, transparent, valid and reliable (noting that there is some tension between validity and reliability). Valid methods of assessment must be employed in order to sample the range of competencies required of a student graduating from this University, at all levels. In order to do this, depending on the purpose, the use of a variety of assessment forms and methods is recommended and may be carried out throughout the year. Assessment should allow students to demonstrate optimal levels of performance. Appropriate formats must be used for the valid testing of competencies and objectives, and adequate sampling with a variety of examiners over time will assist in reliably testing a variety of competencies. It is

acknowledged, however, that assessment is not an overriding aspect of teaching and learning, but is integral to it.

Therefore the assessment of students should be designed to achieve the following purposes:
- To be an educational tool to teach appropriate skills and knowledge
- To encourage continuous learning and detect learning problems
- To determine whether students are meeting, or have met the educational aims and outcomes of a course (including qualifications exit-level outcomes where appropriate) and to give students continuous feedback on their progress
- To determine levels of competence and to inform students on their current competence
- To facilitate decisions relating to student progress
- To provide a measure of student ability for future employers
- To inform teachers about the quality of their instruction
- To allow evaluation of a course (p2).

This policy is premised on the principles of promoting criterion referencing, which compares performance against specified criteria and encourages links with teaching and learning. There is a responsibility to provide criteria that make explicit the constructs of the teaching and to make these available and accessible to the students in as many different ways as possible. There is a need for flexibility and variety in assessment. The shift to criterion-referenced assessment would allow education to make sound judgements about the comparability of qualifications on the basis of scrutinising assessment criteria and the evidence required for their attainment.

In tertiary education in South Africa, pressure to increase the student intake in higher education as well as to improve throughput has a particularly moral dimension. It implies the need to respond to the historical inequalities of the past, by making the higher education sector accessible to previously disadvantaged black and working class communities. This requires the system to be more open, flexible, transparent and responsive to the needs of under-

prepared, adult, lifelong and part-time learners (Harvey, 1993). This in turn, has implications for appropriate assessment practices in higher education. Such assessment practices would incorporate the use of alternative forms of assessment to provide more complete information about what students have learned and are able to do with their knowledge, and to provide more detailed and timely feedback to students about the quality of their learning.

2.3 ASSESSMENT MODELS IN MATHEMATICS EDUCATION

An assessment model emerges from the different aspects of assessment: what we want to have happen to students in a mathematics course, different methods and purposes for assessment, along with some additional dimensions. The first dimension of this framework is WHAT to assess, which may be broken down into: concepts, skills, applications, attitudes and beliefs.

Niss (1993) uses the term *assessment mode* to indicate a set of items in an assessment model that could be implemented in mathematics education. These items include the following:

- The *subject* of assessment i.e. who is assessed
- The *objects* of assessment i.e. what is assessed
- The *items* of assessment i.e. what kinds of output are assessed
- The *occasions* of assessment i.e. when does assessment take place
- The *procedures* and *circumstances* of assessment i.e. what happens, and who is expected to do what
- The *judging* and *recording* in assessment i.e. what is emphasised and what is recorded
- The *reporting* of assessment outcomes i.e. what is reported, to whom.

For the purpose of this study, the focus will be on the *objects* of assessment in the Niss model outlined above i.e. types of mathematical *content* (including methods, internal and external relations) and which types of student *ability* to deal with that content. This varies greatly with the place, the teaching level and

the curriculum, but the predominant content objects assessed seem to be the following:

[a]   *Mathematical facts*, which include definitions, theorems, formulae, certain specific proofs and historical and biographical data.

[b]   *Standard methods* and *techniques* for obtaining mathematical results. These include qualitative or quantitative conclusions, solutions to problems and display of results.

[c]   *Standard applications* which include familiar, characteristic types of mathematical situations which can be treated by using well-defined mathematical tools.

To a lesser extent, objects of assessment also include:

[d]   *Heuristic* and *methods of proof* as ways of generating mathematical results in non-routine contexts.

[e]   *Problem solving* of non-familiar, open-ended, complex problems.

[f]   *Modelling* of open-ended, real mathematical situations belonging to other subjects, using whatever mathematical tools at one's disposal.
       In mathematics, we rarely encounter

[g]   *Exploration* and *hypothesis generation* as objects of assessment.

With regards to the students' ability to be assessed, the first three content objects require knowledge of facts, mastery of standard methods and techniques and performance of standard applications of mathematics, all in typical, familiar situations.

As we proceed towards the content objects in the higher levels of Niss's assessment model, the level of the students' abilities to be assessed also increase in terms of cognitive difficulty. In the proof, problem-solving, modelling and hypothesis objects, students are assessed according to their abilities to activate or even create methods of proof; to solve open-ended, complex problems; to perform mathematical modelling of open-ended real situations and to explore situations and generate hypotheses.

In the Niss assessment model, objects [a] – [g] and the corresponding students' abilities are widely considered to be essential representations of what mathematics and mathematical activity are really about. The first three objects in the list emphasise routine, low-level features of mathematical work, whereas the remaining objects are cognitively more demanding. Objects [a], [b] and [c] are fundamental instances of mathematical knowledge, insight and capability. Current assessment models in mathematics education are often restricted to dealing only with these first three objects. One of the reasons for this is that methods of assessment for assessing objects [a], [b] and [c] are easier to devise. In addition, the traditional assessment methods meet the requirement of validity and reliability in that there is no room for different assessors to seriously disagree on the judgement of a product or process performed by a given student. It is far more difficult to devise tools for assessing objects [d] – [g]. Inclusion of these higher-level objects into assessment models would bring new dimensions of validity into the assessment of mathematics. Webb and Romberg (1992) argue that if we assess only objects [a], [b] and [c] and continue to leave objects [d] – [g] outside the scope of assessment, we not only restrict ourselves to assessing a limited set of aspects of mathematics, but also contribute to actually creating a distorted and wrong impression of what mathematics really is (Niss, 1993).

Traditional assessment models, have, in many cases, been responsible for hindering or slowing down curriculum reform. We should seek alternative assessment models in mathematics education which at the same time allow us to assess, in a valid and reliable way, the knowledge, insights, abilities and skills related to the understanding and mastering of mathematics in its essential aspects; provide assistance to the learner in monitoring and improving his/her acquisition of mathematical insight and power; assist the teacher to improve his/her teaching, guidance, supervision and counselling and to assist curriculum planners, authorities, textbook authors and in-service teacher trainers in shaping the framework for mathematical instruction, while also saving time. Alternative assessment models, such as the PRQ format, can reduce marking loads for

mathematical educators and assessors, and does provide immediate scores to students.

2.4 ASSESSMENT TAXONOMIES

According to the World Book Dictionary (1990), a *taxonomy* is any classification or arrangement. Taxonomies are used to ensure that examinations contain a mix of questions to test skills and concepts. A leader in the use of a taxonomy for test construction and standardisation was Ralph W. Tyler, the "father of educational evaluation" (Romberg, 1992, p19) who in 1931 reported on his efforts to construct achievement tests for various university courses. He claimed to have found eight major types of *objectives*:

● Type 1: information
● Type 2: reasoning
● Type 3: location of relevant data
● Type 4: skills characteristic of particular subjects
● Type 5: standards of technical performance
● Type 6: reports
● Type 7: consistency in application of point of view
● Type 8: character (Tyler, 1931).

At the time, Tyler neither linked these objectives to specific behaviour nor arranged the behaviour in order of complexity. By 1949, however, he had specified seven types of behavior:

[a]   understanding of important facts and principles
[b]   familiarity with dependable sources of information
[c]   ability to interpret data
[d]   ability to apply principles
[e]   ability to study and report results of study
[f]   broad and mature interests
[g]   social attitudes.

The next step was taken by Benjamin Bloom (1956), who organised the objectives into a taxonomy (dedicated to Tyler) that attempted to reflect the distinctions teachers make and to fit all school subjects. In Bloom's *Taxonomy of educational objectives*, objectives were separated by *domain* (cognitive, affective and psychomotor), related to *educational behaviours*, and arranged in hierarchical order from simple to complex:

- Level 1: Knowledge
- Level 2: Comprehension
- Level 3: Application
- Level 4: Analysis
- Level 5: Synthesis
- Level 6: Evaluation.

Bloom's taxonomy has often been seen as fitting mathematics especially poorly (Romberg, Zarinnia & Collis, 1990). It is quite good for structuring assessment tasks, but Freeman and Lewis (1998) suggest that Bloom's taxonomy is not helpful in identifying which levels of learning are involved. They, however, give an alternative which divides into headings not far removed from Bloom's:

- *Routines*
- *Diagnosis*
- *Strategy*
- *Interpretation*
- *Generation* (Freeman & Lewis, 1998).

As Ormell (1974) noted in a strong critique of the taxonomy, Bloom's categories of behaviour "are extremely amorphous in relation to mathematics. They cut across the natural grain of the subject, and to try to implement them – at least at the level of the upper school – is a continuous exercise in arbitrary choice" (p7). All agree that Bloom's taxonomy has proven useful for low-level behaviours (knowledge, comprehension and application), but difficult for higher levels (analysis, synthesis and evaluation). One problem is that the taxonomy suggests that *lower* skills should be taught before *higher* skills. The fundamental problem is the taxonomy's failure to reflect current psychological

thinking on cognition, and the fact that it is based on "the naïve psychological principle that individual simple behaviours become integrated to form a more complex behaviour" (Collis, 1987, p3). Additional criticisms have questioned the validity of the distinction between cognitive and affective objectives, the independence of content from process and the meaning of objectives isolated from any context (Kilpatrick, 1993). Nevertheless, the view of mental abilities and consequently of mathematical thinking and achievement as organised in a linear, hierarchical way has been powerful in 20th Century assessment practice. It has deep roots in our history and our psyches (Romberg *et al.*, 1990).

Since its publication, variants of Bloom's taxonomy for the cognitive domain have helped provide frameworks for the construction and analysis of many mathematics achievement tests (Begle & Wilson, 1970; Romberg *et al.*, 1990). Attacking behaviourism as the bane of school mathematics, Eisenberg (1975) criticised the merit of a task-analysis approach to curricula, because it essentially equates training with education, missing the heart and essence of mathematics. Expressing concern over the validity of learning hierarchies, he argued for a re-evaluation of the objectives of school mathematics. The goal of mathematics, at whatever level, is to teach students to think, to make them comfortable with problem solving, to help them question and formulate hypotheses, investigate and simply tinker with mathematics. In other words, the focus is turned inward to cognitive mechanism.

Smith *et al.* (1996) propose a modification of Bloom's taxonomy called the MATH taxonomy (Mathematical Assessment Task Hierarchy) for the structuring of assessment tasks. The categories in the taxonomy are summarised in Table 2.1.

Table 2.1: MATH Taxonomy.

Group A	Group B	Group C
Factual knowledge	Information transfer	Justifying and interpreting
Comprehension	Applications in new situations	Implication, conjectures and comparisons
Routine use of procedures		Evaluation

(Adapted from Smith *et al.*, 1996)

In the MATH taxonomy, the categories of mathematics learning provide a schema through which the nature of examination questions in mathematics can be evaluated to ensure that there is a mix of questions that will enable students to show the quality of their learning at several levels. It is possible to use this taxonomy to classify a set of tasks ordered by the nature of the activity required to complete each task successfully, rather than in terms of difficulty. Activities that need only a surface approach to learning appear at one end, while those requiring a deeper approach appear at the other end. Previous studies have shown that many students enter tertiary institutions with a surface approach to learning mathematics (Ball, Stephenson, Smith, Wood, Coupland & Crawford, 1998) and that this affects their results at university. There are many ways to encourage a shift to deep learning, including assessment, learning experiences, teaching methods and attitudinal changes. The MATH taxonomy addresses the issue of assessment and was developed to encourage a deep approach to learning. It transforms the notion that learning is related to what we as educators do to students, to how students understand a specific learning domain, how they perceive their learning situation and how they respond to this perception within examination conditions.

The MATH taxonomy has eight categories, falling into three main groups. The first Group A encompasses tasks which could be successfully done using a surface learning approach. Group A tasks will include tasks which students will have been given in lectures or will have practised extensively in tutorials. In Group B tasks, students are required to apply their learning to new situations, or to present information in a new or different way. Group C encompasses the skills of justification, interpretation and evaluation. Tasks in both Groups B and C require a deeper learning approach for their successful completion. The categories of the taxonomy are context specific. For example, proving a theorem when the proof has been emphasised in class is a Group A task while proving the same theorem *ab initio* is a Group C task. The taxonomy encourages us to think more about our attempts at constructing exercises. Whether we act consciously on this influence or simply make changes

instinctively, it provides a useful check on whether we have tested all the skills, knowledge and abilities that we wish our students to demonstrate (Smith *et al.*, 1996).

Recently, work on how the development of knowledge and understanding in a subject area occurs has led to changes in our view of assessing knowledge and understanding. For example, in Biggs (1991) SOLO Taxonomy (Structure of the Observed Learning Outcome), he proposed that as students work with unfamiliar material their understanding grows through five stages of ascending structural complexity:

Figure 2.1: SOLO Taxonomy.

Prestructural	a stage characterised by the lack of any coherent grasp of the material: isolated facts or skill elements may be acquired.
Unistructural	a stage in which a single relevant aspect of the material or skill may be mastered.
Multistructural	a stage in which several relevant aspects of the material or skills are mastered separately.
Relational	a stage in which the several relevant aspects of the material or skills which have been mastered are integrated into a theoretical structure.
Extended Abstract	the stage of 'expertise' in which the material is mastered both within its integrated structure, and in relation to other knowledge domains, thus enabling the student to theorise about the domain.

(Adapted from Biggs, 1991)

The first three stages are concerned with the progressive growth of knowledge or skill in a quantitative sense, the last two with qualitative changes in the structure and nature of what is learned. (Biggs, 1991, p12). According to Biggs (1991), at one end, knowledge and understanding are simple, unstructured and unsophisticated and of use as support for higher order abilities, while at the other end, they are complex, structured and provide the basis for expert performance. In the light of this opinion, Hughes and Magin (cited in Nightingale *et al.,* 1996) regard assessment of isolated fragments of knowledge appropriate

at the earlier stages (perhaps the first two or three) of Biggs's scheme. Only the assessment of higher order abilities would be appropriate at the later stages.

With increased interest in the assessment of higher order abilities, other classifications to improve and assess learning have been developed. In a project at the Queensland University of Technology, a hierarchy of purposes for setting exercises was proposed to the faculty of a mathematics department. The aim of the project was to encourage faculty members to look more critically at their questions and to relate their questions to learning objectives. A classification according to the lecturer's purpose was conceived as a framework for enabling faculty members to think critically about writing questions and about the signals concerning learning that the questions were sending to their students. This classification according to the lecturer's purpose has been described in Figure 2.2 (Hubbard, 1995).

Figure 2.2: Classification according to lecturer's purpose.

1. To learn a formula, practice manipulation, become familiar with notation, state or prove a standard theorem.

2. Any purpose in 1, but set in a context which is mathematically irrelevant.

3. Apply theory to a problem for which a specific model has been provided, show how the model can be used in different situations.

4. Apply results to new kind of problem, develop problem solving strategies.

5. Prepare for a new concept, lead to the development of a concept or extend a concept.

6. Draw conclusions, generalise, make conjectures, reflect on results.

(Adapted from Hubbard, 1995)

In the Queensland project, it was then decided to separate the classifications in order to emphasise the different ways in which lecturer and student might view the questions. This resulted in the learning-required classification. (Figure 2.3)

Figure 2.3: Learning-required classification.

1. Recognition of key words and
    symbols which trigger memorised,
    standard procedures.
    ⬇
2. Some understanding of standard
    procedures so that they can be
    modified slightly for new situations.
    ⬇
3. Ability to explain and justify procedures
    and to form them into a coherent system.
    ⬇
4. Ability to synthesise mathematical
    experiences into strategies for problem
    solving.

(Adapted from Hubbard,1995)

This learning-required classification is based on Crooks (1988) classification, who regards it as a simplification of Bloom's taxonomy. However Crooks' third category 'critical thinking or problem solving' is divided into two categories. These are essentially critical thinking and problem solving but set in a mathematical context. When applying any taxonomy, the mathematical context is important, because learning objectives which are not subject-specific are more difficult for subject specialists to apply.

If we analyse the goals of mathematics education, different levels can be distinguished. A possible categorisation of them is described by Jan de Lange (1994). Because the assessment has to reflect education, these categories can be used both for the goals of mathematics education in general and for the assessment. De Lange (1994) represents the levels of understanding in the form of a pyramid as shown in Figure 2.4.

Figure 2.4: De Lange's levels of understanding.

(Adapted from De Lange, 1994)
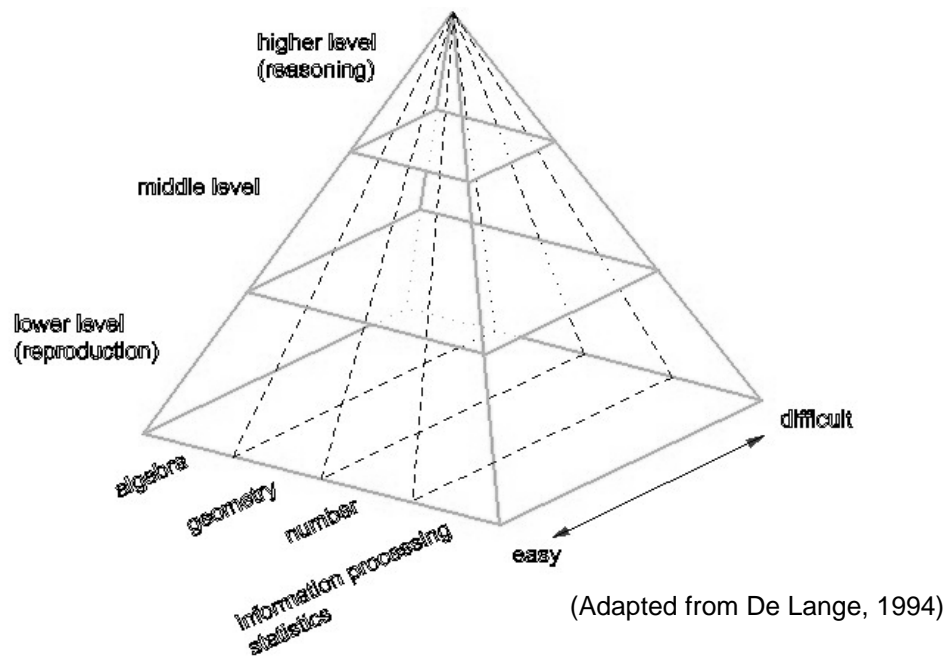
The lower level

This level concerns the knowledge of objects, definitions, technical skills and standard algorithms.

Some typical examples are:

●   adding (easy) fractions

●   solving a linear equation with one variable

●   measuring an angle using a compass

●   computing the mean of a given set of data.

According to De Lange's categorisation, most of traditional school mathematics and traditional tests seem to be at the lower level. One might think that a question at the lower level will be easier than a question at one of the other two levels. But this need not be the case. A question at the lower level can be a difficult one. The difference is that it does not demand much insight; it can be solved by using routine skills or even by rote learning.

The second level

The second level can be characterised by having students relate two or more concepts or procedures. Making connections, integration and problem solving are terms often used to describe this level. Also problems that offer different strategies for solving, or offer more than one approach to solve, are at this level.

For questions at this level careful reading and some good reasoning are needed. There is quite a lot of information to read and students have to make decisions about their selection of strategies.

The third level

The highest level has to do with complex matters like mathematical thinking and reasoning, communication, critical attitude, communication, creativity, interpretation, reflection, generalisation and mathematising. Students' own constructions are a major component of this level.

Assessing content knowledge and understanding, usually at the lower levels of any taxonomy, is often assumed to be far less problematic than assessing the higher order skills and abilities at the higher taxonomy level. Academic staff have a long familiarity with conventional methods of assessing knowledge and understanding, and texts on how to assess knowledge have been in existence for many years (Ebel, 1972; Gronlund, 1976; Heywood, 1989; McIntosh, 1974). However, several researchers of student learning (Dahlgren, 1984; Marton & Saljö, 1984; Ramsden, 1984) have identified an alarming phenomenon whereby numerous students who have done well in examinations intended to test understanding, have been found to still have fundamental misconceptions about basic underlying principles and concepts on which they were supposed to have been tested.

> Some of the most profoundly depressing research on learning in higher education has demonstrated that successful performance in examinations does not even indicate that students have a good grasp of the very concepts which staff members believed the examinations to be testing (Boud, 1990, p103).

In the interests of higher quality tertiary education, a deep approach to learning mathematics is to be valued over a surface approach (Smith *et al.,* 1996). Students entering university with a surface approach to learning should be encouraged to progress to a deep approach. Studies have shown (Ball *et al.,* 1998), that students who are able to adopt a deep approach to study tended to achieve at a higher level after a year of university study.

2.5 ASSESSMENT PURPOSES

Although we appreciate that assessment can have enormous value as a tool for learning and that it provides important data for review, management and planning, we also need to examine different theories of assessment. Different assessment purposes require different assessment theories. There is general agreement that assessment in an educational context can be grouped under three broad traditional purposes: *Diagnostic assessment*, *Formative assessment* and *Summative assessment*, with *Quality assurance* having been added more recently. These will now be defined and discussed in more detail.

2.5.1 Diagnostic assessment

The purpose of diagnostic assessment is to determine the learner's strengths and weaknesses and to determine the learner's prior knowledge (Geyser, 2004). Diagnostic assessment can also be used to determine whether a student is ready to be admitted to a particular learning program and to determine what remedial action may be required to enable a student to progress.

2.5.2 Formative assessment

Boud in Geyser (2004) defines formative assessment as:

> …focused on learning from assessment. Formative assessment refers to assessment that takes place during the process of learning and teaching – it is day-to-day assessment. It is designed to support the teaching and learning

33

process and assists in the process of future learning. It feeds directly back into the teaching-learning cycle. The learner's weaknesses and strengths are diagnosed and (immediate) feedback is provided. It helps in making decisions on the readiness of the learners to do summative assessment. It is developmental in nature, therefore credits of certificates are not awarded (SAQA, 2001, p93).

According to Biggs (2000), the critical feature of formative assessment is the feedback that is given to the students. This feedback is aimed at improving the learning of the student as well as the teaching of the lecturer, motivating students, consolidating work done to date and provides a profile of what a student has learnt.

All formative assessment is diagnostic to a certain degree. Diagnostic assessment is an expert and detailed enquiry into underlying difficulties, and can lead to radical re-appraisal of a learner's needs, whereas formative assessment is more developmental in assessing problems with particular tasks, and can lead to short-term and local changes in the learning work of a learner. Formative learning provides a model for self-directed learning and hence for intellectual autonomy (Brown & Knight, 1994). Students are encouraged to be more autonomous in appraising their performances, learning to be more reflective and to take responsibility for their own learning.

Because formative assessment is intended as the feedback needed to make learning more effective, it cannot simply be added as an extra to a curriculum. The feedback procedures, and more particularly their use in varying the teaching and learning programme, have to be built into the teaching plans, which thereby will become both more flexible and more complex.

The integration of feedback into the curriculum is emphasised very strongly by Linn (1989):

> …the design of tests useful for the instructional decisions made in the classroom requires an integration of testing and instruction. It also requires a clear conception of the curriculum, the goals, and the process of instruction. And it

requires a theory of instruction and learning and a much better understanding of the cognitive processes of learners (p5).

The quote shows how much needs to be done with our current assessment system. Astin (1991, p189) was certain that 'the best principles of assessment and feedback are seldom followed or applied in the typical lower-division undergraduate course'. It seems that there is little scope for formative assessment because too many assessments (especially examinations) do not lead to feedback to the students. In addition, there is the problem of continuous assessments placing increased pressure on staff time with an increase in marking loads. There is also dissatisfaction with the quality of feedback which students often get. These problems are all compounded by the fact that undergraduate classes in tertiary mathematics are usually very large. Large student numbers not only place pressure on administration and marking loads, but also on the effectiveness and quality of feedback to the students. A major improvement in assessment systems would be to examine departmental policies for generating feedback to students. There is a shortage of research into the way that students use the feedback that they do get. The practice of formative assessment must be closely integrated with curriculum and pedagogy and is central to good quality teaching (Linn, 1989).

2.5.3 Summative assessment

The term *'summative'* implies an overview of previous learning. Summative assessment is used to grade students at the end of a unit, or to accredit at the end of a programme (Biggs, 2000). Summative assessment is used to provide judgement on students' achievements in order to:

- establish a student's level of achievement at the end of a programme
- grade, rank or certify students to proceed to or exit from the education system
- select students for further learning, employment, etc
- predict future performance in further study or in employment
- underwrite a 'license to practise' (Brown & Knight, 1994, p16).

The overview of previous learning involved in summative assessment could be obtained by an accumulation of evidence collected over time, or by test procedures applied at the end of the previous phase which covered the whole area of the previous learning. Beneath the key phrases here of 'accumulation' and 'covered', lies the problem of selecting that information which is most relevant for summative purposes. It is through summative assessment that educators exert their greatest power over their students.

Because the purposes of assessment often remain vague and implicit, there is a danger that the different assessment purposes, i.e. summative, formative or diagnostic become confused and conflated and as a consequence, assessment often fails to play a truly educational role (Harlen & James, 1997). For example, an over-stretched lecturer may set a test for formative purposes and then, through lack of time and energy, decide to use the results for summative purposes. Not only is this kind of practice unfair to students, but it also undermines the developmental potential of assessment. Students are entitled to be informed beforehand how their assessment results will be used. A further consequence of confusing the different purposes of assessment is that lecturers sometimes assume that they can add up a series of formative assessment results (eg. classmarks) in order to make a summative judgement. In assessing students it is advisable to keep the formative and summative purposes separate. This is because the reliability concerns of summative assessment are far greater than they are for formative assessment and confusion of the two may result in unfair assessment practices. A common and legitimate practice is to use the evidence derived from formative assessment indirectly to inform professional judgements made about students in difficult summative circumstances. The cycle of formative and summative assessment as illustrated in Figure 2.5 (Makoni, 2000) suggests that rather than understanding the formative and summative purposes of assessment as dichotomous, we should view them as two ends of a continuum (Brown, 1999).

Figure 2.5: Cycle of formative and summative assessment

certification

establishment of learning
outcomes & learning contract

summative
assessment

feedback

the learning
process

feedback

formative
assessment

evidence gathering

interpreting & recording
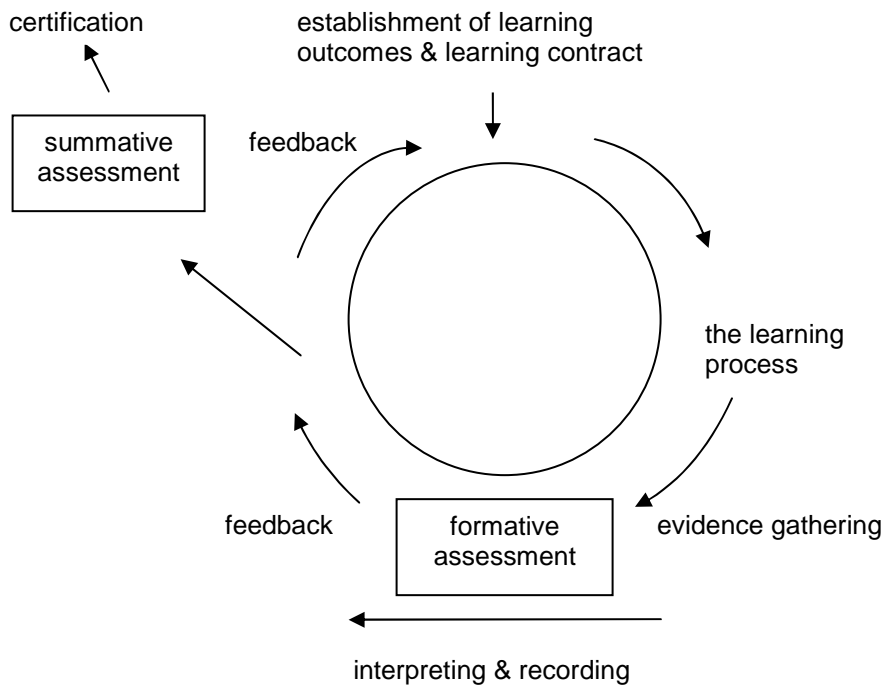
(Adapted from Luckett & Sutherland, 2000, p112)

2.5.4 Quality assurance

One further purpose of assessment needs to be mentioned, and that is how assessment contributes to institutional management. Summative (and to a lesser extent formative) assessment can also be used for quality assurance of the educational system. Here assessment is used to provide judgement on the educational system in order to:

- provide feedback to staff on the effectiveness of their teaching
- assess the extent to which the learning outcomes of a programme
  have been achieved
- evaluate the effectiveness of the learning environment
- monitor the quality of an education institution over time (Brown, Bull & Pendlebury, 1997; Yorke, 1988).

Although often neglected, this type of assessment is crucial. Erwin (1991, p119) said that "for the typical faculty [lecturer] or student affairs staff member, the

major value of assessment is to improve existing programmes". The results of assessment and testing for accountability should be presented and communicated so that they can serve the improvement of educational institutions.

2.6 SHIFTS IN ASSESSMENT

There are tensions between the different purposes of assessment and testing, which are often difficult to resolve, and which involve choices of the best agencies to conduct assessments and of the optimum instruments and appropriate interpretations to serve each purpose. For example, if we are clear on the purpose of each assessment we design, then we will be in a position to make sound judgements about 'the what' and 'the how' of the assessment instrument. Finally, it is worth noting that assessment, together with face-to-face teaching, course design, course management and course evaluation, is part of the generic task of teaching. The phrase 'teaching, learning and assessment' often makes assessment look like an afterthought or at least a separate entity. In fact, teaching and feedback (formative assessment) merge, while assessment is an ongoing and necessary part of helping students to learn.

Geyser (2004) summarises the paradigm shift that is currently under way in tertiary education as follows:

> Traditionally, assessment has been almost entirely summative in nature, with a final explanation and educator as the sole and unconditional judge. Traditional assessments have often targeted a learner's ability to demonstrate the acquisition of knowledge (that is, achievement), but new methods are needed to measure a learner's level of understanding within content area and the organization of the learner's cognitive structure (that is, learning). The main shift in focus can be summarised as a shift away from assessment as an add-on experience at the end of learning, to assessment that encourages and supports deep learning. It is now important to distinguish between learning *for* assessment and learning *from* assessment as two complementary purposes of assessment (p90).

This shift means that we need to move away from assessing how well students can reproduce content knowledge, towards a situation where we learn how to assess the integration and application of knowledge skills, and maybe even attitudes in unfamiliar as well as familiar contexts. Taking this idea one step further, Luckett and Sutherland (2000) are of the opinion that:

> Conventional ways of assessing students such as the unseen three hour exam, are no longer adequate to meet these demands. We can no longer justify testing again and again the same restricted range of skills and abilities; we can no longer get away with simply requiring students to write about performance, instead of getting them to perform in authentic contexts (p201).

New trends in assessment in higher education demand that we begin to assess generic and applied competencies as well as traditional knowledge bases. Hence the need to collect evidence, via assessment, that shows how well (or badly, or if at all) our students have been able to understand, integrate and apply the knowledge, skills and values specified in our course outcomes. A shift in assessment is related to a shift between the types of assessment discussed in section 2.5. We will have to be innovative and try out a range of new assessment approaches and methods, ensuring that we do indeed assess all of our intended learning outcomes and that our assessments add value to students' learning.

> Assessment will be seen as natural and helpful, rather than threatening and sometimes a distraction from real learning as in traditional models (Jessup, 1991, p136).

2.7   ASSESSMENT APPROACHES

Assessment approaches work best where learning outcomes have been articulated in advance, shared with students and assessment criteria agreed. Questions about the purpose of assessment arise, especially questions related to formative as opposed to summative purposes. Assessment approaches which are integrated into a course, not 'bolted-on' are desirable – this implies both staff and curriculum development.

Before going on to describe alternative *question formats,* I will briefly outline a range of *assessment approaches* which are important to think about prior to selecting a specific method and designing a specific instrument. A number of different methods may be appropriate to any one approach, or combination of approaches, depending on one's purpose, learning outcomes and teaching and learning context.

2.7.1   The traditional approach

In the traditional approach it is taken for granted that assessment follows teaching and that the aim of assessment is to discover how much has been learned.

Here the lecturer or examiner is usually considered to be the only legitimate assessor. Students are assessed strictly as individuals in competition with each other in a highly controlled environment and strict measures to avoid cheating are employed. Learning is viewed quantitatively in terms of the amount of teaching which has been absorbed. There is little interest in the specifics of which questions has been correctly answered. Common methods used in this approach include examinations, essays, pen-and paper tests and reports.

Literature review has revealed that more recently certain interesting alternative approaches to assessment in undergraduate mathematics have been explored (Cretchley & Harman, 2001; Anguelov, Engelbrecht & Harding, 2001; Hubbard, 2001; Wood & Smith, 2001). In the overview of approaches that follow, innovative variations will be discussed.

2.7.2   Computer-based (online) assessment

In an age of increasing access to computers and to university education, new technologies have become an exciting medium for the delivery and assessment of courses at the tertiary level.

There can be no doubt that increasing technological support for much that had to be done by hand, will not only impact on the way we do mathematics, but even determine the very nature of some of the mathematics that we do (Cretchley & Harman, 2001, p160).

Engelbrecht and Harding (2004) found that 'many teachers of mathematics still shy away from granting technology the same significant role in the assessment process' (p218).

The following statement by Smith (as cited in Anguelov, Engelbrecht and Harding, 2001) is very descriptive with regard to the motives for technological forms of assessment:

> Courses in mathematics that ignore the impact of technology on present and future practices of science, engineering and mathematics perpetrate a fraud upon our students. Technology should be used not because it is seductive, but because it can enhance mathematical learning by extending each student's mathematical power. Calculators and computers are not substitutes for hard work, but challenging tools to be used for productive ends (p190).

The use of computers in assessment can solve the problem of providing detailed, individualised feedback to large student numbers. This approach is often based on a mastery learning model, in which students receive immediate feedback and can repeat or progress at their own pace. In a study conducted by Senk, Beckmann and Thompson (1997), teachers pointed out that technology allowed them to deal with situations that would have involved tedious calculations if no technology had been available. They explained that "not-so-nice", "nasty", or "awkward" numbers arise from the need to find the slope of a line, the volume of a silo, the future value of an investment or the 10th root of a complex number. Additionally, some teachers of Algebra II classes noted how technology influenced them to ask new types of questions, how it influenced the production of assessment instruments and how it raised questions about the accuracy of results (Senk, Beckmann & Thompson, 1997, p206).

I think you have to ask different kinds of things… When we did trigonometry, you just can't ask them to graph $y = 2 \sin x$ or something like that. Because their calculator can do that for them… I do a lot of going the other way around. I do the graph, and they write the equation… The thing I think of most that has changed is just the topic of trigonometry in general. It's a lot more application type things…given some situation, an application that would be modeled by a trigonometric equation or something like that [Ms. P].

I use it [the computer] to create the papers, and I can do more things with it…not just hand-sketched things. I can pull in a nice polynomial graph from *Mathematica*, put it on the page, and ask them questions about it. So, in the way, it's had a dramatic effect on me personally… We did talk about problems with technology. Sometimes it doesn't tell you the whole story. And sometimes it fails to show you the right graph. If you do the tangent graph on the TI-81, you see the asymptotes first. You know, that's really an error. It's not the asymptote [Mr. M].

The role of information technology in educational assessment has been growing rapidly (Barak & Rafaeli, 2004; Beichner, 1994; Hamilton, 2000). The high speed and large storage capacities of today's computers makes computerised testing a promising alternative to paper-and-pencil measures. Assessment tasks should include life-like, authentic or situated activities (Cumming & Maxwell, 1999). For many disciplines, including mathematics, computer technology can be seen as part of such a context (Groen, 2006). Web-based testing systems offer the advantages of computer-based testing delivered over the Internet. The possibility of conducting an examination where time and pace are not limited, but can still be controlled and measured, is one of the major advantages of web-based testing systems (Barak & Rafaeli, 2004; Engelbrecht & Harding, 2004). Other advantages include the easy accessibility of on-line knowledge databases and the inclusion of rich multimedia and interactive features such as colour, sound, video and simulations. Computer-based online assessment systems offer considerable scope for innovations in testing and assessment as well as a significant improvement of the process for all its stakeholders, including teachers, students and administrators (McDonald, 2002). In a web-based study

conducted by Barak and Rafaeli (2004), MBA students carried out an online Question-Posing Assignment (QPA) that consisted of two components: Knowledge Development and Knowledge Contribution. The students also performed self- and peer-assessment and took an online examination. Findings indicated that those students who were highly engaged in online question-posing and peer-assessment activity received higher scores on their final examination compared to their counter peers. The results provide evidence that web-based activities can serve as both learning and assessment enhancers in higher education by promoting active learning, constructive criticism and knowledge sharing.

Online assessment holds promise for educational benefits and for improving the way achievement is measured. Computer technology has come to play central roles in both learning objectives and instructional environment in tertiary mathematics. While the use of online assessment may seem a logical progression in this regard, it is perhaps not as widely used as it could be. Online assessment can be a valuable investment with efficiencies in marking, administration and resource use (Engelbrecht & Harding, 2004; Greenwood, McBride, Morrison, Cowan & Lee, 2000; Lawson, 1999). In a study conducted by Groen (2006) in the Department of Mathematical Sciences, University of Technology, Sydney, Australia, it was found that marking of computer-based tests was no more time-consuming than marking a paper-based test. Feedback was individualised, easy to supply and immediately accessible to students. Further, copying appeared no more or less possible than for a paper test. In addition, question item banks provided a valuable record of the components of assessment and provide a library of questions. Appropriate design of online assessments tasks and support activities can also foster other positive learning outcomes including competence in the use of, written and electronic communication, critical though, reasoned arguments, problem solving and information management, as well as the ability to work collaboratively. Further online assessment offers an authentic environment under which to assess the computer laboratory skills that feature strongly in many mathematics subjects and in professional practice (Groen, 2006).

2.7.3 Workplace- and community-based/learnership assessment

Where employers are increasingly involved in workplace- and community-based learning and assessment, as is the case with nursing, social work, teaching and tailor-made programmes, employers are more involved in assessment issues, often coming to realise how complex and costly they can be. The workplace- and community-based learnership assessment approach gives students an opportunity to apply their knowledge and skills in a real-world context and to learn experientially. This approach is considered highly beneficial for the development of professional skills and competences as opposed to the learning of knowledge and theory in isolation from context or application. Typically, in such approaches, supervisors or mentors assess performances, but students are also required to submit a written report or portfolio to their lecturer (Brown & Knight, 1994).

2.7.4 Integrated or authentic assessment

Concerns about validity heralded the new era in assessment dating from the 1960s to the present. From the beginning of the historical record to the nineteenth century, measurement in education was quite crude. During the nineteenth century, educational measurement began to assimilate, from various sources, the ideas and the scientific and statistical techniques which were later to result in the psychometric testing period, dating from about 1900 to the 1960s. Dating from the 1960s to the present is the policy-programme evaluation period. Tyler's model of evaluation in education prevailed until the 1970s, when his approach was found inadequate as a guide for policy and practice.

The earliest signs of the new era in assessment were small shifts away from *norm-referenced* towards *criterion-referenced* assessment. The standardised norm-referenced test based on behaviourism assures that one knows isolated pieces of knowledge. Such a test asks students to respond to a variety of questions about specific parts of mathematics, some of which the student knows

and some not. Responses are processed by summing the number of correct responses to indicate how many parts of mathematical knowledge a student possesses and the totals for an individual student compared to those of other students. Criterion-referenced assessment is also based on behaviourism (Niss, 1993). However, criterion-referenced assessment establishes standards (criteria) for specific grades or for passing or failing. So a student who meets the criteria gets the specified result. Competency standards may be used as the basis of criteria-referenced assessment. Mastery learning is another example: students must demonstrate a certain level of achievement or they cannot continue to the next stage of a subject or program of study. The goal is for everyone to meet an established standard.

The problem with both approaches is that neither yields information about the inter-relationships among the parts of knowledge held by a student. Both approaches can reinforce the idea that mere right answers are adequate signs of achievement. What is required is authentic assessment: 'contextualised complex intellectual challenges, not fragmented and static bits or tasks' (Wiggins, 1989, p711). Authentic assessment (Lajoie, 1991), based on constructivist notions, begins with complex tasks which students are expected to work on for some period of time. Their responses are not just answers; instead they are arguments which describe conjectures, strategies and justifications.

Integrated assessment calls on the students to demonstrate that they are:

> …able to pull together and integrate the different bits of information, skills and attitudes that they have developed from across a [whole qualification] as a whole. Integrated assessment therefore involves the design and judgement of learner performances that can be used as evidence from which to infer capability (the integration of theory and practice) and to demonstrate that the purposes of a programme as a whole has been achieved (Luckett & Sutherland in Makoni, 2000, p111).

An authentic test not only reveals student achievement to the examiner, but also reveals to the test-taker the actual challenges and standards of the field (Wiggins, 1989). To design an authentic test, we must first decide what the

actual performances are that we want students to be good at. Authentic assessments can be developed by determining the degree to which each student has grown in his or her ability to solve non-routine problems, to communicate, to reason and to see the applicability of mathematical ideas to a variety of related problem situations (Niss, 1993). In other words, authentic assessment tasks call on students to demonstrate the kind of skills that they will need to have in the 'real world'. Baron and Boschee (1995) argue that authentic assessment relates to assessing complex performances and higher-order skills in real-life contexts:

> Authentic assessment is contextualised, involves complex intellectual changes, and does not involve fragmented and static bits or tasks. The learner is required to perform real-life tasks (p25).

Authentic assessment is performance-based, realistic and set within contexts that students will encounter beyond the educational setting.

Learning is multidimensional and integrated. Integrated assessment is needed to ensure that students can bring together and integrate all the knowledge, skills and attitudes they have gleaned from a programme as a whole. Outcomes-based education requires integrated assessment of competence, which is described as consisting of three dimensions:

- knowledge/*foundational* competence – knowing and understanding what and why

- skills/*practical* competence – knowing how, decision making ability; and

- attitudes and values/*reflexive* competence – the ability to learn and adapt through self-reflection and to apply knowledge appropriately and responsibly (Luckett & Sutherland, 2000, p111).

Reflexive competence is the ability to integrate performance and decision making with understanding and with the ability to adapt to change and unforeseen circumstances, and to explain the reasons behind these adaptations.

Authentic or integrated assessment is particularly appropriate for professional and applied courses. It should be used throughout the curriculum, particularly at the degree exit level. It may also be used at modular level in order to ensure that the specific learning outcomes listed in course outlines are achieved holistically. A scaffolded research project in the discipline is the primary vehicle for this to happen. This could integrate skills from across various disciplines. Diagrammatically, this can be represented as:

Figure 2.6: Integrated assessment.

Skills/practical competence
(knowing how, decision-making ability)

Knowledge/foundational competence
(knowing & understanding what and why)

Integrated assessment of knowledge in use

Attitudes & values/ reflexive competence
(the ability to learn and adapt through self-reflection and to apply knowledge appropriately and responsibly)
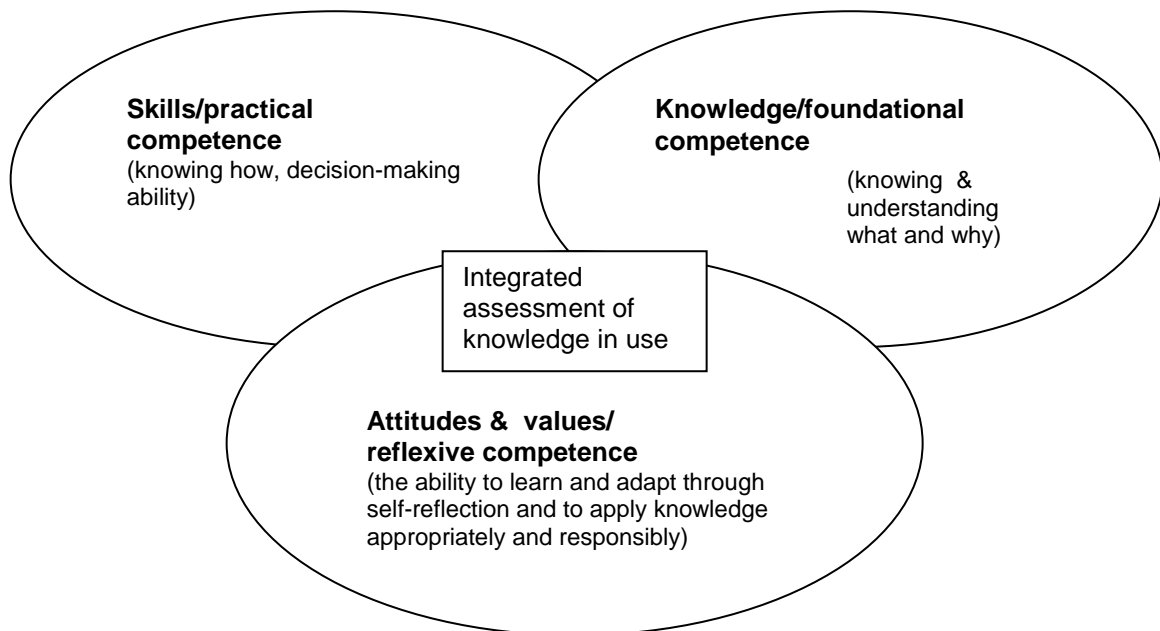
(Adapted from Luckett & Sutherland, 2000, p111)

The controversy about this sort of assessment is centred primarily around its reliability. For assessment to be reliable, it should yield the same results if it is repeated, or different markers should make the same judgements about students' achievements. Because integrated assessment involves a complex task with many variables, the judgement of the overall quality of the performance is more likely to be open to interpretation than an assessment of a simpler task. In a truly authentic and criterion-referenced education, more time would be spent teaching and testing the student's ability to understand and internalise the criteria of genuine competence than in a norm-referenced situation. In higher education, it does not necessarily mean a shift to more external forms of assessment, but it will mean that the unquestioned relationship between a

course and the assessment 'which forms part of it' will be open to critical scrutiny from an outcomes-oriented perspective. The positive aspect is that assessment will be related to outcomes in a discipline which can be publicly justified to colleagues, to students and to external bodies. We are now seeing moves to a holistic conception: no longer can we think of assessment merely as the sum of its parts, we need to look at the impact of the total package of learning and assessment (Knight, 1995). The assessment challenge we face in mathematics education is to give up old, traditional assessment methods to determine what students know which are based on behavioral theories of learning, and develop integrated or authentic assessment procedures that reflect current epistemological beliefs about what it means to know mathematics and how students come to know.

2.7.5   Continuous assessment

Continuous assessment takes place concurrently with, and is often integrated into, the teaching/learning unit at issue. This approach involves assessing students regularly in a manner that integrates teaching and assessment; it uses feedback from each assessment to inform further teaching and the construction of the next assessment. It is usually formative and developmental in purpose, using a range of assessment methods in which the lecturer is not always the sole judge of quality. Its primary purpose is to inform students (and their parents) about their performance so as to help them control and adjust their learning activity. An almost equally important purpose is to inform the teacher about the outcome of his/her teaching in general in order to adjust it if desirable – and specifically in relation to the individual student in order to advise and influence his/her actual or potential association with mathematics. Continuous assessment suggests a cyclical process through which a multi-facetted, holistic understanding of the learner can be developed. If used summatively, continuous assessment should involve summing up the evidence about a learner through the exercise of professional judgement. It should not simply mean adding up a series of test marks that are all given equal weight (Luckett & Sutherland, 2000).

2.7.6   Group-based assessment

This approach recognises that all learning takes place in a social context and that professional identity is best developed through interaction with a community of professionals. In this approach, students are required to work in teams. They may be assessed as a group or individually.   This approach allows one to assess the learning process as well as its product. In group-based assessment, the assessor relies on peer-assessment to tap into attitudes and skills such as accountability, effort and teamwork. A typical approach is to calculate the final mark as the sum of a peer mark for process and a group mark for product. Peers allocate a mark to each individual in the group for process skills and the lecturer allocates a group mark for the learning product (Luckett & Sutherland, 2000).

2.7.7   Self-assessment

Assessment systems that require students to use higher-order thinking skills such as developing, analysing and solving problems instead of memorising facts are important for the learning outcomes (Zohar & Dori, 2002). Two of these higher-order skills are reflection on one's own performance – *self-assessment*, and consideration of peers' accomplishments – *peer assessment* (Birenbaum & Dochy, 1996; Sluijsmans, Moerkerke, van-Merrienboer & Dochy, 2001). Both self- and peer-assessment seem to be underrepresented in contemporary higher education, despite their rapid implementation at all other levels of education (Williams, 1992). Larisey (1994) suggested that the adult student should be given opportunities for self-directed learning and critical reflection in order to mirror the world of learning beyond formal education.

In the self-assessment approach students are invited to assess themselves against a set of given or negotiated criteria, usually for formative purposes but sometimes also for summative purposes. The aim of this type of assessment is to provide students with opportunities to develop the skills of thoughtful, critical

self-reflection. Self-assessment gives students a greater ownership of the learning they are undertaking. Assessment is not then a process done to them, but is a participative process in which they are themselves involved. This in turn tends to motivate students, who feel they have a greater investment in what they are doing.

Self-assessment can be a central aspect of the development of lifelong learning and professional competence, particularly if students are involved in the generation and development of the assessment criteria and are required to justify the marks they give themselves (Boud, 1995). Self-assessment has proved to be an excellent means of getting students to take responsibility for their own learning and to become more reflective and effective learners (Luckett & Sutherland, 2000). Boud (1995) developed this further by arguing that traditional assessment practices neither matched the world of work, nor encouraged effective learning. "Self-assessment", he argued, "is fundamental to all aspects of learning. Learning is an active endeavour and thus it is only the learner who can learn and implement decisions about his or her own learning: all other forms of assessment are therefore subordinate to it" (Boud, 1995, p109).

On graduation, students will be expected to practice self-evaluation in every area of their lives, and it is a good exercise in self-development to ensure that these abilities are extended (Brown & Knight, 1994). The goal of self-assessment is to promote the reflective student, one who has a degree of independence and who is therefore well placed to be a lifelong learner.

2.7.8 Peer-assessment

In peer-assessment students are involved in assessing their peers using a wide range of assessment methods, always under the guidance of the lecturer. The lecturer acts more as an external examiner, checking for reliability and is ultimately responsible for the final allocation of marks.

Criterion-referenced assessment makes this approach possible: the explaining, discussing and even negotiating of the assessment criteria and what will count as evidence for their attainment can be an extremely valuable learning experience for students. Using peer-assessment makes the process much more one of learning, because learners are able to share with one another the experiences that they have undertaken. For peer-assessment, ideas can be interchanged and effective learning will take place (Luckett & Sutherland, 2000).

Experiencing peer-assessment seems to motivate deeper learning and produces better learning outcomes (Williams, 1992). Peer-assessment can deepen students understanding of the subject, develop their evaluative and reflective skills and their groupwork and task management skills. Peer-assessment is probably the best means of assessing how individual students work in teams. Given the importance which employers put upon the ability to work as part of a team, it is important that learners in higher education are exposed to situations which require them to respond sensitively and perceptively to peers' work.

> Through peer-assessment students would be learning, which is, as we repeatedly argue, the main purpose of assessment (Brown & Knight, 1994, p60).

2.8   QUESTION FORMATS

New forms of assessment and question formats are not goals in and of themselves. The major rationale for diversifying mathematics assessment is the value that the diversification has as a tool for the improvement of our teaching and the students' learning of mathematics. Lynn Steen in *Everybody Counts* (Mathematical Sciences Education Board, 1989, p57) makes the point that 'skills are to mathematics what scales are to music or spelling is to writing. The objective of learning is to write, to play music, or to solve problems – not just to master skills'. As assessment policies change, so too must our assessment practices and instruments. Mathematics tests cannot only be vehicles used to assess the memorisation and regurgitation of rote skills. Assessment driven by

problems and applications will naturally subsume the more routine skills at the lower levels of thinking. Again from *Everybody Counts*, we know that:

> Students construct meaning as they learn mathematics. They use what they are taught to modify their prior beliefs and behaviour, not simply to record the story that they are told. It is students' acts of construction and invention that build their mathematical power and enable them to solve problems they have never seen before (p59).

Today's needs demand multiple methods of assessment, integrally connected to instruction, that diagnose, inform and empower both teachers and students.

2.9 CONSTRUCTED RESPONSE QUESTIONS AND PROVIDED RESPONSE QUESTIONS

Questions used for assessment can be classified into two broad categories – *Constructed Response Questions* (CRQs) where students have to construct their own response and *Provided Response Questions* (PRQs) where the student has to choose between a selection of given responses. This terminology was introduced by Engelbrecht and Harding in 2003. In a constructed response format, the student produces a product such as a case study report or lab study, engages in a process or performance such as a social work interview or a musical performance, or exhibits a personal trait such as some leadership ability (Engelbrecht & Harding, 2003; Haladyna, 1999). In mathematics, CRQs or free-response items (Braswell & Jackson, 1995) include questions in open-ended format (Bridgeman, 1992), essays, projects, short answer questions (paper-based or online), portfolios and paper-based or online assignments. Communication in mathematics has become important as we move into an era of a *thinking curriculum* (Stenmark, 1991). In a constructed response format, writing in mathematics becomes vital. Mathematics writing may take on many forms. It may be a separate activity, or may be part of a larger project. Journals, reports of investigations, explanations of the processes used in solving a problem, portfolios or responses to CRQs all become part of what students do daily in the mathematics class as well as what is reviewed for

assessment purposes. The traditional three-hour, unseen constructed response *examination* constitutes an important component of any undergraduate mathematic assessment programme. However, where clear criteria are absent, the marking of such examinations for summative purposes is unreliable (Luckett & Sutherland, 2000) and time-consuming. Methods of assessment within the examination framework can be varied to assess a wider range of cognitive skills and to achieve higher levels of reliability. For example, short answer questions are easier to mark reliably, can be designed to test a wide range of knowledge and are not that time consuming to mark; assignments in which students are given a specified period to deliver a product are closer to real-world conditions and allow more time for thought; open-book examinations and tests are also more authentic and assess what students can do with information.

> Examinations can be used as opportunities for problem-solving if an unseen exam question is, for example, linked to case studies that require students to apply the material that they have had to prepare for the examination to different situations (Hounsell, McCulloch & Scott, 1996, p115).

In a provided response or fixed-response format (Ebel & Frisbie, 1986; Osterlind, 1998; Wesman, 1971), the student chooses among available alternatives. PRQs include multiple choice questions (MCQs), multiple-response questions, matching questions, true/false questions, best answers and completing statements. A true/false question can be classified as a particular type of two option multiple choice. Matching questions, in which students are asked to match items, can be designed to test knowledge and reasoning. In the 'complete the statement' type of PRQ, the student is given an incomplete statement. He/she must then select the choice that will make the completed statement correct. PRQs are sometimes referred to as *objective* tests, and such tests, far from diminishing the curriculum or distorting teaching, enable teachers to diagnose learners' difficulties and individualise their instruction (Kilpatrick, 1993). Others argue that objective tests have driven other forms of assessment out of academic institutions, trivialised learning and warped instruction (Resnick, 1987; Romberg *et al.*, 1990). A common concern is that the use of PRQs encourages rote learning and memorising of discrete bits of information, rather

than developing an overall deeper understanding of the topic. Many examples exist of PRQs, however, that emphasise understanding of important mathematical ideas and generally involve integrating more than one mathematical concept (Gibbs, Habeshaw & Habeshaw, 1988; Lawson, 1999; Johnstone & Ambusaidi, 2001; Smith *et al.*, 1996). This discussion will be expanded on in subsequent sections.

In a study conducted by Engelbrecht and Harding (2003), it is reported that students at the University of Pretoria performed better in online PRQs than in online CRQs, on average, and better in paper CRQs than in online CRQs. It was thus recommended that it is important to use a combination of question types when setting an online paper. In contrast to paper CRQs, online CRQs also mostly have the problem of little or no partial credit. Various strategies have been developed to adapt PRQs to give credit for partial knowledge (Friel & Johnstone, 1978), to reduce the effect of guessing (Harper, 2003) and to find indications of reasoning paths of students.

CRQs offer at least three major advantages over PRQs. Firstly, they reduce measurement error by eliminating random guessing. Secondly, they allow for partial credit for partial knowledge and thirdly, problems cannot be solved by working backwards from the answer choices. Because this last advantage makes test items more like the kind of problems students must solve in their academic work, this enhances the face validity of the test. A review by Traub and Rowley (1991) suggests that there is evidence that some free-response essay tests measure different abilities from those measured by fixed-response tests, but that when the free response is a number or a few words, format differences may be inconsequential. Another study that focused on mathematical reasoning (Traub & Fisher, 1977) found that there was no evidence that provided response and constructed response mathematics tests measured different traits in eighth-grade students. Martinez (1991) found that constructed response versions of questions that relied on figural and graphical material were more reliable and discriminating than parallel provided response questions. Bridgeman (1992) found that at the level of the individual item, there

were striking differences between the constructed response format and the provided response format. Format effects appeared to be particularly large when the PRQs were not an accurate reflection of the errors actually made by students. In the analysis of the individual items, 71% of the examinees answered the easiest item correctly in the constructed-response format, while 92% got it correct in the multiple choice format. According to Bridgeman (1992), this is caused not only by the opportunity to guess, but also by the implicit corrective feedback that is part of the multiple choice format. In other words, if the answer computed by the examinee is not among the answer choices in a multiple choice format, the examinee knows that an error was made and may try a different strategy to compute the correct answer. Such feedback may reduce trivial computational errors. However, despite the impact of format differences at the item level, total test scores in the constructed response and provided response formats appeared to be comparable. Both formats ranked the relative abilities of students in the same order, gender and ethnic differences were neither lessened nor exaggerated and correlations with other test scores and college grades were about the same. Bridgeman (1992) reminds us that tests do more than assign numbers to people. They also help to determine what students and teachers perceive as important:

> Test preparation for an examination with an open-ended answer format would have to emphasize techniques for computing the correct answer, not methods for selecting among five answer choices. Thus, with the grid-in format, coaching and test preparation should become synonymous with sound instructional strategies that are designed to foster understanding of basic mathematical concepts. Ultimately, the decision to accept or reject open-ended answer formats may rest as much on these non-psychometric considerations as on any small differences in test reliability or validity (Bridgeman, 1992, p271).

Assessment for broader educational and societal uses calls for tests that are comprehensive in breadth and depth. Both breadth and depth can be covered by including a large number of questions for assessment using a variety of question formats, such as CRQs and PRQs, including the multiple choice format. Both open-ended and fixed-response assessment formats have a place

to ensure that assessment remains open and congenial to all students (Engelbrecht & Harding, 2004).

2.10 MULTIPLE CHOICE QUESTIONS

The multiple choice test, first invented in 1915, was derived from the tradition of intelligence testing. Intelligence tests, which were to influence the construction of numerous subsequent tests, put mental ability on a scale from low to high. Tasks were arranged in increasing order of difficulty, and the examinee received a score based on the point at which successful performance began to be outweighed by unsuccessful performance. Intelligence tests were instituted in many societies to meet the need for selection into specialist or privileged occupations. One of the first uses of multiple choice testing was to assess the capabilities of World War I military recruits. Criticisms of multiple choice testing became prominent in the late 1960s, notably with the publication by Hoffman (1962) of *The Tyranny of Testing*. The strongest criticisms arose from the growing body of research into effective learning (Gifford & O'Connor, 1992). Here, the evidence indicated that learning is a complex process which cannot be reduced to a routine of selection of small components (Black, 1998). The multiple choice test was further justified by the prevailing emphasis on managing learning through specification of behavioural objectives. These objective tests provided an economical and defensible way of meeting the social needs of an expanding society (Black, 1998). The importance and nature of the function of objective testing changed as societies evolved, from serving education for a small elite, through working with the larger numbers and wider aspirations of a middle class, to dealing with the needs and problems of education for all.

Multiple choice questions (MCQs) have been the most developed of all objective tests. They are applicable to a wide range of disciplines. There is a long history of their use in medicine (Freeman & Byrne, 1976). In undergraduate education, they are generally used within formal examination settings in which a large number of questions are used. They also tend to be used in classes where

enrolment numbers are large. MCQs are attractive to those looking for a faster way of assessing students arising from their ease of marking (Hibberd, 1996). MCQs are easy to mark by hand or by computer, either through optically marked response sheets, directly online or a template. This means that rapid feedback can be given to students, and it also gives the lecturers better records of what students do and do not know which makes it easier to identify major areas of attention.

Many variations of multiple choice form have been used. Wesman (1971) defines the following eight types: the correct answer variety, the best answer variety, the multiple response variety, the incomplete statement variety, the negative variety, the substitution variety, the incomplete alternatives variety and the combine response variety. Extended matching items/questions are also types of multiple choice questions, with the main difference being that there are two or more scenarios. The principle of this type of MCQ is that each scenario should be roughly similar in structure and content, and each scenario has one 'best' answer from amongst the series of answer options given. This variation of MCQ is often used in medical education and other healthcare subject areas to test diagnostic reasoning. Research has shown that students exposed to this variation of MCQ format have a greater chance of answering incorrectly if they cannot synthesise and apply their knowledge (Case & Swanson, 1989).

MCQs are useful for both summative and formative purposes. Use of MCQs as part of an assessment portfolio is extremely valuable and is particularly useful for initial diagnostic purposes. Its strength as a diagnostic test lies in its capacity to detect at a very early stage, any significant gaps in knowledge of an individual student (Hibberd, 1996). The printed or displayed individual results can be given to each student together with directions to relevant supplementary material. The global results from the tests can inform and assist in directing tutorial assistance or other help. Also, they may be used to assist in future planning of lectures, seminars and classes or in more general use for revision purposes. Their use in teaching improves *test-wiseness* (Brown, Bull & Pendlebury, 1997), as well as learning and thereby increases the reliability of

57

the assessment procedure. Sometimes increasing test-wiseness is thought to be questionable, yet if one is going to assess learning in a particular way, then one should give students the opportunities to learn and to be assessed in that way. Ebel and Frisbie (1986) justified test-wiseness by stating that more errors are likely to originate from students who have too little rather than too much skill in test taking. Brown, Bull and Pendlebury (1997) indicate that the use of MCQs in improving test-wiseness can also develop the self-confidence of the students being assessed.

MCQs provide an important way of evaluating the mathematical ability of a large class of students, but they need more care in setting than the more conventional CRQs requiring full written solutions (Webb, 1989). There are several well documented rules to guide the construction of such questions (Gronlund, 1988; Nightingale *et al.*, 1996; Webb, 1989). Carefully constructed MCQs can assess a wide variety of skills and abilities, including higher-order thinking skills. MCQs involve the following terminology:

Item:	the term for the whole MCQ, including all answer choices.
Stimulus material:	the text, diagram, table, graph etc. on which the item is based.
Stem:	either a question or an incomplete statement presenting the problem for which response is required.
Options or alternatives:	all the choices in an item.
Key:	the correct answer or best option.
Distracters:	the incorrect answers or options other than correct answers.
Item set:	a number of items all of which are based around the same stimulus material.

(Adapted from Hughes & Magin, 1996, p152)

Sample Item

If u and v are orthogonal (i.e. perpendicular), then $\| u - v \|^2 =$ Stem

A. $(\| u \| + \| v \|)^2$

B. $(\| u \| - \| v \|)^2$ Distracters   Options   Item

C. $\| u \|^2 - \| v \|^2$

D. $\| u \|^2 + \| v \|^2$ Key

(MATH 109 Tutorial Test 3, August 2004, University of the Witwatersrand.)

Creating a good MCQ starts with a description of the skills, abilities and knowledge to be tested in the form of written specifications. Once the test specifications are prepared, test questions that assess the skills, abilities and/or knowledge must be constructed.

Advice on setting MCQs:

- The item as a whole should test one or more important learning outcomes, processes or skills. The commonest faults found in MCQ items are *irrelevance* and *triviality* (McIntosh, 1974). McIntosh suggests that both of these faults can be avoided only through a process of ensuring that all questions are related to previously established learning outcomes and that the answering of each question requires application of knowledge, understanding or other abilities which have been identified as important course outcomes.

- The stem should be stated in a positive form, wherever possible. Diagrams and pictures can be an economical way of setting out the question situation. A complex or lengthy stem can be justified if it can serve as the basis for several questions.

- The options should all be similar to one another in numbers of words and style, both for directness and to avoid giving clues, whether genuine or false.

- Questions should be checked by several experts to ensure that there are no circumstances or legitimate reasoning by virtue of which any of the distracters could be correct; to look for unintended clues to the correct option; and to ensure that the key really is correct. The main challenge in setting good MCQs is to ensure that the distracters are plausible so that they can represent a significant challenge to the student's knowledge and understanding (Kehoe, 1995).

- Hughes and Magin (1996), advocate using simple words and clear concepts in order to avoid making mathematics tests highly dependent upon students' ability to read.

2.10.1 Advantages of MCQs

MCQs, although often criticised, still form the backbone of most standardised and classroom tests (Fuhrman, 1996). There is a large literature in the field of psychometrics, the psychological theory of mental measurement, that confirms there are good reasons for using multiple choice testing (Haladyna, 1999).

The major justifications offered for their widespread use include the following (Tamir, 1990):
- they permit coverage of a wide range of topics in a relatively short time
- they can be used to measure different levels of learning
- they are objective in terms of scoring and therefore more reliable
- they are easily and quickly scored and lend themselves to machine scoring
- they avoid unjustified penalties to students who know their subject matter but are poor writers

- they are suitable for item analysis by which various attributes can be determined such as which items on a test were too easy or too difficult or ambiguous (Isaacs, 1994; Wesman, 1971).

It is a common misconception that MCQs can test only factual recall. They can be used to test many types of learning from simple recall to high-level skills like making inferences, applying knowledge and evaluating (Adkins, 1974; Aiken, 1987; Haladyna, 1999; Isaacs, 1994; Oosterhof, 1994; Thorndike, 1997; Williams, 2006). These testing experts point out that while multiple choice tests are quick and easy to score, good multiple choice items which test high-level skills are more difficult and time consuming to develop. The design of MCQs is challenging if one wishes to assess deep learning. It is possible to test higher-order thinking through well-developed and researched MCQs, but this requires skill and time on the part of those designing the test.

MCQs can provide a good sampling of the subject matter of concern, and therefore, an adequate and dependable sample of student responses. Given the same time for assessment, free-response items usually sample a smaller number of topics and therefore, tend not to be as reliable as tests made up of many short questions (Fuhrman, 1996). Reliable multiple choice assessments can be ideal if comprehension, application and analysis of content is what one wants to test (Johnson, 1989). Johnson (1989) suggests two ways that higher level MCQs can be introduced into the assessment programme for a curriculum. One way is to make sure that the curriculum includes problem solving skills such as interpreting data, making predictions, assessing information, performing logical analyses, using scientific reasoning or drawing conclusions, and to include questions of this nature in tests.   Another way is to combine mathematics content with process. In order to do this, you need to examine concepts currently tested in the curriculum and think of ways to restructure items so that they require students to apply concepts, analyse information, make inferences, determine cause and effect or perform other thoughtful processes.

By writing questions that assess your students' higher levels of ability, you are really testing their unlimited potential (Johnson, 1989). Johnson (1989) cautions that classroom tests should also include some items written at the knowledge and comprehension levels, since students need to have a certain base of facts and information 'before they are able to reach other plateaus of applying skills and analyzing and evaluating data' (p61).

According to Elton (1987), the reason why MCQs demand so much more than just memory is quite different. It has to do with the brevity of the question and not with the fact that a correct answer has to be chosen. Brief questions can be set in such a way that the student can be asked to think for about two minutes. If he/she thinks wrongly, nothing much is lost, as he/she can go on to the next question. However, if one expects the student to think constructively for 25 minutes or an hour and if he/she then goes wrong in the first five minutes, the penalty is much greater.

MCQs give the instructor the ability to obtain a wide range of scores for better discrimination among students. If fine discrimination among students is desired, MCQs offer the ability to obtain a wide range of scores, because the test is made up of many separately scored parts (Fuhrman, 1996).

With multiple choice tests, it is easier to frame questions so that all students will address the same content. The student must deal with the responses made available. Although this does increase the risk of the student answering correctly by merely recognising or even guessing the correct answer, at least objective scoring is made easier (Hibberd, 1996). CRQs provide less structure for the student, and a common problem is that *test-wise* students can overwhelm the marker with pages of unrelated discourse that may at first glance appear to signify understanding (Fuhrman, 1996).

A further advantage of MCQs, in particular for large groups of students, is that of the reduction in cost and time. The cost savings is most significant in mass testing such as for large lecture courses or standardised testing. MCQs are

quick to mark and provide for ready analyses and comparisons between groups (Hibberd, 1996). High quality MCQs are not easy to construct, but the time spent in constructing them can be offset against the time saved in marking. If one has a large number of students (and not enough tutors) to frequently and objectively assess using CRQs, MCQs can be appropriate for some assessments, especially if subject-matter knowledge is emphasised in the course. Since MCQs can be machine scored, they can be used to assess when scoring must be done quickly, thus being both cost and time effective.

In addition to being a legitimate testing mode, the problem oriented multiple choice examination has pragmatic advantages. First, it makes cheating by copying more difficult. With the multiple choice format it is easy to create duplicate exams with answers, and questions renumbered, making copying very difficult. Secondly, all scoring can be done by machine, eliminating unfair subjective evaluations.

2.10.2 Disadvantages of MCQs

Graham Gibbs (1992) claims that one of the main disadvantages of MCQs is that they do not measure the depth of student thinking. They are 'often used to test superficial learning outcomes involving factual knowledge, and that they do not provide students with feedback' (p31). Further, he argues that this disadvantage is not inherent in the tests in that 'it is possible to devise objective tests which involve analysis, computation, interpretation and understanding and yet which are still easily marked' (p31). A common concern expressed when using MCQs is that students are encouraged to adopt a surface learning approach, rather than developing a deep approach to learning the topic (Black, 1998; Resnick & Resnick, 1992).

Bloom (1956) himself wrote such tests 'might lead to fragmentation and atomisation of educational purposes such that the parts and pieces finally placed into the classification might be very different from the more complete objective with which one started' (p5).

Many educators believe that the use of objective tests such as MCQs, while providing inexpensive assessment of large groups of students, may be a factor in lowering achievement in mathematics. The California Mathematics Council's (CMC) analysis of publishers' tests, for example, indicated that this assessment mode did not provide information about student understanding of graphs, probability, functions, geometric concepts or logic, focusing instead on rote computation (CMC and EQUALS, 1989). In another study, Berg and Smith (1994) challenge the validity of using multiple choice instruments to assess graphing abilities. They argue that from the viewpoint of a constructivist paradigm, multiple choice instruments are an invalid measure of what subjects can actually do, and equally important, the reasons for doing so. However, as shown by many authors (Gronlund, 1988; Johnson, 1989; Tamir, 1990), as the focus turns away from the *correct answer* variety (where one of the options is absolutely correct while the others are incorrect) to the *best answer* variety (where the options may be appropriate or inappropriate in varying degrees and the examinee has to select the *best*, namely the most appropriate option), the picture changes dramatically. Now the student is faced with the task of carefully analysing the various options, each of which may present factually correct information, and of selecting the answer which best fits the context and the data given in the item's stem. MCQs of this kind cater for a wide range of cognitive abilities. When compared with open-ended CRQs, although they do not require the student to formulate an answer, they do impose the additional requirement of weighing the evidence, provided by the different options. The correct answers require analytical skills, knowledge of relevant theories and judgement, all cognitively high level items within the assessment models.

A criticism, mentioned earlier, is that MCQs are very time consuming to write. Andresen, Nightingale, Boud & Magin (1993) estimated that the development time is such that it would take three years before a course with 50 students a year was showing a saving in staff time. If reliability is at a premium, then many rewrites and plentiful piloting are needed. A department will want to build up a substantial bank of MCQs so that a cohort of students gets a different item on a

topic than did the students in the past two years. One suggestion to build up a bank of MCQs is to use them for formative purposes, in peer- and self-assessment, perhaps with computer or tutor support.   Such a study was conducted by Barak and Rafaeli (2004) in which graduate MBA students were required to author questions and present possible answers relating to topics taught in class. The students were required to share these questions online with their classmates. The online question-posing assignment required students to be actively engaged in constructing instructional questions, testing themselves with their fellow students' questions (self-assessment) and assessing questions contributed by their peers (peer-assessment).   Although standardised item banks of mathematics questions at the tertiary level are freely available, these are problematic in that they are standardised to specific contexts and may contain linguistic features and other concepts which are unfamiliar to students attending universities in South Africa.   If used, such questions will have to be modified and refined to suit the South African context.

Another objection to the whole principle of multiple choice is that MCQs are not characteristic of the real world (Bork, 1984). Education often criticise multiple choice tests because such tests are rarely 'authentic' (Fuhrman, 1996). Webb (1989) relates a comment made by Peter Hilton on this very issue about MCQs:

> …the very idea is highly artificial. Nowhere in real-life mathematics, let alone real life, is one ever faced with a problem together with five possible solutions, exactly one of which is guaranteed to be correct (p216).

Fuhrman (1996) argues that when a real world task is one that requires choosing the 'correct' or 'best' answer from a limited universe of answers, multiple choice tests can be used. But if the real world task is one that requires the performance of a skill, such as a laboratory skill or writing skill, MCQs are not usually appropriate.

Webb's defence in this case is that even so MCQs serve as a diagnostic tool and not a real-life event. The distracters in a multiple choice item function much like one of the standard procedures in a Piagetian classical interview. There,

when the interviewer is not fully satisfied even when the child gives a correct answer, understanding is checked by suggesting an alternative answer. Thus, the distracters in a good multiple choice item serve as such alternatives.

In designing MCQs, a recognised strategy is to select plausible distracters. If these are chosen on the basis of representing common errors in understanding the topic, patterns of wrong choices can have useful diagnostic value. Most test setters use their experience of frequently encountered misconceptions when deciding on plausible distracters.

The danger of this practice, however, is that when a student gets to an answer on grounds of a misconception and finds his wrong answer as one of the distracters, the student believes that he answered correctly. The student often feels that his mathematical prowess is intact until he receives feedback on his response, thereby reinforcing the misconception (Engelbrecht & Harding, 2003). This view is supported by Webb (1989) who proposes that distracters should be devised that

> …look feasible, but which could not have been obtained by means of a correct strategy incorporating a minor algebraic error (p217).

When distracters based on misconceptions are included, immediate feedback is advisable if MCQs are used in formative assessment. The MCQs must be written in a manner that does not give away the correct answers. The MCQ test must also feature a good overall balance of well written items clearly correlated to the learning outcomes of the course (Johnson, 1989).

The rigidity of the marking scheme for MCQs is criticised. Several authors have reported that about one third of students choosing the correct option in a multiple choice question do so for a wrong reason (Tamir, 1990; Treagust, 1988; Johnstone & Ambusaidi, 2001). We assume that when a student makes a wrong choice, it indicates a certain lack of knowledge or understanding, or that the student reveals a misconception. However, it is possible for students to have the correct understanding, but to make a minor calculation error.

In general, several options are available for the modification of test items in order to address these issues (Johnstone & Ambusaidi, 2001). Treagust (1988) developed a two-tier testing methodology for the probing of conceptual understanding. MCQs treat minor and major errors as equal and do not make provision for partial credit. There have been several ingenious attempts made to score MCQs to allow for partial knowledge (Friel & Johnstone, 1978; Johnstone & Ambusaidi, 2001). Some of these ask the students to rank all the responses in the question from the best to the worst. In other cases students are given a tick (✓) and two crosses (✗) and asked to use the crosses to label distracters they know to be wrong and the tick to choose what they think is the best answer. They get credit for eliminating the wrong, as well as for choosing the correct. The rank order produced when these devices are applied to multiple choice tests and the rank order produced by an open-ended test correlate to give a value of about 0.9; almost a perfect match. This underlines the importance of the examiner having the means of detecting and rewarding reasoning (Johnstone & Ambusaidi, 2001). You could also give partial credit for a partially correct option on Learning Management Systems such as Blackboard (Engelbrecht & Harding, 2006).

2.10.3 Guessing

Another (well researched) concern when using MCQs is the possibility of *guessing*. It is always possible to guess at an answer so that the probability of obtaining correct answers in items comprising of four options by purely random selection is 25%. The probability of choosing the correct answer randomly gets lower if there are a sufficient number of distracters. True/false questions are rarely a good idea.

Different evaluators have taken different positions regarding the way the problem of guessing should be addressed. Guessing can be counteracted by negative marking or penalty marking whereby each wrong answer leads to marks being lost. A rational student who is not sure of the answer to a question

will therefore not answer it, incurring no penalty. A wrong answer penalty would strongly discourage guessing. Aubrecht and Aubrecht (1983) argue that although they would like to discourage *random* guessing, they believe that there is an important pedagogical reason to encourage *reasoned* guessing. Active involvement on the part of the student in sifting through the answers on the test, even if the wrong answer is eventually chosen, prepares the student to understand the correct answer when it is explained. If students can correctly eliminate some distracters, this method of reasoned guessing, they will do better than if they guess randomly. A wrong answer penalty in MCQs reduces the effect of guessing (Harper, 2003) and finds indications of reasoning paths for students (Johnstone & Ambusaidi, 2001).

At some institutions, however, negative marking is prohibited. Using negative marking also requires knowledge of the probability for guessing the correct answer. This may be beyond the statistical competence of many question designers, particularly if the test includes multiple response questions or matching questions for which the process is more complex. Harper (2003) developed a method for post-test correction for guessing. His method enables the test designer to do a post-test correction to neutralise the impact of guessing.

An alternative approach to eliminate guessing is the use of *justifications* (Tamir, 1990). The term *justification* is assigned to reasons and arguments given by a respondent to a multiple choice item for the choice made. When students are required to justify their choice in MCQs, they have to consider the data in all the options and explain why a certain option is better than others. In addition, there is the *back-wash* effect when requiring justifications for multiple choice items. In other words, students who know that they may be asked to justify their choices will attempt to learn their subject matter in a more meaningful way and in more depth so that they will be prepared to write an adequate and complete justification. Justifications to choices in multiple choice items significantly increase the information that test results provide about students' knowledge.

Their contribution is made by:

- identifying misconceptions, missing links and inadequate reasoning among students who correctly choose the best answer
- gaining better understanding of notions held by students who choose certain distracters.

2.10.4 In defense of multiple choice

Seen as a part of an overall strategy of assessment, MCQs have a great deal to commend them. Much of the criticism levelled at multiple choice tests focuses on poorly worded answers which penalise the better student and that the correct answer may be guessed. Neither of these faults is inherent in the multiple choice test itself, but only in the way in which it is used. The primary focus of a mathematics testing methodology based on an active, constructivist view of learning is on revealing how individual students think about key concepts in mathematics. Rather than comparing students' responses with a correct answer to a question, the emphasis should rather be on understanding the variety of responses that students make to a question and inferring from those responses students' level of conceptual understanding. In defense of multiple choice tests, they provide faster ways of assessing the large numbers of first year undergraduate students studying tertiary mathematics and test scores can be highly reliable. This research study has concentrated mostly on MCQs, and not on the other types of PRQs. As discussed in the literature review, MCQs enable one to sample rapidly a student's knowledge of mathematics and they may be used to measure deep understanding. Literature search has revealed that alternative types of MCQs encourage a deep approach to learning as they require students to solve a problem by utilising their knowledge and intellectual skills. Traditional *factual recall* MCQs can be modified to both assist student learning and to better assess the students' progress towards understanding.

A sophistication of the standard multiple choice test is available through the use of computer adaptive testing. Here, the questions to be presented to a student at any point during a test can be chosen on the basis of the quality of the

answers supplied up to that point. This can mean that each student can avoid spending time on items which give little useful information because they are far too difficult or far too easy (Scouller & Prosser, 1994).

Biggs (1991) points out that the use of MCQs in very large classes provides a form of continuous assessment and feedback:

> students knowing how they have done on a multiple choice test can provide more feedback than is otherwise available…and that it is also possible to provide computerised tutorial feedback for students when they give incorrect answers to multiple choice questions (p31).

The inclusion of multiple choice formats in assessment lessens the burden of heavy teaching loads coupled with large student numbers experienced by academic staff, particularly in the early undergraduate years. This enables academic staff to perform their duties as teachers and researchers in academic institutions.

The challenge, then, is to find out enough about student understanding in mathematics to design assessment techniques that can accurately reflect these different understandings.

2.11 GOOD MATHEMATICS ASSESSMENT

From a methodological point of view, mathematics assessment for broader education and societal uses calls for tests that are comprehensive in breadth and depth (Ramsden, 1992). With regard to the importance of assessment, Ramsden (1992) says that:

> From our students' point of view, assessment always defines the actual curriculum. In the last analysis, that is where the curriculum resides for them, not in the lists of topics or objectives. Assessment sends messages about the standard and amount of work required, and what aspects of the syllabus are most important. Too much assessed work leads to superficial approaches;

clear indications of priorities in what has to be learned, and why it has to be learned, provide fertile ground for deep approaches (p187).

Whether we focus on examinations or on other forms of assessment, we can use a range of techniques to assess the nature and extent of student learning. Our decisions about which forms of assessment we choose are likely to be affected by the particular learning context and by the type of learning outcome we wish to achieve (Wood, Smith, Petocz & Reid, 2002).

Essentially, good mathematics assessment practices:

- *encourage meaningful learning* when tasks encourage understanding, integration and application
- *are valid* when tasks and criteria are clearly related to the learning objectives and when marks or grades genuinely reflect students' levels of achievement
- *are reliable* when markers have a shared understanding of what the criteria are and what they mean
- *are fair* if students know when and how they are going to be assessed, what is important and what standards are expected
- *are equitable* when they ensure that students are assessed on their learning in relation to the objectives
- *inform teachers about their students' learning* (Biggs, 2000; Brown & Knight, 1994; Wood *et al.*, 2002).

It is also possible (and desirable) to characterise the quality of a test as a whole. In this context, *quality* is defined as the extent to which the test measures what we wish it to measure, and the degree to which it is consistent as an instrument for this measurement (Niss, 1993). The first of these characterises the *validity* of the test: the second of these is the *reliability*. Measuring quality in terms of reliability and validity can and should be done for any type of assessment. *Good assessment* must be both reliable and valid (Fuhrman, 1996). This definition is part of the "common wisdom" of psychometrics (Haladyna, 1999). A reliable assessment is one which consistently achieves the same results with the same

(or similar) cohort of students. Qualitatively, a reliable measure is one that provides consistent scores. There are several ways to determine the reliability of a measure. One type of reliability is defined as the level of agreement between test scores for a test given on several occasions. Reliability can be expressed analytically, and using performance data, calculated for any scored test. Various factors affect reliability: the number and quality of the questions, including ambiguous questions, too many options within a question paper, the type of examination environment, the type of test administration directions, vague marking instructions, the objectivity of scoring procedures, poorly trained markers and the test-security arrangements (Nightingale *et al.*, 1996).

An assessment is valid when it accurately measures what it intends to measure. Validity is determined in a variety of ways, depending on the purpose of the test. For example, for a test that is intended to assess subject matter, the validity of the test content can be confirmed by linking the items to the important concepts in the curriculum. A valid test is built by ensuring that each question is linked to a specific item that is included in the curriculum. Often the description of the skills/knowledge to be tested is too broad to permit the measurement of each and every concept listed. In this case, a valid test should sample the subject matter in a way that ensures the broadest possible representation of the subject in the examination. For a test used for predictive purposes, for example to predict success in an academic programme, the validity can be confirmed by correlating performance on the test to some measure of actual success attained (Black, 1998).

A student's mathematical understanding, for example, of linear functions or the capacity to solve non-routine examples, is a "mental concept" (Romagnano, 2001), and as such can only be observed indirectly. Objectivity in mathematics assessment would be desirable if we could have it, but according to Kerr (1991), is a myth. Romagnano (2001) is of the opinion that all assessments of students' mathematical understanding are subjective. Good mathematics assessment should not be defined in terms of its objectivity or subjectivity. A more useful way to characterise good mathematics assessment methods would be with

respect to their *consistency* (or reliability) and the *meaning* (or validity) of the information they provide. When a consistent method is used by different teachers to assess the knowledge of a given student, the teachers' assessments will agree. When two students have roughly the same level of understanding of a set of mathematical ideas, consistent assessment of these students' understandings will be roughly equal as well. Good mathematics assessment methods provide teachers with information about student understanding of specific mathematical ideas and how this understanding changes over time, information that can be used to make appropriate curriculum decisions.

> The Assessment Principle: Assessment should support the learning of important mathematics and furnish useful information to both teachers and students.
>
> *-Principles and standards for school mathematics (NCTM, 2000)*

The National Council of Teachers of Mathematics (NCTM, 2000) evaluation standards suggest that:

- student assessment be integral to instruction
- multiple means of assessment be used
- all aspects of mathematical knowledge and its connections be assessed
- instruction and curriculum be considered equally in judging the quality of a programme.

According to Webb and Romberg (1992), good mathematics assessment practices are those in which students can:

- learn to value mathematics
- develop confidence
- communicate mathematically
- learn to reason mathematically
- become mathematical problem solvers (p39).

Assessment should be a means of fostering growth toward high expectations and should support high levels of student learning. When assessments are used in thoughtful and meaningful ways, students' scores provide important information that, when combined with information from other sources, can lead

to decisions that promote student learning and equality of opportunity (NCTM, 2000).

2.12 GOOD MATHEMATICS QUESTIONS

The types of questions that we set reflect what we, as mathematics educators, value and how we expect our students to direct their time (Wiggins, 1989). In striving to set questions of good quality, assessors need to be able to measure how good a mathematics question is. *Good* mathematics questions are those that help to build concepts, alert students to misconceptions and introduce applications and theoretical questions.

> When students are asked to puzzle and explain, to apply their knowledge in an unfamiliar context, they must construct meaning for themselves by relating what they know to the problem at hand. In other words, they must act like mathematicians. This kind of activity encourages them in the belief that mathematics is primarily a reasonable enterprise, founded in the relationships apparent in everyday life and accessible to all students, whatever age or level of ability (Massachusetts Department of Education, 1987, p41).

According to Romberg (1992) the criteria for measuring *good mathematics questions* can be traced to three main concerns:

1. Test questions must reflect the current view of the nature of mathematics. This view emphasises understanding, thinking, and problem solving that require students to see mathematical connections in a situation-based problem and to be able to monitor their own thinking processes to accomplish the task efficiently. This requires that test questions have the following characteristics:

- They assess thinking, understanding and problem solving in a situational setting as opposed to algorithmic manipulation and recall of facts.

- They assess the interconnection among mathematical concepts and the outside world.

2. Test questions must reflect the current understanding of how students learn. The current view of instruction and learning assumes that students

are active learners and engage in creating their own meaning during the instructional process. This requires that test questions have the following characteristics:

They must:

- be engaging
- be situational and based upon real-life applications
- have multiple-entry points in the sense that students at various levels in their mathematical sophistication should be able to answer the question
- allow students to explore difficult problems and students' explorations are rewarded
- allow students to answer correctly in diverse ways according to their experiences, rather than requiring a single answer
3. Test questions must support good classroom instruction and not lend themselves to distortion of curriculum. Good curriculum practices require that test questions have the following characteristics
- They must be exemplars of good instructional practices
- They should be able to reveal what students know and how they can be helped to learn more mathematics (p125).

Hubbard (2001) suggests that good mathematics questions are those that require students to reflect on results, in addition to obtaining them. Good questions specifically encourage students to develop relational understanding, a process approach and higher-level learning skills. Further, students' solutions to good questions should indicate what kind of intellectual activity they engaged in to answer the questions. Good questions direct students to think, as well as to do (Hubbard, 2001).

Asking the right question is an art to be cultivated both by educators and by students, for teaching and learning as well as for assessment. Good questions and their responses will contribute to a climate of thoughtful reflectiveness (Niss, 1993). Stenmark (1991) has suggested a list of possible characteristics of good open-ended questions to open new avenues of thinking for students.

- *Problem Comprehension*

Can students understand, define, formulate or explain the problem or task? Can they cope with poorly defined problems?

- *Approaches and Strategies*

Do students have an organised approach to the problem or task? How do they record? Do they use tools (diagrams, graphs, calculators, computers, etc.) appropriately?

- *Relationships*

Do students see relationships and recognise the central idea? Do they relate the problem to similar problems previously done?

- *Flexibility*

Can students vary the approach if one approach is not working? Do they persist? Do they try something else?

- *Communication*

Can students describe or depict the strategies they are using? Do they articulate their thought processes? Can they display or demonstrate the problem situation?

- *Curiosity and Hypotheses*

Do students show evidence of conjecturing, thinking ahead, checking back?

- *Self-assessment*

Do students evaluate their own processing, actions and progress?

- *Equality and Equity*

Do all students participate to the same degree? Is the quality of participation opportunities the same?

- *Solutions*

Do students reach a result? Do they consider other possibilities?

- *Examining results*

Can students generalise, prove their answers? Do they connect the ideas to other similar problems or to the real world?

- *Mathematical learning*

Did students use or learn some mathematics from the activity? Are there indications of a comprehensive curriculum? (p31).

Questions might also assess a student's understanding of a specific mathematical topic. Such focused mathematics questions can be developed according to instructional needs.

Retaining unsatisfactory questions is contrary to the goal of good mathematics assessment (Kerr, 1991). This view is consistent with the NCTM Evaluation Standards proposal that 'student assessment be integral to instruction' (NCTM, 1989, p190). By thinking of instruction and assessment as simultaneous acts, educators optimise both the quantity and the quality of their assessment and their instruction and thereby optimise the learning of their students (Webb & Romberg, 1992).

2.13 CONFIDENCE

When the National Council of Teachers of Mathematics (NCTM) published its *Curriculum and evaluation standards for school mathematics* in 1989, many of the recommended assessment methods were different from those routinely used in mathematics classrooms of the 1980s. For example, one such recommended assessment method was having students write essays about their understanding of mathematical ideas and using classroom observations and individual student interviews as methods of assessment. The document, *Evaluation Standard* 10 – *Mathematical Disposition* (NCTM, 1989), maintains

that it is also important to assess students' *confidence*, interest, curiosity and inventiveness in working with mathematical ideas. Corcoran and Gibb (1961) and other writers in the 1950s and the 1960s argued similar points (as cited in the National Council of Teachers of Mathematics Yearbook, 1961):

> One of the best indications of the mastery of a subject possessed by a pupil is his ability to make significant comments or to ask intelligent questions about the subject… Another indication of achievement in a field is interest in that field… Still another indication of achievement is the degree of confidence displayed when work is assigned or undertaken (Spitzer, pp193-194).

> Appraisal ideally includes many aspects of learning in addition to acquisition of facts and skills. It includes the student's attitude toward the work; the nature of his curiosity about the ingenuity with mathematics; his work habits and his methods of recording steps toward a conclusion; his ability to think, to exclude extraneous data, and to formulate a tentative procedure; his techniques and operations; and finally, his feeling of security with his answer or conclusion (Sueltz, pp15-16).

Using only the results of multiple-choice tests can lead to incorrect conclusions about what a student does or does not know (Webb, 1989). As Johnson (1989) indicated, if students can write clearly about mathematical concepts, then they demonstrate that they understand them. In a study conducted by Gay and Thomas (1993), with 199 seventh- and eighth-grade students that focused on students' understanding of percentage, about one-fourth of the students had no explanation to support their correct choice to the multiple choice question. It is possible that this lack of response gives some indication of the number of students who simply guessed correctly. It is also possible that these students lacked confidence in their reasoning and chose not to give any explanation (Gay & Thomas, 1993). Students need to have a reason for making decisions and solving problems in mathematics and the confidence to share that reasoning with others (Webb, 1994).

It is well documented that mathematical attitude is one of the strongest predictors of success in the mathematical sciences (McFate & Olmsted, 1999; Wagner, Sasser & DiBiase, 2002). There are, however, a number of non-cognitive factors such as study habits (consistent work), motivation (interest and desire to understand presented material) and self-confidence that may be equally or more important in the prediction of student success (Angel & LaLonde, 1998).

The extent of students' awareness of their strengths and weaknesses is known to be associated with their success or lack of success in some areas of mathematical performance. For example, in the literature on mathematical problem solving (Campione, Brown & Connell, 1988; Krutetskii, 1976; Schoenfeld, 1987), the successful problem solvers are described as those students who have a collection of powerful strategies available to them and who can reflect on their problem-solving activities effectively and efficiently. In contrast, descriptions of unsuccessful problem solvers tend to portray them as students who have command of fewer strategies and who do not function in a self-reflective or self-evaluative manner (Kenney & Silver, 1993).

Students' ability to monitor their learning is one of the key building blocks in self-regulated learning, which, in turn, is an essential requirement for success at tertiary level (Isaacson & Fujita, 2006). Students who are skilful at academic self-regulation understand their strengths and weaknesses as learners as well as the demands of specific tasks. Students who are expert learners know when they have mastered, or not mastered, the required academic tasks and can adjust their learning accordingly (Isaacson & Fujita, 2006). Such students are said to have high metacognitive ability. The inability to do so is especially harmful in the case of poor performers who become victims of an assessment regime that they do not understand and which they perceive themselves to be unable to control. Isaacson and Fujita (2006) have shown that low achieving students have lower metacognitive knowledge monitoring abilities. They are less able to predict their performance after writing a test, rely more on time spent on studying than on mastery of concepts to decide their confidence for success,

are less likely to adjust their self-efficacy depending on feedback received from taking a test and show the largest discrepancy between their actual performance and their expected performance, satisfaction goals and pride goals. Tobias and Everson (2002) have found that the ability to differentiate between what is known (learned) and unknown (unlearned) is an important ingredient for success in all academic settings.

Metacognition has two components: it refers to knowledge about cognition and regulation of one's own cognitive processes (Baker & Brown, 1984). The ability to know how well one is performing through monitoring and checking of outcomes of learning (self-assessment) is an essential requirement for the planning and control of appropriate behaviour to ensure mastery of subject content. Self-reflection and self-assessment of the confidence of a student in answering a test item, whether PRQ or CRQ, encourages sense making and autonomy.

A number of studies have been reported where metacognitive ability of students was assessed and correlated with test performance by means of confidence judgement indicating the likelihood that the answers provided to each multiple choice question was correct (Carvalho, 2007; Sinkavich, 1995). Carvalho (2007) investigated the effects of test types (free response/short answers and multiple choice tests) on students' performance, confidence judgements and the accuracy of those judgements. The results showed that the difference between performance and judgement accuracy was significantly larger for multiple choice than for short answer tests in undergraduate psychology. Students were significantly more confident in multiple choice than in short-answer tests, but their judgements were significantly more accurate in the short answer than in the multiple choice tests. In addition, upon repeated exposure to a short-answer test format both the performance and confidence of students increased, whereas that was not the case for multiple choice testing. Carvalho suggested a possible explanation for this observation is that multiple choice tests may require tasks of lower cognitive demand, such as recognition, as compared to the higher demand of recall and self-construction of responses. This may tempt students

into reduced metacognitive activity. They do not need to engage as deeply with the content and their mastery of the material in order to make an accurate judgement (Pressley, Ghatala, Woloshyn, & Pirie, 1990). Carvalho (2007) suggested that the continuous pairing of high confidence and low accuracy levels observed for multiple choice assessment could negatively affect students' self-regulation of learning. If they do not understand the reasons why their judgements are consistently inaccurate despite their feeling of confidence, they may start to feel that they have no control over their learning and its relationship to the outcomes of assessment. When students are asked to express their confidence in the correctness of answers provided during assessment they are required to engage in the metacognitive activity of judging their conceptual understanding and/or mastery of skills and proper application to the task at hand.

Assessment in mathematics must build learners' confidence and competence (Anderson, 1995). As we look for increased achievement and motivation in our mathematics classrooms, we must acknowledge and develop self-assessment of confidence as one of the many ways to include authentic assessment as a key element in the learning process. The confidence index (CI), which is an indication of confidence, is discussed in Section 5.2.2.

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.
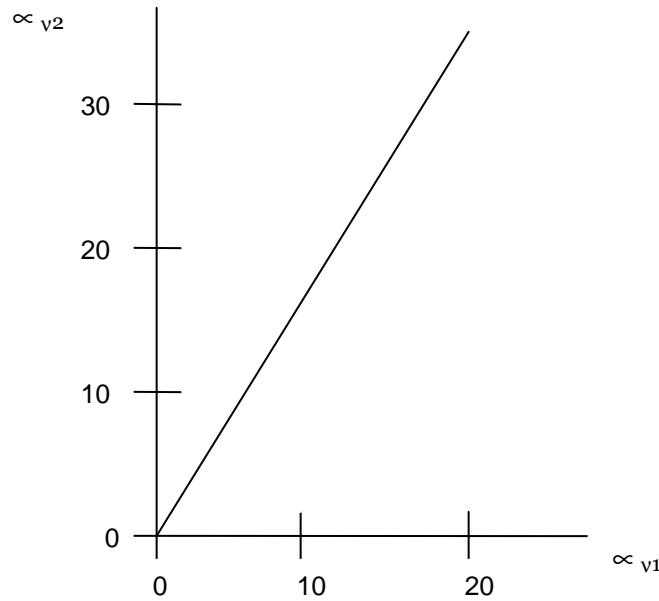
(Source: Rasch ,1980)

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

$$P(\alpha_{vi}) = e^{-\lambda_{vi}} \frac{(\lambda_{vi})^{\alpha_{vi}}}{\alpha_{vi}!}$$
(1.1) ; 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.

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

92

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*

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

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: $P(\alpha_{vi}, \alpha_{vj} | \alpha_{vi} + \alpha_{vj}; \theta_v, \delta_i, \delta_j) = g(\delta_i, \delta_j)$. 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.

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

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.

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

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.

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

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.
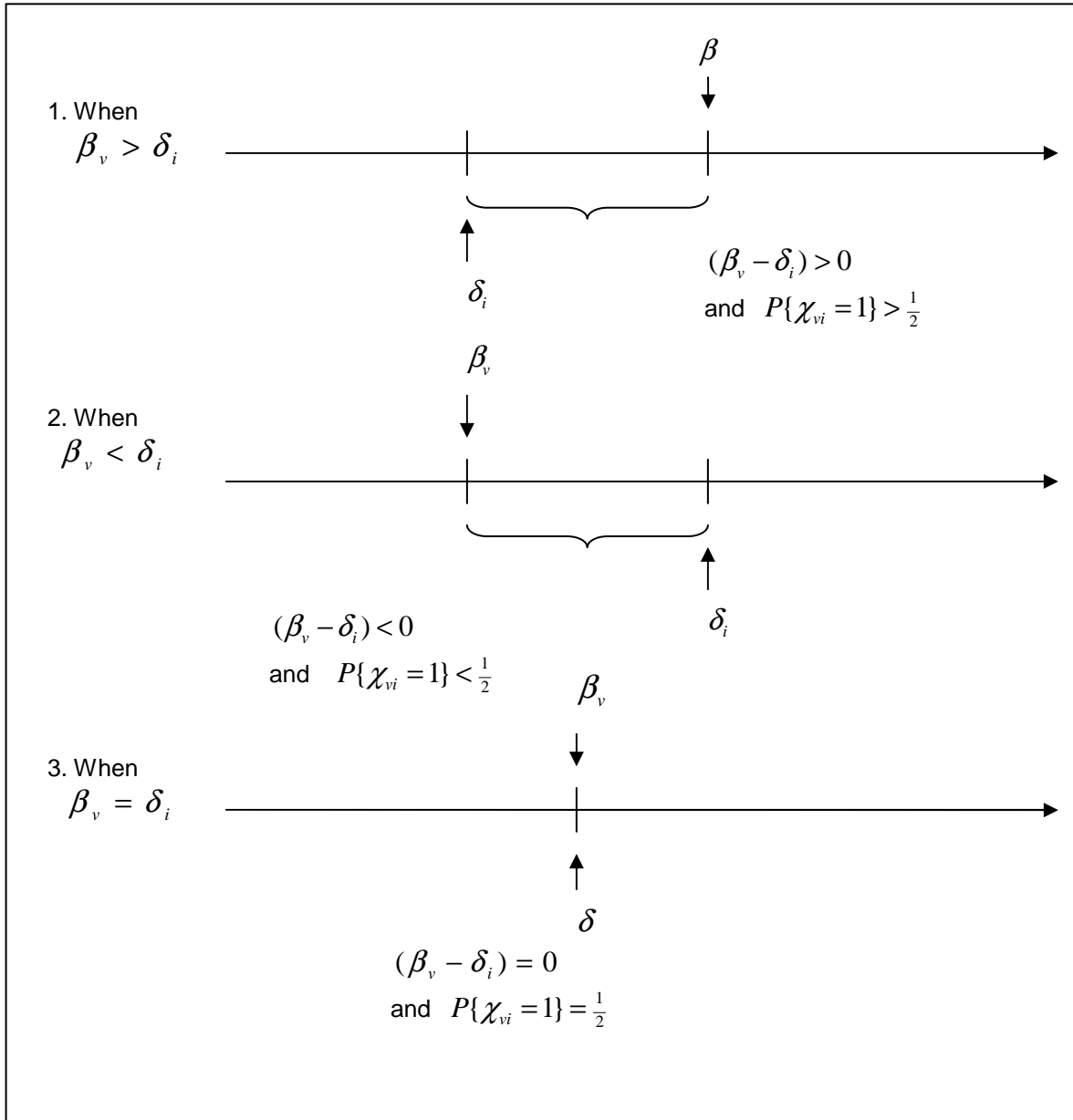
The curve in Figure 3.3 summarises the implications of Figure 3.2 for all reasonable relationships between probabilities of correct responses and differences between person ability and item difficulty. This curve specifies the conditions a response model must fulfill. The difference $(\beta_v - \delta_i)$ 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.
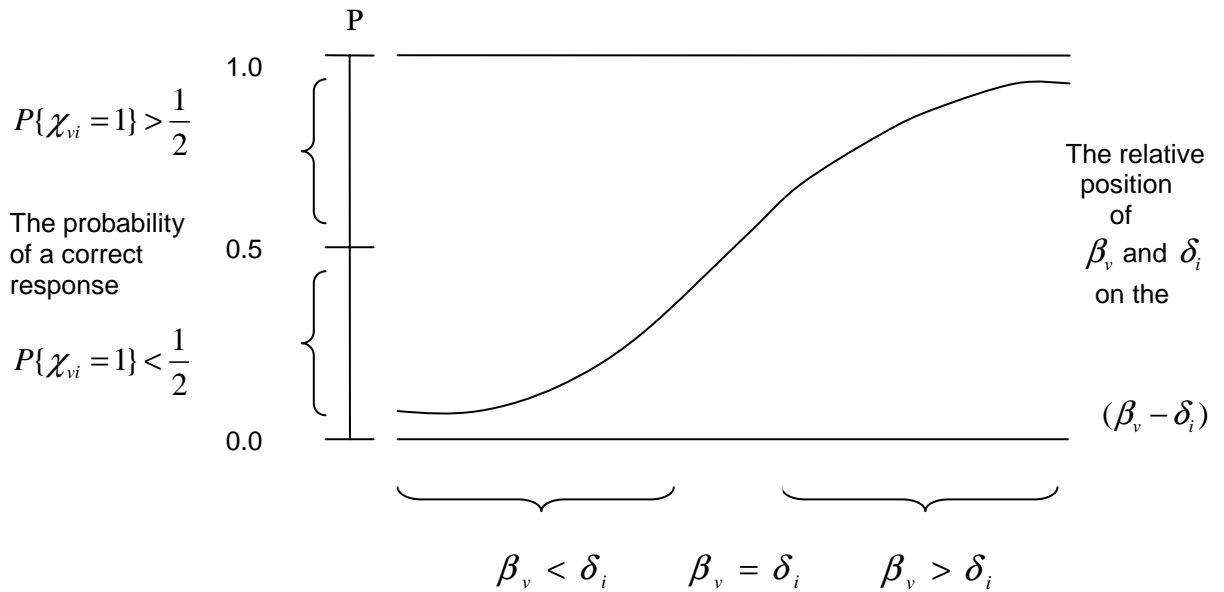
$$P\{\chi_{vi} = 1 | \beta_v, \delta_i\} = f(\beta_v - \delta_i)$$

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 v 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.
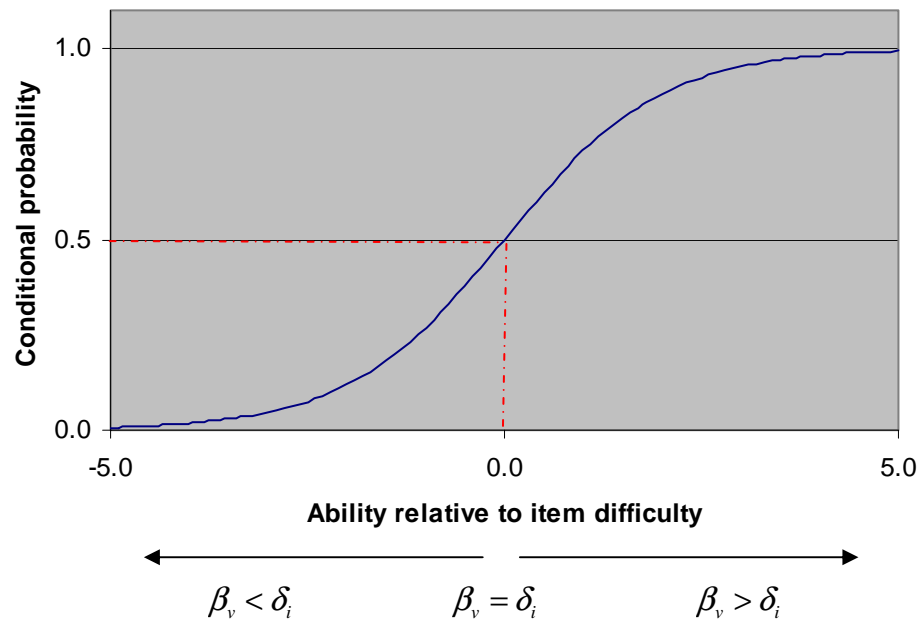
The item characteristic curve provides the opportunity to directly establish the probability of a person of ability β_v answering an item of difficulty δ_i correctly. For example, if in Figure 3.4 a person with ability $\beta_v = 0.0$ interacts with an item of difficulty $\delta_i = 0.0$ the probability 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	B	C	D
COMPLETE GUESS	PARTIAL GUESS	ALMOST CERTAIN	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 $x = a$.

A	B	C	D
COMPLETE GUESS	PARTIAL GUESS	ALMOST CERTAIN	CERTAIN

b. Let $[\![x]\!]$ be the greatest integer less than or equal to x.

(i) Show that $\lim_{x \to 2} f(x)$ exists if $f(x) = [\![x]\!] + [\![-x]\!]$.

A	B	C	D
COMPLETE GUESS	PARTIAL GUESS	ALMOST CERTAIN	CERTAIN

(ii) Is $f(x) = [\![x]\!] + [\![-x]\!]$ 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

114

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.

CHAPTER 4: QUALITATIVE INVESTIGATION

In this chapter I address the third research subquestion:

What are student preferences regarding different assessment formats?

4.1 QUALITATIVE DATA ANALYSIS

According to Schumacher and McMillan (1993), qualitative data analysis is primarily an inductive process of organising the data into categories and identifying patterns (relationships) among the categories. Unlike quantitative procedures, most categories and patterns emerge from the data, rather than being imposed on the data prior to data collection.

4.2 QUALITATIVE INVESTIGATION

In the qualitative component of my research study, I relied upon the qualitative method of interviewing. The format of the interview was described in section 3.3.1. In qualitative research, the role of the researcher in the study should be identified and the researcher should provide clear explanations to the participants. As researcher and interviewer, I investigated what the interviewees experienced being exposed to alternative assessment formats in their undergraduate studies and how they interpreted these experiences. The interview questions were presented in section 3.3.1.

In this section, I present the data that was gathered, in the form of interviews and an analysis of the data. The qualitative data findings are presented as a narration of the interviewees' responses. The data is used to illustrate and substantiate the third research subquestion of this research study related to student preferences i.e. What are student preferences regarding different assessment formats? Analysis is often intermixed with presentation of the data, which are usually quotes by the interviewees.

The issues discussed in this section focus on how a group of first year tertiary students, registered for the Mathematics I Major course at the University of the Witwatersrand, view the different assessment formats, both PRQ and CRQ, that they have been exposed to in their assessment programme. Relevant quotes from each interview were selected and will be discussed to highlight the most important beliefs, attitudes and inner experiences that this group of students had concerning the different assessment formats in their assessment programme.

- ## In favour of alternate assessment formats

The interviewee was a Chinese female student with an October class record of 70%. The following extract from her interview illustrates that this student enjoyed both the PRQ and CRQ formats of assessment.

Interviewer: You saw that a percentage of your tests was multiple choice and a percentage was always long questions and your tutorial tests were only multiple choice. Did you like those different formats?

Candidate: Ja, I did, 'cos multiple choice gives you an option of , y'know, the right answer's there somewhere so it kind of relieves you a bit and then you balance it off with a nice, um, long question so it's not... you aren't just depending on your luck but you're also applying your knowledge and I think that's.. that's cool.

This candidate was an average to high achieving student with a good work ethic. She attended all her classes and tutorials and often came for additional assistance. She had a positive attitude towards the different assessment formats, explaining that she liked both PRQs and CRQs as 'they balanced each other off'. She felt secure with both formats since in the MCQs she knew that one of the options provided was the correct answer, and the CRQs provided the opportunity to apply her knowledge which she felt very comfortable with.

- ## MCQs test a higher conceptual level

The interviewee was a black male student with an October class record of 81%. The following extract from his interview illustrates the student's perceptions of the different learning approaches he believed to have used for PRQs and CRQs.

123

Interviewer: Do you feel that the mark you got for the MCQ section is representative of your knowledge?

Candidate: (Laughs) Well, it depends, I mean, if I got a low mark then it means that I don't understand anything and it's not exactly like that. So, I wouldn't say it represents my knowledge or anything like that.

Interviewer: So what does it represent?

Candidate: (Laughs) Well, it simply means that maybe I didn't understand all the concepts very very well. I'm not digging deep into the concept, I'm just doing it on the surface, that's all.

Interviewer: I see and is that what multiple choice probes?

Candidate: I think so.

Interviewer: Deeper?

Candidate: Ja, ja. It requires a lot of knowledge because some questions are very short and we take the long way trying to do it and we run out of time. So you really need to understand what you are doing in multiple choice.

This candidate was a high achieving student who performed consistently well throughout the MATH109 course. He was of the opinion that MCQs are not fully representative of his mathematical knowledge as he approaches MCQs on the surface, rather than adopting a deeper learning approach towards MCQs. However, he does admit that some MCQs do test a higher conceptual level of understanding and for such MCQs, one requires a good mathematical knowledge. He also mentions the problem that MCQs testing higher cognitive skills are time consuming, and if you do not have a good understanding of the concept you could 'run out of time'.

- CRQs provide for partial credit

The interviewee was a coloured female student with an October class record of 81%. The following extract from her interview illustrates that this student prefers CRQs to PRQs because of the factor of partial credit.

Interviewer: Which type of question do you prefer?

Candidate: Um.. overall, I have to say traditional because in a way if you are doing an MCQ question and you get an answer and it doesn't appear there, you like sort of... your heart sinks, you know, it's like oh my word, what have I done wrong? But um... you know, also in traditional… ja, you can't be right… you don't know if you're completely wrong or if you're right and you know that at least you'll get some marks along the way for doing what you could. So… but, overall, I do prefer the traditional questions because, ja, you can freestyle. (Laughs).

This candidate was a high achiever and an independent student. Earlier on in the interview she had stated that she liked both assessment formats because:

it's good that we get asked different ways because it shows that we really understand and we know how to apply. It's not just doing it like out of routine.

When I probed her about the assessment format she preferred, she chose the CRQ format for the reason that if your answer to an MCQ was incorrect no marks were awarded, but even if your answer to a CRQ was incorrect, you could get partial marks for method. She also mentioned that since there was no negative marking in the MCQs, she always felt encouraged to answer these, even if at her first attempt her answer did not correspond to any of the provided options.

- Confidence plays an important role in assessment

The interviewee was a white female student with an October class record of 58%. The following extract from her interview illustrates that this student had little confidence in her performance in the mathematics tests and examinations, both PRQ and CRQ.

Interviewer: Do you have confidence in answering questions in maths tests which are different to the traditional types of questions?

Candidate: Fluctuated. Bit of a roller coaster.

Interviewer: Can you explain what you mean?

Candidate: It's got a lot to do with mental blocks as well. I prepared a lot more for the June test and my head was more around it. Mark really helped me. I was sort of in the Resource Centre lots and he really helped me get my head around it.

This candidate was an average ability student, struggling to cope with the pressures of her first year studies, as well as getting used to residence life away from her family. This candidate's performance in the two types of assessment was very erratic. In the April test, she scored poorly in the MCQs, in the June test she scored higher in the MCQs than in the CRQs and in the September test she again scored poorly in MCQs. She justified this fluctuation due to her having 'mental blocks' about the MCQs which she appeared to have little confidence in. She did admit that her performance was also strongly linked to the amount of preparation before each test. For the June test, she received a lot of extra assistance from the tutor in the Mathematics Resource Centre which not only helped her to gain a greater understanding of the content material, but also improved her confidence. It was pointed out that none of the students had been exposed to the PRQ format in their secondary school education, and so this assessment format was totally unfamiliar to them. The students thus lacked the confidence which they had gained with the CRQ assessment format in their secondary education, in which the predominant assessment format in the mathematics tests and examinations was the traditional, long open-ended question. The candidate was of the opinion that she would have performed better in the MCQs if she had had more exposure to this format, thereby increasing her confidence in this assessment format.

Another interesting quote from the candidate, linked to confidence, was the fact that she regarded the MCQs as more challenging than the CRQs.

Interviewer: In your school background were you exposed to different types of questions in Mathematics?

Candidate: We were, um, not as like... not such a broad spectrum but we were. We didn't really do MCQ as such in Maths but um... I

think it… ja… the MCQs are definitely challenging because, I don't know, in most subjects they are, you know, like…

Interviewer: What makes them challenging?

Candidate: I actually… it's weird because whenever you write a test and then people are like "Is it MCQ or long questions?" If you say it's long questions people are like phew… you know...

Interviewer: Okay.

Candidate: With MCQ it's like, "Oh my word!" because I think also, besides the fact that you're limited to one choice out of four, five, um… in long questions you can express yourself more because it's not like this or that, you know, there is some inbetween.

● MCQs require good reading and comprehension skills

The interviewee was a coloured male student with an October class record of 59%. The following extract from his interview illustrates his opinion on the importance of visual (graphical) PRQs and CRQs.

Interviewer: How would you ask questions in Maths tests if you were responsible for the course?

Candidate: Well, the way it's been done is great, I think, um, because it's not… it's not the old boring do the sum, do that sum, there's a whole lot of variations within the course which is great and it shouldn't be boring…

Interviewer: Okay.

Candidate: …but it… I think this is good.

Interviewer: Are there any other types of questions you could recommend that could be incorporated into Maths?

Candidate: Um, no. Well, maybe reading of graphs.

Interviewer: Okay.

Candidate: And finding the intercepts and the… say if this is increasing or decreasing and…

Interviewer: More graph interpretation questions?

Candidate: Yes.

This candidate was an average performing student who showed a very positive attitude towards the variety of assessment formats in the mathematics course. Earlier on in the interview he expressed his beliefs why he did not seem to perform well in the MCQ assessment format. He felt that it was due to the phrasing of the questions. So this student linked his poor performance to his reading and comprehension inabilities. He recommended that more visual (graphical) items should be included in the different assessment formats. He was of the opinion that such types of questions did not rely on reading and comprehension skills as much as the more theoretical questions.

Interviewer: When you looked at the multiple choice questions, what was it about them that you think made you perform badly?

Candidate: I think it was just the phrasing in different ways 'cos you phrased the question differently to what we expected. You didn't expect to… to see that type of question, but it was tricky.

- PRQ format lends itself to guessing and cheating

The interviewee was a black male student with an October class record of 43%. The following extract from his interview illustrates the student's opinion about the guessing factor involved in MCQs.

Interviewer: Which types of questions do you prefer in Maths?

Candidate: Uh, I like long questions. Ja, I like long questions very much. I don't like MCQs.

Interviewer: Why?

Candidate: Uh, MCQs… what can I say about them? Ja, sometimes they are like deceiving 'cos maybe when you want to work out… work out the solution then you say, "Ah, I can't do this thing," you just maybe choose an answer randomly, but on long questions you… you are trying to make sure that, at least, you get a solution, you see, so that's why I don't like MCQs 'cos somewhere we are not working as students. You just say, "Oh, I don't get it," then I tick A, but on long questions you are trying by all means to get that six marks or five marks.

Interviewer: Oh, so it's guessing?

Candidate: Ja! Ja, guessing, guessing.

This candidate was a low achieving student who was not in favour of the alternate assessment formats. He believed that his poor performance was linked to the inclusion of the PRQ format in the mathematics tests and examinations. He went on to explain that he preferred the traditional long CRQs to the MCQs as he considered MCQs as questions that promote guessing. He believed that if you did not have any options to choose from, you would be more careful in your working out of the solution. He expressed the opinion that 'we are not working as students' with MCQs, because if he cannot arrive at one of the solutions in the options, he simply guesses the answer, whereas with the CRQs, he would try to achieve the allocated marks by 'trying all means' at finding the solution. He did not consider guessing as a fair method of arriving at a solution. In fact, later on in the interviewee, he hinted to the fact that he thought CRQs were more reliable as it was more difficult to cheat with CRQs than with MCQs.

Candidate: …another point because MCQs, there's.. there's a great possibility of cheating.

Interviewer: Okay.

Candidate: 'Cos if you can't get something you just look to the person next to you. Oh, you just copy.

● Alternate formats add depth to assessment

The interviewee was an Indian female student with an October class record of 68%. The following extract from her interview illustrates the student's opinion about the proportion of PRQs and CRQs that should be included in mathematics tests and examinations.

Interviewer: What percentage of questions should be MCQ and what percentage should be long questions?

Candidate: I think about seventy percent should be MCQ and the rest should be long questions because it's... sometimes it's harder to understand than MCQ questioning despite understanding the knowledge, you know, understanding the maths and the theory

that you get 'cos it's very tricky sometimes. But I think it separates like your A's from your B's, you know, your like seventy-fives from your sixties. It's a good way to see what type of student you are.

This candidate was an average performing student who confessed that in mathematics the MCQ format had actually raised her marks. She explained that with MCQs, 'there's a whole technique to be learnt', and she felt confident that she had mastered this technique. She expressed the opinion that a greater percentage of MCQ should be included in mathematics tests and examinations as she believed that this type of assessment format separated the distinction 'A' candidates from the good 'B' candidates. So in her opinion, the performance of the students in the MCQs was a good measuring stick of their overall mathematical ability.

- Diagnostic purpose

The interviewee was an Indian male student with an October class record of 75%. The extract from his interview illustrates this candidate's opinion on how MCQs could be used for diagnostic purposes.

Interviewer:	Do you like the different formats of assessment in your maths tests?
Candidate:	Um, no, it's okay, but… Ja I think that… no, the papers have been up to standard so far. I don't think there really is a problem, especially like, um, the MCQs I felt really like gives you… it really tests your understanding of how to, you know, of all your calculations and stuff. I don't really think there's a problem with the way we've been tested so far.
Interviewer:	Which type of questions do you prefer, MCQs or traditional long questions?
Candidate:	Well, personally, I don't like the MCQs because sometimes you think you've got the right answer but, you know, you might have made a mistake somewhere in your calculations. You saw it or your right answer there then… but I think that the MCQs are

130

probably designed that way. Like you would have probably picked up what kind of mistakes we would have made so… so I think, ja, there should be a variety of different questions.

This candidate was amongst the top achieving students in the class. He liked the challenging questions and expressed the opinion that these could be of the PRQ or CRQ format. For this candidate it was not about the format of the question, but rather the cognitive level of skills required to answer the question. He felt that the MCQs had the diagnostic purpose of really testing understanding of knowledge and of methods of solving. With MCQs, an incorrect distracter chosen by the student is often a good indicator of the 'kind of mistakes we would have made' in the CRQs, thus identifying any misconceptions that the student might have. This candidate felt that a variety of different questions was necessary to diagnose common errors.

● Distracters can cause confusion

The candidate was a white male student, with an October class record of 37%. In the extract, the student expresses the frustrations he experienced with MCQs if two of the distracters were very similar to each other.

Interviewer: Which type of questions do you prefer in Maths?

Candidate: I feel more confident with the long questions than short questions, ja, than multiple choice 'cos multiple choice… two answers can be really close and you think about what you could have done wrong or what could be…if it is actually right then keep on going over it and over it and then you end up choosing one and end up being wrong.

This candidate was a poorly performing student, who admitted earlier in the interview that he had not been taking his studies seriously. He had not been attending classes regularly and had not studied for his tests. He did not have any preference for the type of assessment format, although he did feel more confident with the CRQ format. His lack of confidence in the MCQs was linked to the fact that often the distracters were very similar to each other and he found

it difficult to make the correct choice. He did not have enough confidence to trust his calculation of the correct answer, and when faced with the situation of two answers very close in value or nature to each other, he doubted his calculation. This lack of confidence was also evident in his performance in the CRQ format.

In summary, a qualitative analysis of these interviews appears to indicate that there were two distinct camps; those in favour of PRQs and those in favour of CRQs. Those in favour of PRQs expressed their opinion that this assessment format did promote a higher conceptual level of understanding; greater accuracy; required good reading and comprehension skills and was very successful for diagnostic purposes. Those against PRQs were of the opinion that they encouraged guessing; gave no credit for incorrect responses; that students lacked confidence in this format linked to the choice of distracters and that PRQs promoted a surface learning approach.

Those in favour of CRQs were of the opinion that this assessment format promoted a deeper learning approach to mathematics; required good reading and comprehension skills; partial marks could be awarded for method and students felt more confident with this more traditional approach. Those against CRQs generally felt that they were time consuming; did not provide any choice of distracters as a guide to a method of solution and that their poor performance in this assessment format was linked to their reading, comprehension and problem-solving inabilities.

From the students' responses, it seems as if the weaker students prefer CRQs. These students expressed a lack of confidence in PRQs, with one of the interviewees justifying her lack of confidence in this assessment format as a 'mental block'. The weaker students seemed to perform better in CRQ assessment format, thus resulting in a greater confidence in this format. The attitudes of weaker students to the PRQ format illustrate the important role that confidence plays in assessment. Weaker ability students also felt threatened by the fact that if their answer to an MCQ was incorrect, no marks were awarded,

whereas with CRQs, partial marks were awarded even if the answer was incorrect. Weaker students often lack the necessary reading and comprehension skills required to answer MCQs successfully. One of the weaker students opposing MCQs felt that the PRQ format lends itself to 'guessing and cheating'. The weaker ability students also expressed their frustration with MCQs if two or more of the distracters were very similar to each other. They felt that distracters can cause confusion, and this in turn would affect their performance.

The results from the qualitative investigation highlighted the most important beliefs, attitudes and inner experiences that this group of students of various mathematical abilities had concerning the PRQ and CRQ assessment formats in their mathematics assessment programme. These results address the research subquestion regarding the student preferences with respect to the different assessment formats.

CHAPTER 5:   THEORETICAL FRAMEWORK

In this chapter, I identify an assessment taxonomy consisting of seven *mathematics assessment components*, based on the literature. I attempt to develop a theoretical framework with respect to the mathematics assessment components and with respect to three measuring criteria: *discrimination index*, *confidence index* and *expert opinion*. The theoretical framework forms the foundation against which I construct the proposed model for measuring how *good* a mathematics question is. In this way, the first two research subquestions are addressed:

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

- 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?

I also elaborate on the parameters used in my research study for judging a test item. Finally, I describe the model developed for my research for measuring a good question.

In Section 5.1, I wish to elaborate on the proposed mathematics assessment components which were originally identified in this study from the literature. I also identify and discuss question examples, both PRQs and CRQs, within each mathematics assessment component.

In Section 5.2, I elaborate on the parameters I have identified for judging a test item.

In Section 5.3, I develop a model for measuring how good a mathematics question is that will be used both to quantify and visualise the quality of a mathematics question.

5.1 MATHEMATICS ASSESSMENT COMPONENTS

Based on the literature reviewed on assessment taxonomies in Section 2.4 and adapting Niss's assessment model for mathematics (Niss, 1993) reviewed in Section 2.3, I propose an assessment taxonomy pertinent to mathematics. This taxonomy consists of a set of seven items, hereafter referred to as the *mathematics assessment components*. In this research study, I investigated which of the assessment components can be successfully assessed in the PRQ format, and which can be better assessed in the CRQ format. To assist with this process, I used the proposed hierarchical taxonomy of seven mathematics assessment components, ordered by the cognitive level, as well as the nature of the mathematical tasks associated with each component. This mathematics assessment component taxonomy is particularly useful for structuring assessment tasks in the mathematical context. The proposed set of seven mathematics assessment components are summarised below:

(1)   Technical
(2)   Disciplinary
(3)   Conceptual
(4)   Logical
(5)   Modelling
(6)   Problem solving
(7)   Consolidation

Corresponding to Niss's assessment model (Niss, 1993) reviewed in Section 2.3, in this proposed set of seven mathematics assessment components, questions involving manipulation and calculation would be regarded as *technical.* Those that rely on memory and recall of knowledge and facts would fall under the *disciplinary* component. Assessment components (1) and (2) include questions based on mathematical facts and standard methods and techniques. The *conceptual* component (3) involves comprehension skills with algebraic, verbal, numerical and visual (graphical) questions linked to standard applications. The assessment components (4), (5) and (6) correspond to the

135

logical ordering of proofs, *modelling* with translating words into mathematical symbols and *problem solving* involving word problems and finding mathematical methods to come to the solution. Assessment component (7), *consolidation*, includes the processes of synthesis (bringing together of different topics in a single question), analysis (breaking up of a question into different topics) and evaluation requiring exploration and the generation of hypothesis.

Comparing with Bloom's taxonomy (Bloom, 1956), reviewed in Section 2.4, components (1) and (2) would correspond to Bloom's level 1: *Knowledge*. This lower-order cognitive level involves knowledge questions, requiring recall of facts, observations or definitions. In assessment tasks at this level, students are required to demonstrate that they know particular information. Components (3) and (4) correspond to Bloom's level 2: *Comprehension* and level 3: *Application*. These middle-order cognitive levels involve comprehension and application type questions which call on the learner to demonstrate that she/he comprehends and can apply existing knowledge to a new context or to show that she/he understands relationships between various ideas. Mathematics assessment components (5), (6) and (7) all correspond to Bloom's highest cognitive levels: level 4: *Analysis*; level 5: *Synthesis* and level 6: *Evaluation*. These levels involve tasks requiring higher-order skills such as analysing, synthesising and evaluating. At this cognitive level, the learner is required to go beyond what she/he knows, predict events and create or attach values to ideas. Problem solving might be required here where the learner is required to make use of principles, skills or his/her own creativity to generate ideas.

A modification of Bloom's taxonomy, adapted for assessment, called the MATH taxonomy (Smith *et al.*, 1996) was discussed in Section 2.4 in the literature review. The MATH taxonomy has eight categories, falling into three main groups. Group A tasks include those tasks which require the skills of factual knowledge, comprehension and routine use of procedures. In the proposed mathematics assessment component taxonomy, assessment components (1) and (2) -Technical and Disciplinary, would correspond to these Group A tasks. In the MATH taxonomy Group B tasks, students are required to apply their

learning to new situations, or to present information in a new or different way. Such tasks require the skills of information transfer and applications in new situations, and would correspond to assessment components (3) - Conceptual and (4) - Logical. The third group in the MATH taxonomy, Group C encompasses the skills of justification, interpretation and evaluation. Such skills would relate to the mathematics assessment components (5) - Modelling, (6) - Problem solving and (7) - Consolidation. One of the main differences between Bloom's taxonomy and the MATH taxonomy is that the MATH taxonomy is context specific and is used to classify tasks ordered by the nature of the activity required to complete each task successfully, rather than in terms of difficulty.

Using Bloom's taxonomy and the MATH taxonomy, the proposed mathematics assessment components can be classified according to the cognitive level of difficulty of the tasks as shown in Table 5.1

Table 5.1:   Mathematics assessment component taxonomy and cognitive level of difficulty.

Mathematics assessment components	Cognitive level of difficulty
1. Technical 2. Disciplinary	Lower order / Group A
3. Conceptual 4. Logical	Middle order / Group B
5. Modelling 6. Problem solving 7. Consolidation	Higher order / Group C

Table 5.2 summarises the proposed mathematics assessment components and the corresponding cognitive skills required within each component. These skills were identified by the researcher, based on the literature review, as being the necessary cognitive skills required by students to complete the mathematical tasks within each mathematics assessment component.

Table 5.2: Mathematics assessment component taxonomy and cognitive skills.

Mathematics assessment Components	Cognitive skills
1. Technical	• Manipulation • Calculation
2. Disciplinary	• Recall (memory) • Knowledge (facts)
3. Conceptual	Comprehension: • algebraic • verbal • numerical • visual (graphical)
4. Logical	• Ordering • Proofs
5. Modelling	Translating words into mathematical symbols
6. Problem solving	Identifying and applying a mathematical method to arrive at a solution
7. Consolidation	• Analysis • Synthesis • Evaluation

5.1.1 Question examples in assessment components

In the following discussion, one question within each mathematics assessment component has been identified according to Table 5.2, from the MATH109 tests and examinations. The classification of the question according to one of the assessment components was validated by a team of lecturers (experts) involved in teaching the first year Mathematics Major course at the University of the Witwatersrand. In addition, the examiner of each test or examination was asked to analyse the question paper by indicating which assessment component best represented each question. In this way, the examiner could also verify that there was a sufficient spread of questions across assessment components, and in particular, that there was not an over-emphasis on questions in the technical and disciplinary components. This exercise of indicating the assessment component next to each question also assisted the moderator and external examiner to check that the range of questions included all seven mathematics assessment components, from those tasks requiring lower-order cognitive skills to those requiring higher-order cognitive skills.

Assessment Component 1: Technical

If $z = 3 + 2i$ and $w = 1 - 4i$, then in real-imaginary form $\dfrac{z}{w}$ equals:

A. $\dfrac{-5}{17} + \dfrac{14i}{17}$

B. $\dfrac{5}{15} - \dfrac{14i}{\sqrt{15}}$

C. $3 - 4i$

D. $\dfrac{11}{17} + \dfrac{14i}{17}$

MATH109 August 2005, Tutorial Test, Question 5.

In this t*echnical* question, students are required to manipulate the quotient of complex numbers, z and w, by multiplying the numerator and denominator by the complex conjugate $\overline{w}$, and then to calculate and simplify the resulting quotient by rewriting it in the real-imaginary form, $\alpha + bi$.

Assessment Component 2: Disciplinary

If $f(x) = \dfrac{\sin x}{x}, x \neq 0,$ which of the following is true?

A. f is not a function.

B. f is an even function.

C. f is a one-to-one function.

D. f is an odd function.

MATH109 March 2005, Tutorial Test A, Question 1.

In this *disciplinary* question, students have to recall the definitions and properties of a function, an even function, a one-to-one function and an odd function, in order to decide which one of the given statements correctly describe the given function $f(x)$. Such a question requires the cognitive skill of memorising facts and then remembering this knowledge when choosing the best option.

In the following discussion, three question examples have been chosen to illustrate three of the comprehension type cognitive skills: verbal, numerical and visual (graphical), that are required by students to complete the tasks within the conceptual mathematics assessment component.

Assessment Component 3: Conceptual

State why the Mean Value Theorem does not apply to the function $f(x) = \dfrac{2}{(x+1)^2}$

on the interval $[-3, 0]$

   A. $f(-3) \neq f(0)$

   B. f is not continuous

   C. f is not continuous at $x = -3$ and $x = 0$

   D.   Both A and B

   E.   None of the above

MATH109 June 2006, Section A: MCQ, Question 7.

In the above *conceptual* question, the student is required to apply his/her knowledge of the Mean Value theorem to a new, unfamiliar situation which requires that the student selects the best *verbal* reason why the Mean Value theorem does not apply to the function $f(x)$ and the interval given in the question. This question requires a comprehension of all the hypotheses of the Mean Value theorem and tests the students' understanding of a situation where one of the hypotheses to the theorem fails.

Assessment Component 3: Conceptual

$$\lim_{x \to \infty} \left(1 + \frac{2}{x} \right)^x =$$

   A. 2

   B. e^2

   C. ∞

   D. 1

   E. Does not exist

MATH109 November 2005, Section A: MCQ, Question 2.

In the *conceptual* question above, the student is required to apply his/her knowledge of the definition of Euler's number e, which is defined in lectures as:
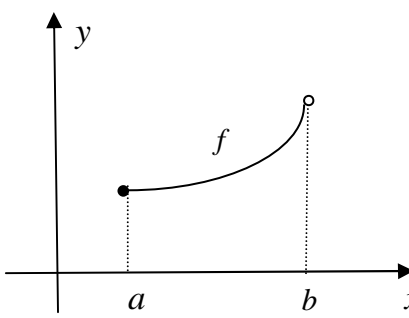
$$\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x = e$$

They need to make a conjecture and extrapolate from this definition to choose the best *numerical* option for $\lim_{x \to \infty} \left(1 + \frac{2}{x}\right)^x$.

This result had not been discussed in class, and hence is not a familiar result to the students.

Assessment Component 3: Conceptual

Determine from the graph of $y = f(x)$ whether f possesses extrema on the interval $[a, b]$

A. Maximum at $x = a$; minimum at $x = b$.

B. Maximum at $x = b$; minimum at $x = a$.

C. No extrema.

D. No maximum; minimum at $x = a$.

MATH109 May 2006, Section A: MCQ, Question 1.

In this *graphical conceptual* question, students are required to apply their knowledge of the Extreme Value theorem and the definition of relative extrema on an interval I. There is no algebraic calculation necessary of the values of the extrema on the closed interval $[a,b]$. The Extreme Value theorem is an existence theorem because it tells of the existence of minimum and maximum values, but does not show how to find these values. Students need to examine the graph of

the given function f and consider how f behaves at the end points as well as how the continuity (or lack of it) has affected the existence of extrema on the given interval. The choice of the correct option is assisted by having a *visual* figure when the decision is made.

Assessment Component 4: Logical (PRQ)

Decide whether Rolle's theorem can be applied to $f(x) = x^2 + 3x$ on the interval $[0, 2]$.

If Rolle's theorem can be applied, find the value(s) of c in the interval such that $f'(c) = 0$. If Rolle's theorem cannot be applied, state why.

    A. Rolle's theorem can be applied; $c = \dfrac{-3}{2}$

    B. Rolle's theorem can be applied; $c = 0, c = 3$

    C. Rolle's theorem does not apply because $f(0) \neq f(2)$

    D. Rolle's theorem does not apply because $f(x)$ is not continuous on $[0, 2]$

<div align="right">MATH109 May 2006, Section A: MCQ, Question 5.</div>

This *logical* PRQ firstly requires the student to recall the conditions of Rolle's theorem to decide whether Rolle's theorem can be applied to the given function. Such a decision requires the conceptual skill of *ordering* the conditions stated in the proof of Rolle's theorem, and checking that the three conditions of:
(i) continuity on $[0, 2]$, (ii) differentiability on $(0, 2)$ and (iii) $f(0) = f(2)$, are met. Once the decision is made, the student can proceed to the second part of the question which requires the student to find the value(s) of c in $(0, 2)$ such that $f'(c) = 0$. The logical ordering of the conditions of Rolle's theorem leads to the student realising that since the last condition is not met i.e. $f(0) \neq f(2)$, Rolle's theorem does not apply.

A further example within the logical assessment component has been provided below, this example being a constructed response question appearing in MATH 109 June 2006, Section C: Calculus.

Assessment Component 4: Logical (CRQ)

(a) In the proof of the following theorem, the order of the statements is incorrect. Give a correct proof of the theorem by reordering the statements. You need only list the statement numbers in their correct order.

Theorem:

If a function f is continuous on the closed interval $[a,b]$ and F is an antiderivative of f on the interval $[a,b]$, then $\int_a^b f(x)dx = F(b) - F(a)$

① Since F is the antiderivative of f, $F'(c_i) = f(c_i)$

② $\therefore f(c_i) = \dfrac{F(x_i) - F(x_{i-1})}{\Delta x_i}$

③ $\therefore \displaystyle\sum_{i=1}^{n} f(c_i)\Delta x_i = \sum_{i=1}^{n} \left[F(x_i) - F(x_{i-1}) \right] = F(b) - F(a)$

④ By the Mean Value theorem, there exists $c_i \in (x_{i-1}, x_i)$ such that

$$F'(c_i) = \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}}$$

⑤ Divide the closed interval $[a, b]$ into n subintervals by the points

$$a = x_0 < x_1 < x_2 < ... < x_{i-1} < x_i < ... < x_{n-1} < x_n = b$$

⑥ Taking the limit as $n \to \infty$, $F(b) - F(a) = \displaystyle\lim_{n \to \infty} \sum_{i=1}^{n} f(c_i)\Delta x_i = \int_a^b f(x)dx$

⑦ $F(b) - F(a) = \displaystyle\sum_{i=1}^{n} \left[F(x_i) - F(x_{i-1}) \right]$

⑧ $\therefore f(c_i)\Delta x_i = F(x_i) - F(x_{i-1})$

Correct order: (Only list the statement numbers.)

(b) What is the theorem called?

(c) What kind of series is the series on the right hand side of statement ⑦?

MATH109 June 2006, Section C: Calculus, Question 4.

This *logical* CRQ requires the students to recall the proof of the Fundamental Theorem of Calculus. Although the proof is given, the statements appear in the incorrect order. The students are required to reorder the given statements to correct the proof. Such a reordering process involves the cognitive skill of logical *ordering*.

Assessment Component 5: Modelling (CRQ)

Following the record number in attendance during the opening day of the Rand Easter show this year, organisers are planning a special event for the opening eve in 2007. Murula.com will sponsor a ten-seater jumbo jet, carrying all eight members of the organisation committee, to fly in a western direction at 5000 m/minute, at an altitude of 4000 m, over the show grounds that evening.

In order to ensure that all people participating in this event will be able to follow the jet from the surface at the show grounds, a special 10 000 W searchlight will be installed at the main entrance gate to keep track of the plane. The searchlight is due to be kept shining on the plane at all times.

N

W ← → E

S

⇐ Plane

4000 m

θ

Searchlight

x m
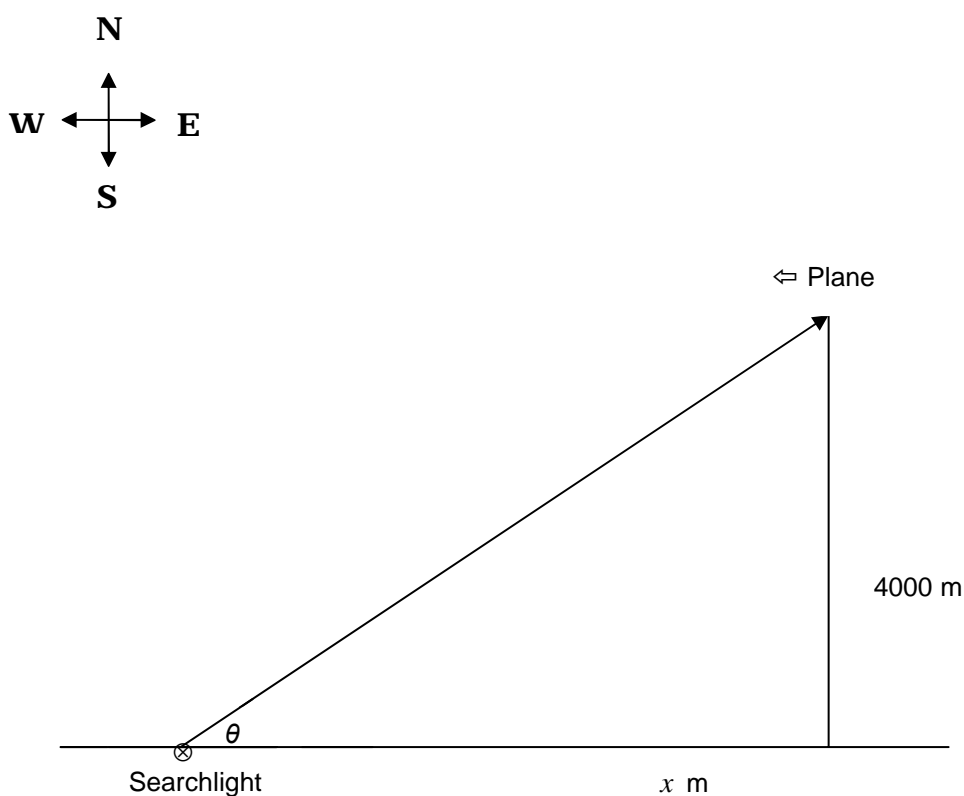
What will be the rate of change of the angle of the searchlight when the jet is due east of the light at a horizontal distance of 2000 m?

MATH109 May 2006, Section C: Calculus, Question 2.

In this *modelling* CRQ, students are required to translate the words into mathematical symbols and to use related rates to solve the real-life problem. To solve the related-rate problem, students firstly have to identify all the given quantities as well as the quantities to be determined. A sketch has been provided which can assist students to identify and label all these quantities. Secondly, students have to write an equation involving the variables whose rates of change either are given or are to be determined. Thirdly, using the Chain Rule, both sides of the equation must be implicitly differentiated with respect to time. Finally, all known values for the variables and their rates of change must be substituted into the resulting equation, so that the required rate of change can be solved for.

In modelling type questions, students have to develop a mathematical model to represent actual data. Such a procedure requires two conceptual skills: accuracy and simplicity. This means that the student's goal should be to develop a model that is simple enough to be workable, yet accurate enough to produce meaningful results.

Assessment Component 6: Problem solving (PRQ)

Which of the following is an antiderivative for $f(x) = x \cos x$?

A. $F(x) = \dfrac{1}{2} x^2 \cos x + 4$

B. $F(x) = \dfrac{1}{2} x^2 \sin x + 5$

C. $F(x) = x \sin x + \cos x - 1$

D. $F(x) = x \cos x + \sin x - 2$

E. None of the above.

MATH109 June 2006, Section A: MCQ, Question 5.

In this *problem solving* MCQ, the student is required to find his or her own method to arrive at the solution. Firstly, the student has to know what the

antiderivative of a function is in order to decide on a method. The solution can be arrived at by either integrating $f(x)$ using the technique of integration by parts, since $f(x)$ is a product of two differentiable functions, or by differentiating each function $F(x)$ provided in the distracters, using the Product Rule, until the original function $f(x)$ is obtained.

Assessment Component 6: Problem solving (CRQ)

This question deals with the statement

$P(n): n^3 + (n+1)^3 + (n+2)^3$ is divisible by 9, for all $n \in \mathrm{N}, n \geq 2$

(1.1) Show that the statement is true for $n = 2$.

(1.2) Use Pascal's triangle to expand and then simplify $(k+3)^3$.

(1.3) Hence, assuming that $P(k)$ is true for $k > 2$ with $k \in \mathrm{N}$, prove that $P(k+1)$ is true.

(1.4) Based on the above results, justify what you can conclude about the statement $P(n)$.

MATH109 June 2006, Section B: Algebra. Question 1.

In the *problem solving* CRQ, the students are required to use the principle of Mathematical Induction to prove that the statement $P(n)$ is true for all natural numbers $n \geq 2$. The CRQ has been subdivided into smaller subquestions involving different cognitive skills to assist the student with the method of solving using mathematical induction. In subquestion (1.1), the students need to establish truth for $n = 2$ by actually testing whether the statement $P(n)$ is true for $n = 2$. Hence (1.1) assess within the technical mathematics assessment component. Subquestion (1.2) involves a numerical calculation, the result of which will be used in the proof by induction. Hence (1.2) also assesses within the technical assessment component. In subquestion (1.3), students are required to complete the proof by induction, by assuming the inductive

146

hypothesis that $P(k)$ is true for $k > 2, k \in \mathrm{N}$, and proving that $P(k+1)$ is true.

Since subquestion (1.3) requires the cognitive skills of identifying and applying the principle of Mathematical Induction to arrive at a solution, (1.3) assesses within the problem solving mathematics assessment component. Subquestion (1.4) concludes the proof by requiring the students to justify that both of the conditions of the principle hold, and therefore by the principle of induction $P(n)$ is true for every $n \geq 2, n \in \mathrm{N}$. Hence (1.4), requiring no more than a simple manipulation, assesses within the technical assessment component. This problem solving CRQ illustrates that often those questions involving higher order cognitive skills subsume the lower order cognitive skills.

Assessment Component 7: Consolidation (PRQ)

Let $y = f(x) = \cos(\arcsin x)$. Then the range of f is

   A. $\{y \mid 0 \leq y \leq 1\}$

   B. $\{y \mid -1 \leq y \leq 1\}$

   C. $\{y \mid -\dfrac{\pi}{2} < y < \dfrac{\pi}{2}\}$

   D. $\{y \mid -\dfrac{\pi}{2} \leq y \leq \dfrac{\pi}{2}\}$

   E.   None of the above.

In the assessment component of *consolidation*, questions require the conceptual skills of analysis and synthesis and in certain cases evaluation. In the MCQ under discussion, students are required to *analyse* the nature of the function f, being a composition of both the functions $\cos x$ and $\arcsin x$. Within this analysis, consideration of the domain and range of each separate function has to be made. Once all the individual functions have been analysed with their restrictions on their domain and range, all this information has to be *synthesised* in order to make a conclusion about the resulting composite function, and the

restrictions on the domain and range of the composite function. An *evaluation* is finally required of the correct option which best describes the restriction on the range of the composite function.

Assessment Component 7: Consolidation (CRQ)

Let $[\![x]\!]$ be the greatest integer less than or equal to x.

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

  (ii) Is $f(x) = [\![x]\!] + [\![-x]\!]$ continuous at $x = 2$? Give reasons.

In the *consolidation* CRQ provided, students are expected to go beyond what they know about the greatest integer function $[\![x]\!]$. Part (i) requires an *analysis* of the behaviour of the function $f(x)$, being the sum of two greatest integer functions, as x approaches 2. In this analysis, the limit of each individual greatest integer function, $[\![x]\!]$ and $[\![-x]\!]$, needs to be investigated as x approaches 2. *Synthesis* is then required to complete the question, by summing up each individual limit, if they exist.

In part (ii), the student is required to make an *evaluation*, based on the results from part (i). A further condition of continuity needs to be checked i.e. the value of $f(2)$, and together with the result obtained in part (i), the student can make a judgement decision about the continuity of the function at $x = 2$. In this question, a consolidation of both the results from parts (i) and (ii) assists the student to make the overall evaluation. Such techniques of justifying, interpreting and evaluation are considered to be integral to the consolidation assessment component.

5.2 DEFINING THE PARAMETERS

In this research study, in order to define the parameters for developing a model to measure how good a mathematics question is, a few assumptions are made about mathematical questions. Firstly, we assume that the question is clear, well-written and checked for accuracy. We also assume that the question tests what it sets out to do. Issues such as ambiguity etc. are not considered. These are right or wrong and we assume correctness.

For developing a model for measuring a good question (described in section 5.3), we depart from the following four premises:

- A good question should discriminate well. In other words, high performing students should score well on this question and poor performing students are not expected to do well.

- Students' confidence when dealing with the question should correspond to the level of difficulty of the question. There is a problem with a question when it is experienced as misleadingly simple by students and subsequently leads to an incorrect response. In this case, students are over confident and do not judge the level of difficulty of the question correctly. Similarly, there is a problem if a simple question is experienced as misleadingly difficult and students have no confidence in doing it.

- The level of difficulty of the question should be judged correctly by the lecturer. When setting a question, the lecturer judges the level of difficulty intuitively. There is a problem with the question when the lecturer over or underestimates the level of difficulty as experienced by students.

- The level of difficulty of a question does not make it a good or poor question. Difficult questions can be good or poor, just as easy questions can be.

With these premises as background, three parameters were identified:

(i)   Discrimination index

(ii)   Confidence index

(iii)   Expert opinion

Although only these three parameters were used to develop a model to quantify the quality of a question, a fourth parameter was used to qualitatively contribute to the characteristics of a question:

(iv)   Level of difficulty

How these parameters were amalgamated to develop the model will be discussed in section 5.3. In this section we only clarify the parameters.

5.2.1 Discrimination index

The extent to which test items discriminate among students is one of the basic measures of item quality. It is useful to define an *index of discrimination* to measure this quality.   The discrimination index (DI) is computed from equal-sized high and low scoring groups on the test (say the top and bottom 27%) as follows:

$$DI = (C_H - C_L)/N \quad ; \quad \text{where}$$

C_H = number of students in the high group that responded correctly;

C_L = number of students in the low group that responded correctly;

N = number of students in both groups.

Using this definition, the discrimination index can vary from -1 to +1. Ideally, the DI should be close to 1. If equal numbers of 'high' and 'low' students answer correctly, the item is unsuccessful as a discrimination (DI = 0). If more 'low' than 'high' students get an item correct, the DI is negative, a signal for the examiner to improve the question.

For purposes of building up a test bank, a DI value of 0.3 is an acceptable lower limit. Using the 27% sample group size, values of 0.4 and above are regarded as high and less than 0.2 as low (Ebel, 1972).

The proportion of students answering an item correctly also affects its discrimination. Items answered correctly (or incorrectly) by a large proportion of students (more than 85%) have markedly reduced power to discriminate. On a good test, most items will be answered correctly by 30% to 80% of the students.

A few basic rules for improving the ability of test items to discriminate follow:

1. Items that correlate less than 0.2 with the total test score should probably be restructured. Such items do not measure the same skill or ability as does the test on the whole or are confusing or misleading to students. Generally, a test is better (i.e. more reliable) the more homogeneous the items. It is generally acknowledged that well constructed mathematics tests are more homogeneous than well constructed tests in social science (Kehoe, 1995). Homogeneous tests are those intended to measure the unified content area of mathematics.

   A second issue involving test homogeneity is that of the precision of a student's obtained test score as an estimate of that student's "true" score on the skill tested. Precision (reliability) increases as the average item-test correlation increases.

2. Distracters for PRQs that are not chosen by any students should be replaced or eliminated. They are not contributing to the test's ability to discriminate the good students from the poor students. One should be suspicious about the correctness of any item in which a single distracter is chosen more often than all other options, including the answer, and especially so if the distracter's correlation with the total score is positive.

3. Items that virtually everyone gets right are unsuccessful for discriminating among students and should be replaced by more difficult items (Ebel, 1965).

The Rasch model specifies that item discrimination, also called the item slope, be uniform across items. Empirically, however, item discriminations vary. The software package, Winsteps, estimates what the item discrimination parameter

would have been if it had been parameterised. During the estimation phase of Winsteps, all item discriminations are asserted to be equal, of value 1.0, and to fit the Rasch model. As empirical item discriminations never are exactly equal, Winsteps can report an estimate of those discriminations post-hoc (as a type of fit statistic). The empirical discrimination is computed after first computing and anchoring the Rasch measures. In a post-hoc analysis, a *discrimination parameter*, a_i, is estimated for each item. The estimation model is of the form:

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

P_{vix} = probability 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;

$F_x = $ Rasch-Andrich threshold.

In Winsteps, item discrimination is not a parameter. It is merely a descriptive statistic. The Winsteps reported values of item discrimination are a first approximation to the precise value of a_i. The possible range of a_i is $-\infty$ to $+\infty$, where $+\infty$ corresponds to a Guttman data pattern (perfect discrimination) and $-\infty$ to a reversed Guttman pattern. 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. Rasch estimation usually forces the average item discrimination to be near 1.0. An *estimated discrimination* of 1.0 accords with Rasch model expectations. Values greater than 1.0 indicate over-discrimination, and values less than 1.0 indicate under-discrimination. Over-discrimination is thought to be beneficial under classical (raw-score) test theory conventions (Linacre, 2005).

In classical test theory, the ideal item acts like a switch i.e. high performers pass, low performers fail. This is perfect discrimination, and is ideal for sample stratification. Such an item provides no information about the relative performance of low performers, or the relative performance of high performers. Rasch analysis, on the other hand, requires items that provide indication of relative performance along the latent variable as discussed in section 3.4. It is

152

this information which is used to construct measures. From a Rasch perspective, over-discriminating items tend to act like switches, not measuring devices. Under-discriminating items tend neither to stratify nor to provide information about the relative performance of students on those items.

A second important characteristic of a good item is that the best achieving students are more likely to get it right than are the worst achieving students. Item discrimination indicates the extent to which success on an item corresponds to success on the whole test. Since all items in a test are intended to cooperate to generate an overall test score, any item with negative or zero discrimination undermines the test. Positive item discrimination is generally productive, unless it is so high that the item merely repeats the information provided by other items on the test.

5.2.2 Confidence index

The *confidence index* (CI) has its origins in the social sciences, where it is used particularly in surveys and where a respondent is requested to indicate the degree of confidence he has in his own ability to select and utilise well-established knowledge, concepts or laws to arrive at an answer. In the science education literature, as well as the measurement literature (as discussed in section 2.14), a range of studies has considered some aspects of student confidence and how such confidence may impact students' test performance. Students' self-reported confidence levels have also been studied in the field of educational measurement to assess over- and underconfidence bias in students' test-taking practices (Pallier, Wilkinson, Danthiir, Kleitman, Knezevic, Stankov & Robertsw, 2002). In physics education research, Hasan *et al.* (1999) used a confidence index in conjunction with the correctness or not of a response, to distinguish between students' embedded misconceptions (wrong answer and high confidence) and lack of knowledge (wrong answer and low confidence) and to restrict guessing (Table 5.3). The CI is usually based on some scale. For example, in Hasan's (1999) study, a six-point scale (0 – 5) was used in which 0 implies no knowledge (total guess) of methods or laws required for answering a

particular question, while 5 indicates complete confidence in the knowledge of the principles and laws required to arrive at the selected answer. When a student is asked to provide an indication of confidence along with each answer, we are in effect requesting him to provide his own assessment of the certainty he has in his selection of the laws and methods utilised to get to the answer (Webb, 1994).

The decision matrix in Table 5.3 is used for identifying misconceptions in a group of students.

Table 5.3: Decision matrix for an individual student and for a given question, based on combinations of correct or wrong answers and of low or high average CI.

	Low CI	**High CI**
Correct answer	Lucky guess	Sufficient knowledge (understanding of concepts)
Wrong answer	Lack of knowledge	Misconception

(Adapted from Hasan *et al.*, 1999, p296).

If the degree of certainty is low i.e. low CI, then it suggests that guesswork played a significant part in the determination of the answer. Irrespective of whether the answer was correct or wrong, a low CI value indicates guessing, which, in tum, implies a lack of knowledge. If the CI is high, then the student has a high degree of confidence in his choice of the laws and methods used to arrive at the answer. In this situation, if the student arrived at the correct answer, it would indicate that the high degree of certainty was justified. Such a student is classified as having adequate knowledge and understanding of the concept. However, if the answer was wrong, the high certainty would indicate a misplaced confidence in his/her knowledge of the subject matter. This misplaced certainty in the applicability of certain laws and methods to a specific question is an indication of the existence of misconceptions.

Hasan *et al.* (1999) recommend that if the answers and related CI values indicate the presence of misconceptions, then feedback to students can be

154

modified with the explicit intent of removing the misconceptions. Furthermore, the information obtained by utilising the CI can also be used to address other areas of instruction. In particular, it can be used:

- as a means of assessing the suitability of the emphasis placed on different sections of a course

- as a diagnostic tool, enabling the teacher to modify feedback

- as a tool for assessing progress or teaching effectiveness when both pre- and post-tests are administered

- as a tool for comparing the effectiveness of different teaching approaches, including technology-integrated approaches, in promoting understanding and problem-solving proficiency.

In a study conducted by Potgieter, Rogan and Howie (2005) on the chemical concepts inventory of Grade 12 learners and University of Pretoria Foundation year students, the CI indicated general overconfidence of learners about the correctness of answers provided. It also showed that the guessing factor was less serious a complication than anticipated in the analysis of multiple choice items for the prevalence of specific misconceptions. Engelbrecht, Harding and Potgieter (2005) reported that first year tertiary students are also more confident of their ability to handle conceptual problems than to handle procedural problems in mathematics. They argue that the CI cannot always be used to distinguish between a lack of knowledge (wrong answer, low CI) and a misconception (wrong answer, high CI), since students could just be overconfident, or in procedural problems, students with high confidence may make numerical errors.

The literature is divided about whether self-evaluation bias facilitates subsequent performance. In some studies overconfidence appears to be associated with better performance (Blanton, Buunk, Gibbons & Kuyper, 1999), whereas other studies showed no long term performance advantage of overconfidence (Robins & Beer, 2001). Pressley *et al.* (1990) argue that the relationship between self-evaluation bias and subsequent performance depends on the motivational factors contributing to the exaggeration of confidence.

Exaggerated self-reports that are motivated by avoidance of self-protection are associated with poor subsequent performance, whereas exaggeration motivated by a strong achievement motivation is associated with improved future performance.

Ochse (2003) differentiated between overestimators, realists and underestimators based on the projection that students in third-year psychology made of their expected subsequent performance. Ochse found that, on average, overestimators (38% of sample) expected significantly higher marks than both realists and underestimators, were significantly more confident about the accuracy of their estimations, perceived themselves to have significantly higher ability than their peers, but achieved the lowest marks of the three groups (11.5% below class average, 20.6% lower than predicted). Underestimators, on the other hand (17% of sample), achieved the highest marks of the three groups (17.5% above class average, 14.3% above prediction) despite their unfavourable perceptions of their own ability and low confidence in their projected achievements. Ochse suggested that overoptimism may reflect ignorance of required standards and may result in complacency, inappropriate preparation or carelessness. The result of such ignorance is disappointment, frustration and anger when actual performance falls far short of expectations.

It should be noted that research on self-efficacy indicates a strong relationship between self-assessment and subsequent performance. Ehrlinger (2008) has pointed out that this relationship depends on the ability of respondents to control or regulate their actions in order to achieve the desired outcome. The close correlation between prediction of performance and self-efficacy also requires an accurate specification of a specific task.

In this research study, the CI values per item were calculated according to a 4-point Likert scale in which 1 implied a 'complete guess', 2 implied a 'partial guess', 3 for 'almost certain', while 4 indicated 'certain'. In terms of the Rasch model, a Likert scale is a format for observing responses wherein the categories increase in the level of the variable they define, and this increase is uniform for

all agents of measurement. The polytomous Rasch-Andrich rating scale model, discussed in section 3.4.1.3, was used in the Winsteps calculation of the CI.

5.2.3 Expert opinion

For purposes of this study, subject specialists were referred to as *experts* in terms of their mathematical knowledge of the content, as well as their experience in the methodological and pedagogical issues involved in teaching the content. Experts were asked to review test and examination items le in the first-year mathematics major course and to express their opinions on the level of difficulty of these questions. The aim of this exercise was to encourage the experts to look more critically at the questions, both PRQs and CRQs, and to express their opinions on the level of difficulty of each test item, independent of the students' performance in these items i.e. the predicted level of difficulty. The opinions were categorised into three main types using the following scale:

1:    student should find the question easy

2:    student should find the question of average difficulty but fair

3:    student should find the question difficult or challenging.

For the purpose of this study we consider the term *expert opinion* equivalent to *predicted performance.*

While giving their opinions, experts could reflect on the learning outcomes of the course, and on the assessment components corresponding to each test item. Such reflection would assist experts to write questions that guide students towards the kinds of intellectual activities they wish to foster, and raise their awareness of the effects of the kinds of questions they ask on their students' learning. In this context, Hubbard (2001) refers to Ausubel's meaningful learning, Skemp's description of relational understanding, Tall's definition of different types of generalisation and abstraction and Dubinsky and Lewin's reflective abstraction as all investigating in different ways, the kinds of intellectual activities which we desire our students to engage in. The experts involved in giving their opinions were not asked to familiarise themselves with

any of the above research papers. However, it was hoped that because they were successful mathematics thinkers themselves, the task of giving their opinions would enable them to recognise the intellectual activities required to solve different types of questions, in both the PRQ and CRQ formats.

All questions for which the experts expressed their opinion, involved subject matter which was familiar and covered a wide range of teaching and learning purposes. No model examples were given to the experts so that they would not be influenced by the researcher's views. The researcher did explain to the team of experts that their individual opinions would in no way classify questions as good or bad. This was not the intention of the task. To anticipate the problem that experts might have when trying to express their opinions on questions as being easy, average difficulty or challenging, not knowing exactly what information had been provided to students in lectures and tutorials, those involved in teaching the calculus course were asked for their expert opinions on the calculus PRQs and CRQs only, and those involved in teaching the algebra course were asked for their opinions on the algebra PRQs and CRQs only. In this way, the experts were completely familiar with the content, in particular knowing whether a question was identical or similar to one for which a specific model solution had been provided in lectures or tutorials, or whether this was not the case. The mathematical content is important because learning objectives that are not subject specific are more difficult for subject specialists to apply. One of the difficulties experienced by the experts in giving their opinions on how students experience the difficulty level of the test items, is that most experts are accustomed to thinking exclusively about the subject matter of the test item and their own view of mathematics, rather than about what might be going on in the minds of their students as they tried to answer the questions. By giving their opinions, there is an expectation that when experts set assessment tasks in the future, they will be influenced by their experiences and reflect on the purpose of their questions. The wording of the questions needs to reflect what kind of intellectual activity they intend for their students to engage in.

In this study, a panel of 8 experts were asked for their opinions. As this number was too low to apply any Rasch model, the expert opinion per item was calculated as the average of the individual expert opinions given per item. Winsteps will operate with a minimum of two observations per item or person. For statistically stable measures to be estimated, at least 30 observations per element are needed. The sample size needed to have 99% confidence that no item calibration is more than 1 logit away from its stable value is in the range $27 < N < 61$. Thus, a sample of 50 well-targeted examinees is conservative for obtaining useful, stable estimates. 30 examinees/observations is enough for well-designed pilot studies. Hence the Rasch model was not used in the calculation of the expert opinion per item.

5.2.4 Level of difficulty

Student performance was used as an estimate of the level of difficulty of an item, a common practice. The level of difficulty, although not a direct indication of the quality of the question, is a useful parameter when selecting questions to assemble a well-balanced set of questions.

In traditional test theory, difficulty level is defined as:

Difficulty level = number of correct responses/total number of responses.

An item that everyone gets wrong (difficulty level = 0.0) is unsuccessful. Equally unsuccessful is an item that everyone gets right (difficulty level = 1.0). In the Rasch logit-linear models, as discussed in Chapter 3, Rasch analysis produces a single difficulty estimate for each item and an ability estimate for each student. Through the application of this model, raw scores undergo logarithmic transformations that render an interval scale where the intervals are equal, expressed as a ratio or log odds units or logits (Linacre, 1994). A logit is the unit of measure used by Rasch for calibrating items and measuring persons. The difficulty scale starts from easy items (negative logits) and moves to more difficult ones (positive logits).

5.3 MODEL FOR MEASURING A GOOD QUESTION

In this section a model for measuring how good a mathematics question is will be developed that will be used both to quantify and visualise the quality of a good mathematics question.

5.3.1 Measuring criteria

To address the research questions of this study, three measuring criteria, based on the parameters discussed in section 5.2, were identified. These criteria form the foundation of the theoretical framework developed for the purpose of this study, and were used to diagnose the quality of a test item.

(1) *Point measure* as a discrimination index.

(2) *Confidence deviation*: the deviation between the expected students' confidence level and the actual student confidence for the particular item.

(3) *Expert opinion deviation*: the deviation between the expected student performance according to experts and the actual student performance.

(1) Point measure as a discrimination index

According to literature (Wright, 1992), there are numerous ways of conceptualising and mathematically reporting discrimination. The *point measure* and the Rasch discrimination index are two of them. In classical test theory, the point biserial correlation is the Pearson correlation between responses to a particular item and scores on the total test. In the Rasch model, the point measure correlation is a more general indication of the relationship between the performance on a specific item and the total test score, and is computed in the same way as the point biserial, except that Rasch measures replace total scores. It was therefore decided to use the point measure as the measure of discrimination, rather than the Rasch discrimination index. The point measure (rpm) is a number between 0 and 1.

In order to assign the same measuring scale to all three criteria, the discrimination was adapted by subtracting the point measure values (rpm) from 1 (the perfect correlation).

$$\therefore \text{Adapted discrimination} = 1 - rpm \quad (0 \le rpm \le 1)$$

The discrimination was adapted in this way so that the amount of departure of the point measure values from the perfect correlation value of 1 could be investigated. Thus, in this model, the closer the adapted discrimination is to 0, the better the correlation.

(2) Confidence deviation

In this study, the CI values per item were calculated according to a 4-point Likert scale as discussed in section 5.2.2:

        1 : complete guess

        2 : partial guess

        3 : almost certain

        4 : certain

To measure the confidence deviation, the confidence measure (average over the students) for each item was plotted against each corresponding item difficulty. A best fit regression line was fitted to the points, as shown in Figure 5.1.

Figure 5.1: Illustration of confidence deviation from the best fit line between item difficulty and confidence.
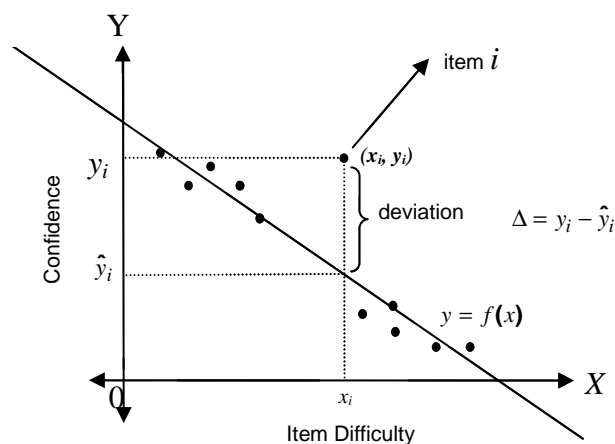
For any given item difficulty, the amount of deviation between the actual confidence measures and the confidence values as predicted by the best fit line, is measured by the vertical distance $|y_i - \hat{y}_i|$, where y_i is the observed confidence value and $\hat{y}_i$ is the predicted confidence value from the best fit line for item i. Small confidence deviation measures (close to 0) represent a small deviation of the confidence index from the item difficulty.

Ideally an item should lie on this regression line and should have a confidence deviation of 0. An item that lies far away from the line indicates that students were either over confident or under confident for an item of that particular level of difficulty.

(3) Expert opinion deviation

In this study, eight experts were asked to give their opinions on the difficulty values per item according to a scale as discussed in section 5.2.3:

      1: student should find the question easy

      2: student should find the question of average difficulty, but fair

      3: student should find the question difficult or challenging.

The expert opinion deviation from the item difficulty was measured by the amount of deviation of the expert opinion (average of eight expert opinions) from the best fit line fitted to the regression between the item difficulties and the expert opinion measures over all the items. As with confidence deviation, the amount of deviation between the observed expert opinion measures (y_i) and the expected expert opinion values $(\hat{y}_i)$ (which we will refer to as expected performance) on the students' actual performance in that item, is represented by the vertical distance from the best fit line for each item, as shown in Figure 5.2. Thus, for the point (x_i, y_i) which lies far from the best fit line, the actual expert opinion on the difficulty level differs greatly from the expected difficulty level which means that for this item i, the experts as a group misjudged the difficulty of the question as per student performance.

Figure 5.2: Illustration of expert opinion deviation from the best fit line between item difficulty and expert opinion.
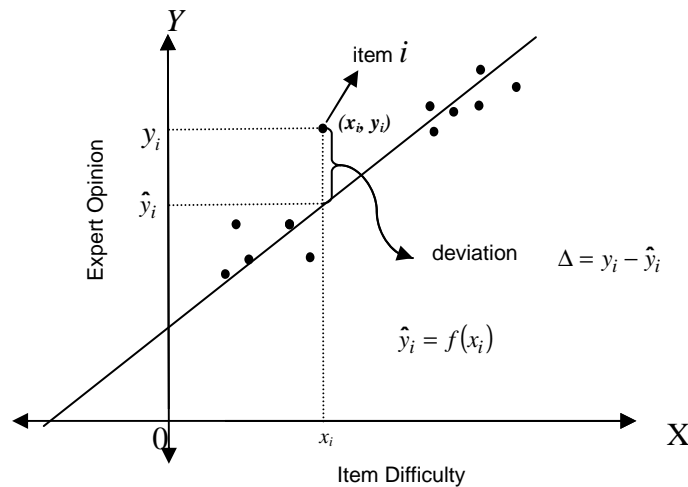
Figures 5.1 and 5.2 show that the larger the deviation of the predicted value from the observed value, the further the observed value is from the regression line and the worse the situation is in terms of an indication of quality.

5.3.2 Defining the Quality Index (QI)

The three measuring criteria discussed in section 5.3 were considered together as an indication of the quality of an item. In future, this will be referred to as the *Quality Index (QI)*. In this study, we do not enter into a debate which of the three measuring criteria are more important. In the proposed QI model, all three criteria are considered to be equally important in their contribution to the overall quality of a question. In order to graphically represent the qualities of a question, 3-axes radar plots were constructed, where each of the three measuring criteria is represented as one of the three arms of the radar plot. In order to compare and plot all three criteria, the measurement direction for the three axes was standardised between 0 and 1. This was done using the transformation formula,

$y = \dfrac{x-a}{b-a}$, where the original scale interval $[a,b]$ is now transformed into the required scale $[0,1]$ on each axis, with a being the minimum value and b the maximum value for each of the respective three criteria. In order to spread out the values between 0 and 1 on each axis, a further normalisation of the data on the interval $[0,1]$ was done.

In Figure 5.3, a visual representation of the three axes of the QI is given. The axes were assigned on an ad hoc basis, with adapted discrimination of the first axis, adapted confidence deviation on the second axis and adapted expert opinion deviation of the third axis. On each axis, the value of 0.5 is indicated as a cut-off point between weak and strong and between small and large. The closer the values are to 0, the more successful the criteria are considered to be in their contribution to the quality of a question.

Figure 5.3: Visual representation of the three axes of the QI.
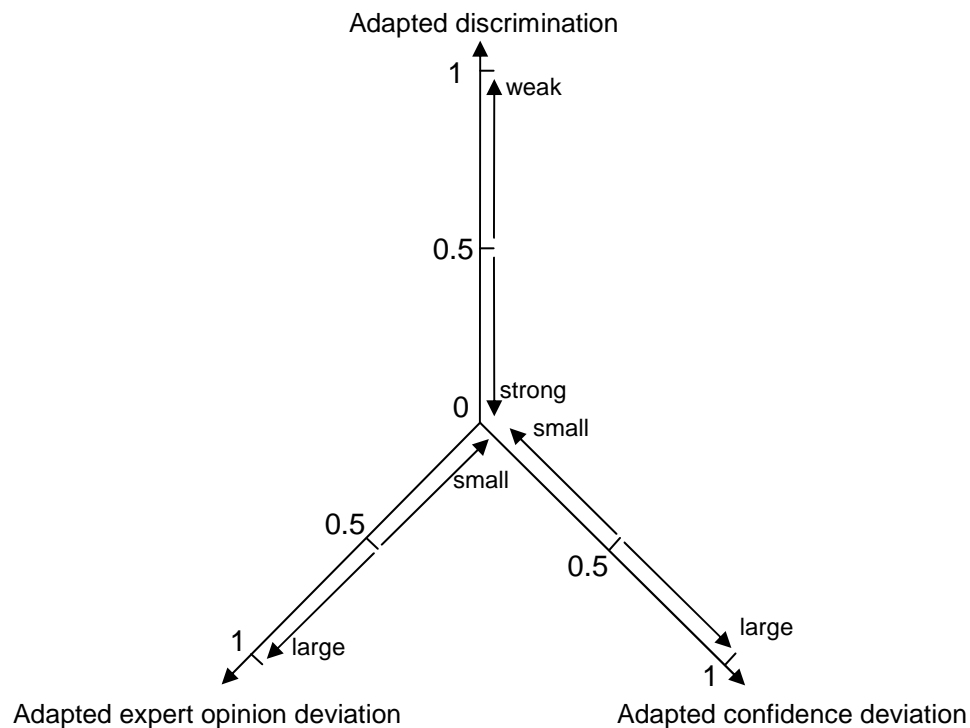
Figure 5.4 depicts an example of a radar plot.

Figure 5.4: Quality Index for PRQ
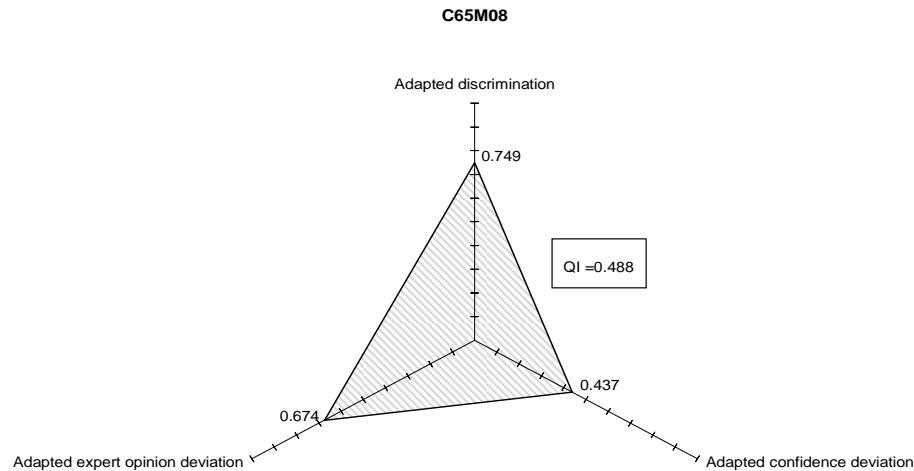
The Quality Index (QI) is defined to be the *area* of the radar plot. The area formula is:

$$QI = \frac{\sqrt{3}}{4}\left[(Discr \times Conf\ dev) + (Conf\ dev \times EO\ dev) + (EO\ dev \times Discr)\right]$$ where

Discr = Adapted discrimination;

Conf dev = Adapted confidence deviation;

EO dev = Adapted expert opinion deviation

The QI combines all three measuring criteria and can now be used to compare the quality of the PRQs with the CRQs within each assessment component. For the proposed model, the smaller the area of the radar plot, i.e. the closer the QI value is to zero, the better the quality of the question. A sample group of test items was used, in total 207 items, of which 94 of the items were PRQs and 113 were CRQs. The median QI value for all the test items was calculated and this value of 0.282 was used as a cut-off value to define the quality of an item as follows:

Good quality　　　　　:　　　QI < 0.282

Poor quality　　　　　:　　　QI ≥ 0.282

If the QI of an item is close to 0.282, the item quality is considered to be moderately good/poor.

In the following two figures an example of a small QI, which constitutes a good quality item, versus an example of a large QI constituting an item of lesser quality are presented.

In Figures 5.5 and 5.6 an example of a small QI, which constitutes a good quality item, versus an example of a large QI constituting an item of lower quality are represented for comparison purposes.

Figure 5.5: A good quality item.

Show that $\displaystyle\sum_{r=0}^{n} \binom{n}{r} (-1)^r = 0$.

CRQ, Algebra, June 2005, Q1b.

A651b (Good quality)

Adapted discrimination

0.213

QI =0.079

0.240

0.291

Adapted expert opinion deviation

Adapted confidence deviation
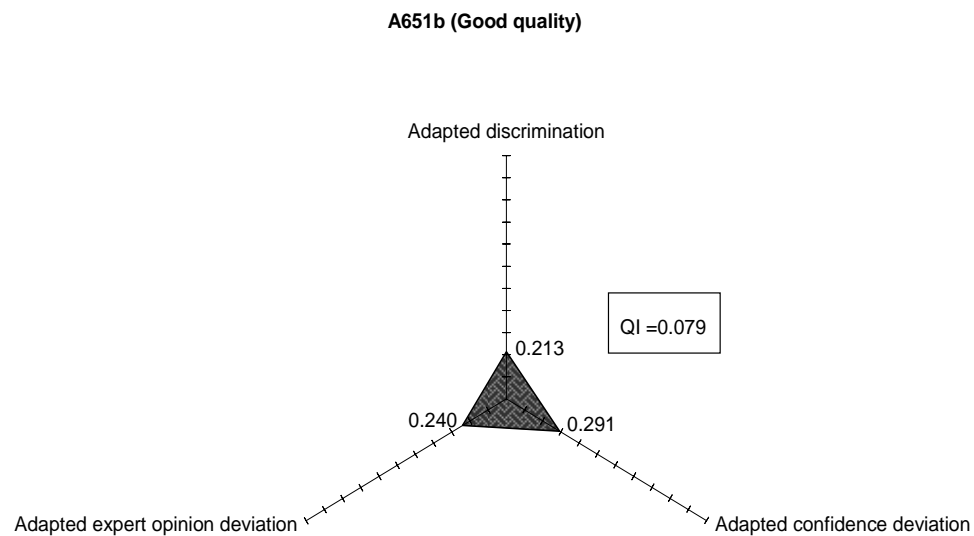
Figure 5.6: A poor quality item.

Consider the following theorem:

Theorem: If a function f is continuous on the closed interval $[a,b]$ and F is an antiderivative

of f on $[a,b]$ then $\int_{a}^{b} f(x)dx = F(b) - F(a)$.

Consider the proof to this theorem:

Proof: Divide the interval $[a,b]$ into n sub-intervals by the points

$a = x_0 < x_1 < ... < x_{n-1} < x_n = b$.

Show that $F(b) - F(a) = \sum_{i=1}^{n} [F(x_i) - F(x_{i-1})]$.

CRQ, Calculus, September 2005, Q3b.

C953b (Poor quality)
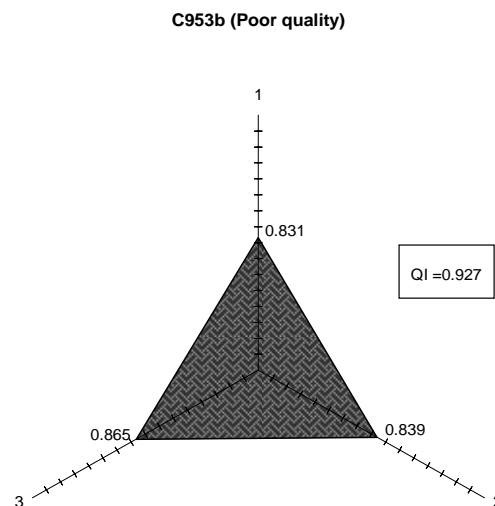
5.3.3 Visualising the difficulty level

Difficulty level is an important parameter, but does not contribute to classifying a question as good or not. Both easy questions and difficult questions can be classified as good.

In this study, the range of difficulty levels over the 207 test items was calculated to be a value of 0.12 using the maximum difficulty value of 4.56 and the

167

minimum difficulty value of -5.56. The standard deviation for this range was calculated to be a value of 1.59. Using these parameters, the distribution of the difficulty levels was investigated by creating a histogram with six intervals of difficulty of 1.5 logits each, as indicated in Figure 5.7.
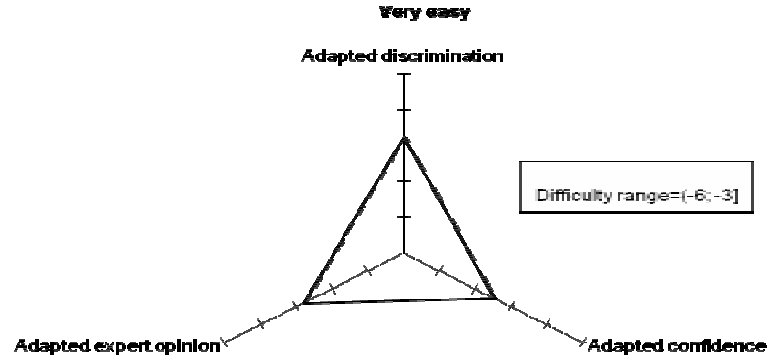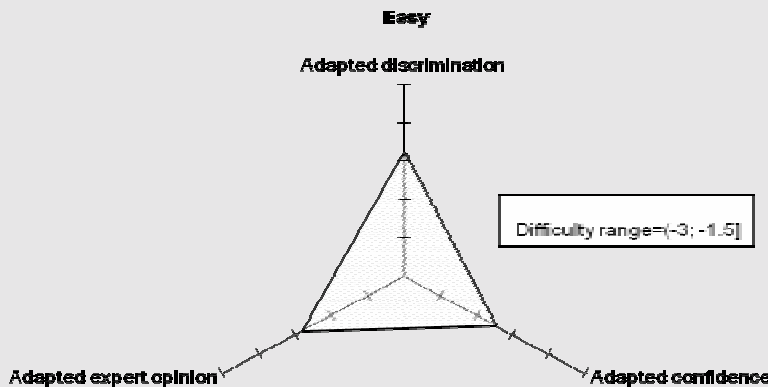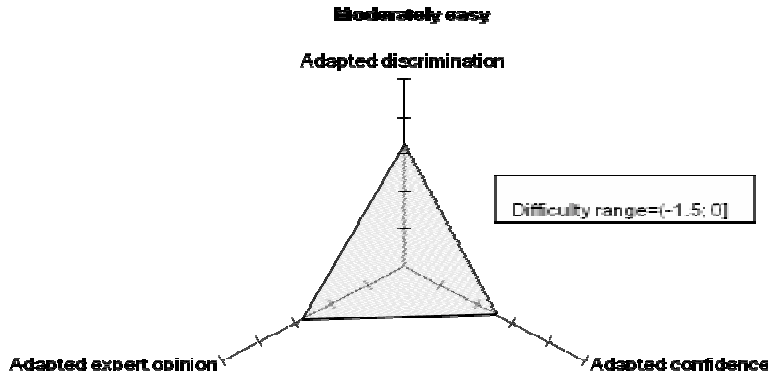
Figure 5.7: Distribution of six difficulty levels.
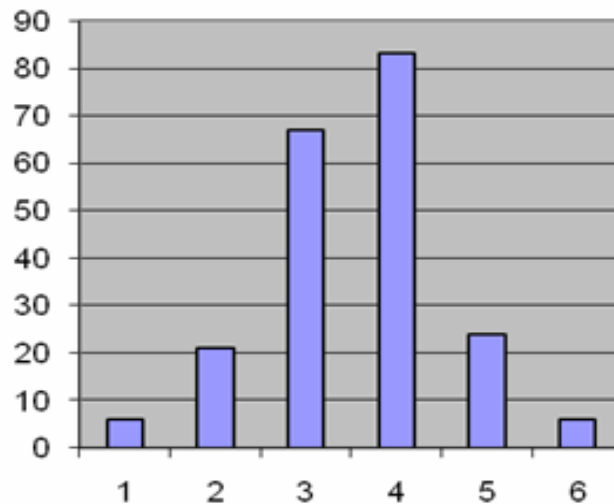
For each of the six intervals, a corresponding shading of the radar chart was chosen to represent the six difficulty levels: very easy; easy; moderately easy; moderately difficult; difficult; very difficult.

Table 5.4 represents the classification and shading of the difficulty intervals. The greater the level of difficulty, the darker the shading of the radar plot, i.e. the intensity of the shading increases from white for the very easy items , through increasing shades of grey to black for the very difficult items. For example, in Figures 5.5 and Figures 5.6 the dark grey shading of the radar plots represents a difficult item. So Figure 5.5 visually represents a difficult, good quality item and Figure 5.6 represents a difficult, poor quality item.

Table 5.4: Classification of difficulty intervals.

Interval	Degree of difficulty	Shading
(-6; -3]	Very easy	Very easy Adapted discrimination Difficulty range=(-6; -3] Adapted expert opinion — Adapted confidence
(-3; -1.5]	Easy	Easy Adapted discrimination Difficulty range=(-3; -1.5] Adapted expert opinion — Adapted confidence
(-1.5; 0]	Moderately easy	Moderately easy Adapted discrimination Difficulty range=(-1.5; 0] Adapted expert opinion — Adapted confidence

169

Interval	Degree of difficulty	Shading
(0; 1.5]	Moderately difficult	
(1.5; 3]	Difficult	
(3; 6]	Very difficult	

In Chapter 6, in the research findings, a quantitative data analysis will be presented. In this chapter, I report on and compare good quality items and poor quality items, both PRQs and CRQs, within each of the seven mathematics assessment components in terms of the Quality Index developed in section 5.3.2.

CHAPTER 6:   RESEARCH FINDINGS

6.1   QUANTITATIVE DATA ANALYSIS

In this chapter on the research findings, an analysis of good quality items and poor quality items, both PRQs and CRQs, in terms of the Quality Index developed in section 5.3.2, within each of the seven mathematics assessment components, will be presented.

6.1.1   Methodology

Stage 1

The traditional statistical analysis of data, supplied by the Computer Network Services (CNS) Division of the University of the Witwatersrand, include the Performance Index, Discrimination Index and Easiness/Difficulty factor per question for all tests (PRQ and CRQ) during the period of study, July 2004 to July 2006.

Raw data, including students' responses to test items and confidence of responses, was obtained from the Computer Network Services (CNS) Division of the University of the Witwatersrand. Spreadsheets were constructed using a 'Mathematica' programme developed by a statistician from the School of Statistics at the University of the Witwatersrand. The following information was captured on every spreadsheet per test:

- students' responses to all test items, both PRQ and CRQ
- students' confidence of responses per test item, both PRQ and CRQ.

The correct answers and mathematics assessment components per test item were also recorded for reference purposes. Student numbers were not recorded on every spreadsheet. In constructing these spreadsheets, records were excluded if:

(i) the student had failed to provide an answer; or

(ii) the student had failed to provide a confidence of response; or

(iii) the student had filled in the MCQ card incorrectly.

It should be noted that in most cases the excluded records were due to (ii) above. The proportion of all the records excluded in this manner ranged between 7,2% and 8,9% across the tests. All subsequent calculations were performed on this filtered data.

For PRQs and CRQs, the Performance Index (PI) per question was equal to the proportion of (filtered) respondents who obtained the correct answer. It should be noted that the "easiness/difficulty" statistic provided on the CNS printouts is equal to the Performance Index i.e. Performance Index = Difficulty Index. An overall Confidence Index (per assessment component) was calculated by averaging the CIs per question for all questions in that assessment component. An overall Performance Index or Difficulty Index (per assessment component) was calculated in a similar manner by averaging the PIs per question for all questions in that assessment component.

Stability (test- retest) 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 cohort of students (Mathematics I Major) in each of the 3 years, 2004, 2005 and 2006 respectively. Internal consistency was achieved by correlating and equating the items in each test to each other, as described under test item calibration in section 6.2.1.

Stage 2

The Rasch model (Rasch, 1960), as discussed in section 3.4.1, was used to evaluate both the attitudinal data (confidence levels) as well as test data. The Winsteps (Linacre & Wright, 1999) Rasch analysis programme was utilised by a data analyst from the University of Pretoria for the quantitative data analysis in this research study. In particular the WINSTEPS® Version 3.55.0 was used to

analyse the data in this study. SAS Version 9 and Microsoft EXCEL 2003 were also used in calculating totals and means.

The Winsteps software, developed by John M Linacre in 2005, constructs Rasch measures from simple rectangular data sets, usually of persons and items. Item types that can be combined in one analysis include dichotomous, multiple choice and partial credit items. Paired comparisons and rank-order data can also be analysed. Missing data is no problem. Winsteps is designed as a tool that facilitates exploration and communication. The structure of the items and persons can be examined in depth. Unexpected data points are identified and reported in numerous ways. Powerful diagnosis of multidimensionality through principal components analysis of residuals detects and quantified substructures in the data. The working of rating scales can be examined thoroughly, and rating scales can be recoded and items regrouped to share rating scales as desired (Linacre, 2002). Measures can be fixed (anchored) at pre-set values (Linacre, 2005).

In order to prepare the data in an ASCII format to import into Winsteps, SAS was used to create ASCII files with a specific layout. Control files were prepared in Winsteps for each part of each test, i.e. the PRQ part, the CRQ part as well as the confidence index part. This was done as the different Rasch models, discussed in section 3.4.1.3, were applicable to the different types of data. These parts of the tests were first analysed separately to check for model "fit". Such "fit" statistics help detect possible idiosyncratic behaviour on the part of respondents and test items. Those respondents who exhibited "misfit" were first investigated for coding errors, and then their raw hard-copy responses were reviewed for evidence of non-attention to the test. Such individuals might be ones who are haphazardly circling responses or those who are guessing and/or miscoding.

Winsteps provides ways of diagnosing problems in the analysis. In the first place the point measure values were considered. Where items exhibited negative point measure values, these items were scrutinised for errors such as an

incorrect key and corrected. If the point measure stayed negative, the item was removed from the analysis. Subsequently, the output tables for person ability and item difficulty were checked for misfitting entries. Person ability tables were considered first.

Misfit

Some explanation in terms of misfitting items or students is in order. One would expect that a student of medium mathematical ability would be able to respond correctly to easier items in the test and incorrectly to the difficult items in a test. Where the item difficulty matches the ability of the student, one would expect the student to answer some of these items correct and some incorrectly. If an item's difficulty corresponds exactly to the student's ability, the probability of success of the student on that item is 0.5, in other words, success or failure is expected equally. The Rasch model assumes this pattern of responses, and the Infit and Outfit mean-square statistics are 1.0. If for example, a student would guess the answer to a difficult item correctly (one that the student should really get wrong) the Outfit statistic would be much larger than 1.0 because it is sensitive to outliers.

The approach used in the analyses of this study's data was that items and persons were accepted as not misfitting when Infit mean-square statistics was from 0.5 to 1.5. Where the values were less than 0.5, too much predictability or overfit was experienced and when the value exceeded 1.5, too much noise was present in the data or a situation of underfit existed. The Infit statistics were considered first, and then the Outfit statistics.

Mean-square statistics indicate the size of the misfit, but the "significance" of the improbability of the misfit is important.

Misfitting persons were deleted, and the analysis was repeated. Another round of misfitting persons were removed from the analysis. Only then were the fit

statistics of the items considered. If an item proved to be problematic in terms of the fit statistics, the item was also removed from any subsequent analysis.

The same procedure was followed to explore the misfitting persons and items in terms of the CRQs and the confidence index.

For the PRQs, the dichotomous Rasch model applies:

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

In the confidence index, the same categories were available throughout and were thus analysed according to the Rasch-Andrich rating scale model:

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

CRQs were analysed through the application of the Partial Credit model:

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

These various Rasch models have already been discussed in more detail in section 3.4.1.

Test item calibration

Through the application of the Rasch family of models it is also possible to put the measures of different tests onto the same scale if certain assumptions are made. The tests can be linked either through common items on the tests or through common students writing the tests. A challenge in terms of the data faced the researcher. Although, as mentioned previously, it was known that the same cohort of students wrote the same tests in a calendar year, the student identification numbers were not available on all the data sets and therefore no linking could take place on a one-to-one basis. The strong assumption was then made that the subject matter of the different tests were distinct and that the tests

could therefore be regarded as independent. In other words, it was assumed that because the subject matter was distinct, students' ability did not improve progressively throughout the year. This assumption led to the decision that all the data could be calibrated together, anchoring the items that were common over the three years. In this way, the item difficulties and the student measures were on the same scale and were deemed directly comparable.

Fit statistics were again considered and if in the combined calibration of items any misfitting items were identified, they were excluded from the analysis. A small number of items misfitted, and this is not to be unexpected in such a large data set.

The same procedure was followed in terms of the CRQs. In order to place the measures of the PRQs and the CRQs on the same scale, a combined calibration of these items was also executed. Another challenge presented itself. At first, when the PRQs and the CRQs were calibrated together, the whole set of CRQs misfitted. It was then decided to recode the partial credit items into dichotomous items in the following way: If a student scored less than half the marks, the student was awarded a 0 for that specific item; if the student scored half or more of the marks on an item, the student was awarded a 1 for the item. The CRQs were therefore eventually analysed through the same model as the PRQs i.e. the dichotomous Rasch model, and the combined calibration of items then produced a set of items that mostly fitted the Rasch model.

Confidence level item calibration

A similar process was followed to determine the item difficulties of the confidence levels. The item difficulty for a rating scale is defined as the point where the top and bottom categories are equally probable (Linacre, 2005).

6.2   DATA DESCRIPTION

Response data from 14 different mathematics tests written between August 2004 and June 2006 were available. Table 6.1 is a representation of the tests written, the number of provided response items (PRQs) per test, the number of constructed response items (CRQs) and the number of students per test. The same cohort of students (Mathematics I Major) wrote the tests in each of the three years, 2004, 2005 and 2006 respectively.

Table 6.1: Characteristics of tests written.

Year	Month	Number of PRQs	Number of CRQs	Number of students
2004	August	10	0	457
2005	March	8	0	410
2005	April Tutorial A	8	0	263
2005	April Tutorial B	8	0	126
2005	May	8	0	403
2005	June	12	17	414
2005	August	10	0	389
2005	September	8	17	387
2005	November	15	18	385
2006	March	8	15	352
2006	April Tutorial A	8	0	245
2006	April Tutorial B	8	0	105
2006	May	8	14	359
2006	June	12	24	348

Out of a total of 221 PRQ and CRQ items, seven items were discarded because their fit statistics indicated that they did not fit the model. Table 6.2 included in the Appendix A5, presents these items with their fit statistics. Another seven items (I115M09 – I115M15) were discarded because the actual items were not available. Finally, 207 items were included in the analyses. The Rasch statistics

for all 207 test items analysed are included in Appendix A6. Confidence level items Rasch statistics are included in Appendix A7.

6.3 COMPONENT ANALYSIS

Examples of questions in the different mathematics assessment components are now presented. Within each of the seven assessment components, both PRQs and CRQs, ranging from easy to difficult, and of good and poor quality are presented. For each item, the question is followed by a radar plot and a table summarising the quality parameters of the test item i.e. item difficulty; discrimination; confidence index; expert opinion and the final quality index, as discussed in the theoretical framework in Chapter 5. Each of the axes of the radar plots are labelled with the corresponding values for discrimination, confidence index and expert opinion. The Quality Index (QI) is displayed alongside the radar plot. The shading of the radar plot corresponds to one of the six item difficulty levels as classified in Table 5.4. The comments briefly summarise the difficulty level, the three measuring criteria as developed in the theoretical framework and the overall quality of the item.

1.   Technical component

A651(a)

Find the constant term in $\left(-x^2 + \dfrac{1}{x} \right)^{12}$
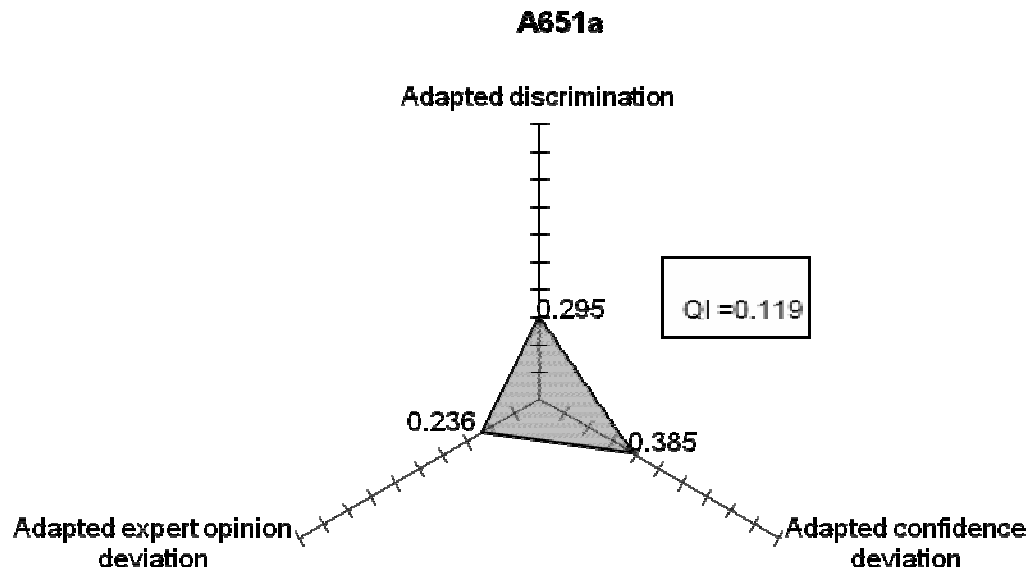
A651a

A651a		Comment
Assessment Component	Technical	
PRQ/CRQ	CRQ	
Item Difficulty	1.10	Moderately difficult
Discrimination	0.295	Discriminates well
Confidence Index	0.385	Small deviation from expected confidence level
Expert Opinion	0.236	Small deviation from expected performance
Quality Index	0.119	Good quality CRQ (excellent)

A652(a)

Write $-2\cos x + 2\sqrt{3}\sin x$ in the form $R\cos(x-\theta)$

A652a

Adapted discrimination

0.501

QI =0.273

0.318

0.574

Adapted expert opinion deviation

Adapted confidence deviation
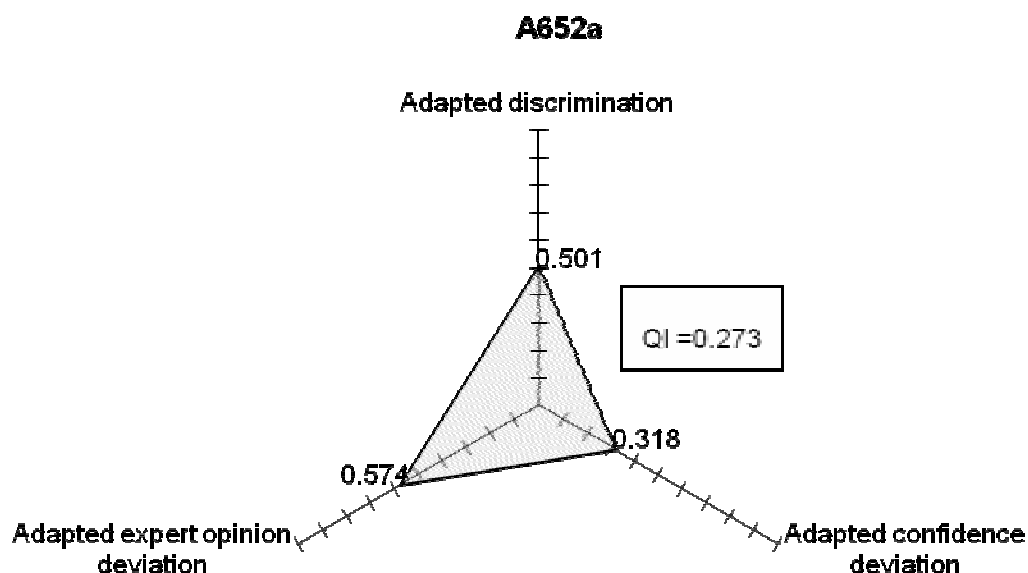
A652a		Comment
Assessment Component	Technical	
PRQ/CRQ	CRQ	
Item Difficulty	-0.33	Moderately easy
Discrimination	0.501	Discriminates fairly well
Confidence Index	0.318	Small deviation from expected confidence level
Expert Opinion	0.574	Large deviation from expected performance
Quality Index	0.273	Good quality CRQ (moderate)

C115MO7

The limit of the sequence $\left\{\dfrac{1}{n!}\left(-5+(-1)^n\right)\right\}$ is

    A. -5

    B. 1

    C. 0

    D. the sequence diverges

<div align="right">PRQ, Calculus, November 2005, Q7</div>

C115M07
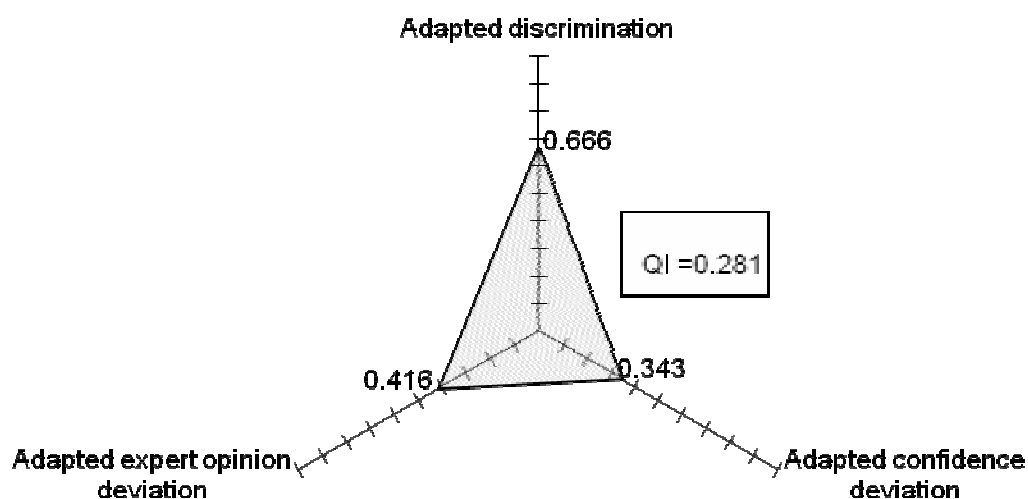
C115M07		Comment
Assessment Component	Technical	
PRQ/CRQ	PRQ	
Item Difficulty	-1.12	Moderately easy
Discrimination	0.666	Does not discriminate well
Confidence Index	0.343	Small deviation from expected confidence level
Expert Opinion	0.416	Small deviation from expected performance
Quality Index	0.281	Good quality PRQ (moderate)

A1155bii

Let $A = \begin{pmatrix} 1 & 1 & 1 \\ ab & bc & ca \\ a+b & b+c & c+a \end{pmatrix}$

For what value(s) of a, b, c does A^{-1} exist?
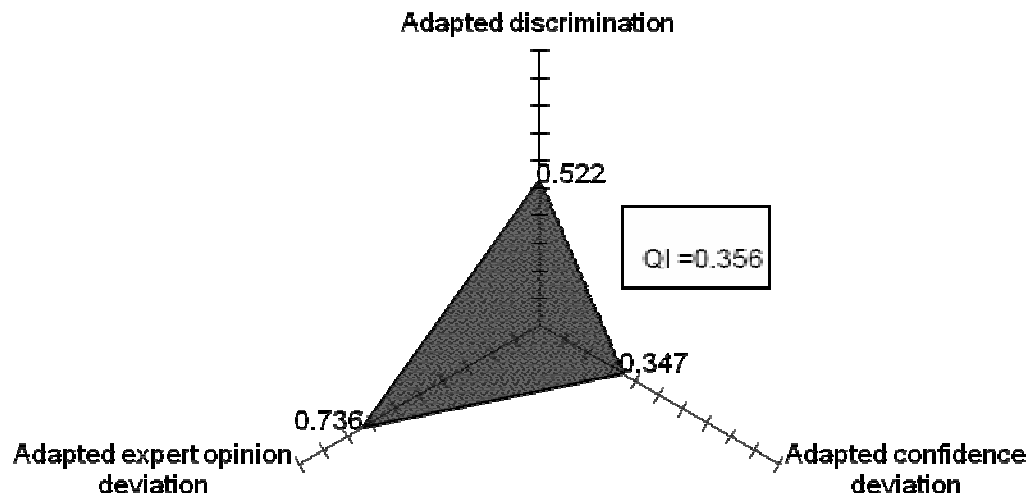
A1155bii

A1155bii		Comment
Assessment Component	Technical	
PRQ/CRQ	CRQ	
Item Difficulty	2.23	Difficult
Discrimination	0.522	Discriminates poorly
Confidence Index	0.347	Small deviation from expected confidence level
Expert Opinion	0.736	Large deviation from expected performance
Quality Index	0.356	Poor quality CRQ

A661.1

$P(n) = n^3 + (n+1)^3 + (n+2)^3$ is divisible by 9

Show that the statement is true for $n = 2$
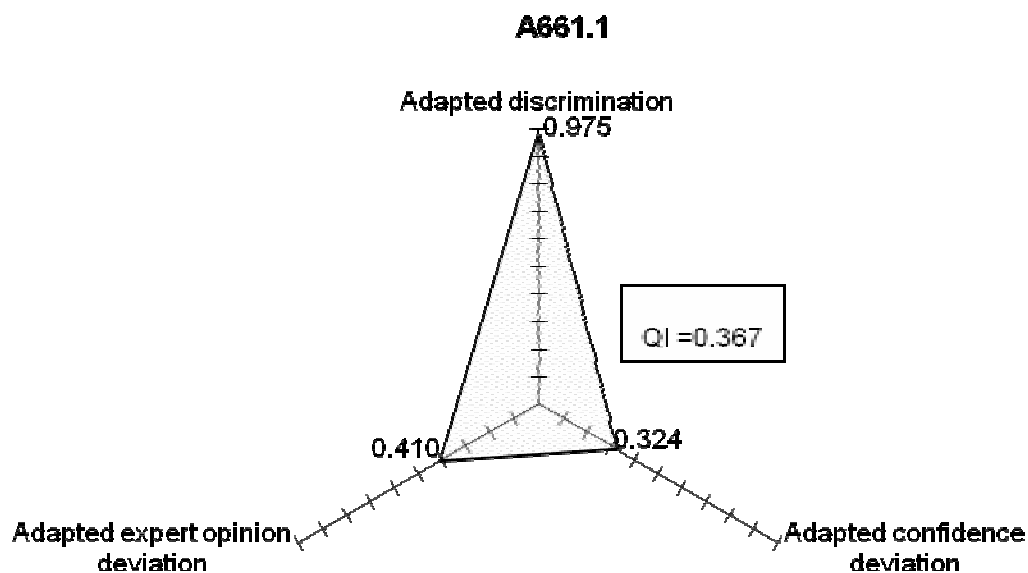
A661.1

A661.1		Comment
Assessment Component	Technical	
PRQ/CRQ	CRQ	
Item Difficulty	-2.35	Easy
Discrimination	0.975	Discriminates weakly
Confidence Index	0.324	Small deviation from expected confidence level
Expert Opinion	0.410	Small deviation from expected performance
Quality Index	0.367	Poor quality CRQ

A56MO2

The exact value of $\arctan(\tan(5\pi/3))$ is

A. $5\pi/3$

B. $-5\pi/3$

C. $-\pi/3$

D. $\pi/3$

E. $2\pi/3$

<div align="right">PRQ, Algebra, May 2006, Q2</div>

A56M02
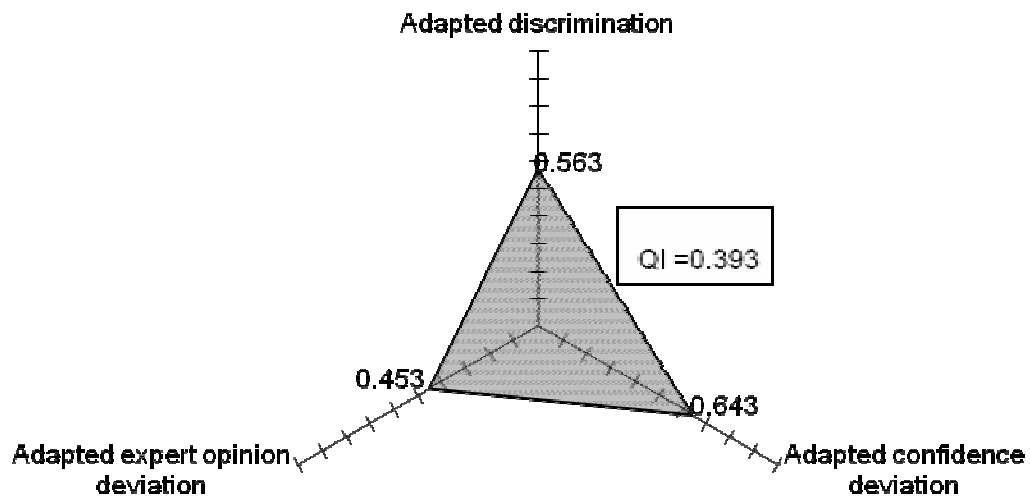
A56M02		**Comment**
Assessment Component	Technical	
PRQ/CRQ	PRQ	
Item Difficulty	0.77	Moderately difficult
Discrimination	0.563	Weak discrimination
Confidence Index	0.643	Large deviation from expected confidence level
Expert Opinion	0.453	Small deviation from expected performance
Quality Index	0.393	Poor quality PRQ

C65M08

If $\int\limits_{3}^{5} g(x)dx = 5$ and $\int\limits_{3}^{5} h(x)dx = -1$, then $\int\limits_{3}^{5} \left(2\,g(x) - 5h(x)\right) dx =$

   A.   5

   B.   15

   C.   7

   D.   0

   E.   -27

PRQ, Calculus, June 2005, Q8

C65M08
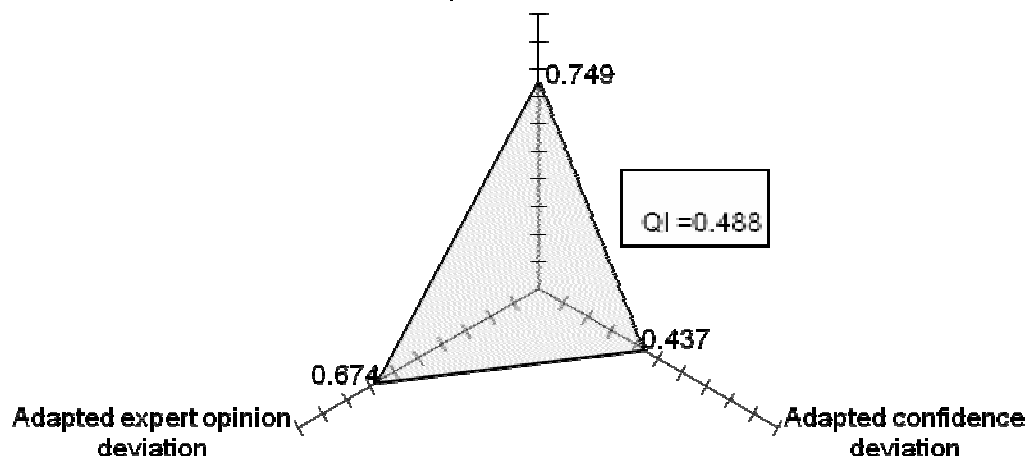
C65M08		Comment
Assessment Component	Technical	
PRQ/CRQ	PRQ	
Item Difficulty	-1.04	Moderately easy
Discrimination	0.749	Weak discrimination
Confidence Index	0.437	Small deviation from expected confidence level
Expert Opinion	0.674	Large deviation from expected performance
Quality Index	0.488	Poor quality PRQ

2. Disciplinary component

A35M08

Let a, b and c be real numbers. Which of the following is the correct statement?

A. $a < b \Rightarrow a + b > b + c.$

B. $a > b \Rightarrow ac > bc.$

C. $|x| > a \Leftrightarrow -a < x < a.$

D. $\sqrt{c^2} = c.$

E. $0 < a < b \Rightarrow \dfrac{1}{b} < \dfrac{1}{a}.$

PRQ, Algebra, March 2005, Q8

A35M08

Adapted discrimination

QI =0.165

0.069

0.355

0.842

Adapted expert opinion
deviation

Adapted confidence
deviation
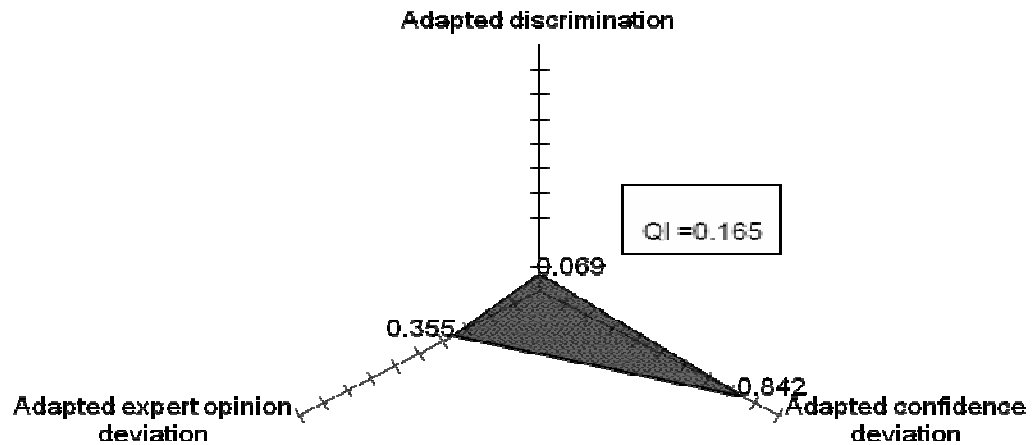
A35M08		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	PRQ	
Item Difficulty	2.25	Difficult
Discrimination	0.069	Discriminates very well
Confidence Index	0.842	Large deviation from expected confidence level
Expert Opinion	0.355	Small deviation from expected performance
Quality Index	0.165	Good quality PRQ

C363b

Prove, using the Intermediate Value Theorem, that there is a number exactly 1 more than its cube.

C363b
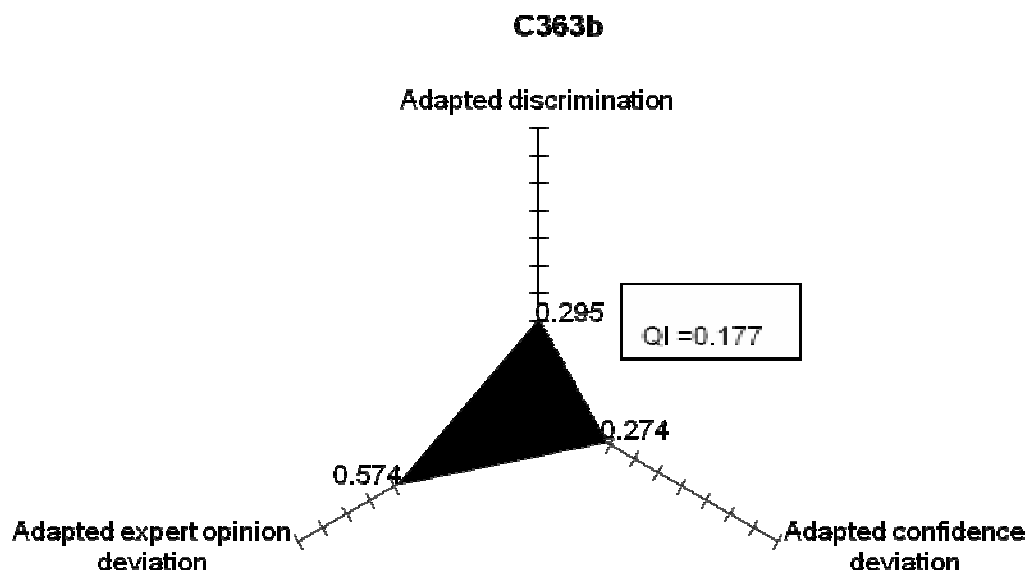
C363b		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	CRQ	
Item Difficulty	3.94	Very difficult
Discrimination	0.295	Discriminates well
Confidence Index	0.274	Small deviation from expected confidence level
Expert Opinion	0.574	Large deviation from expected performance
Quality Index	0.177	Good quality CRQ

C561a(i)

A bacterial colony is estimated to have a population of $P(t) = \dfrac{24t + 10}{t^2 + 1}$ million, t hours after the

introduction of a toxin.

At what rate is the population changing 1 hour after the toxin is introduced?

C561ai
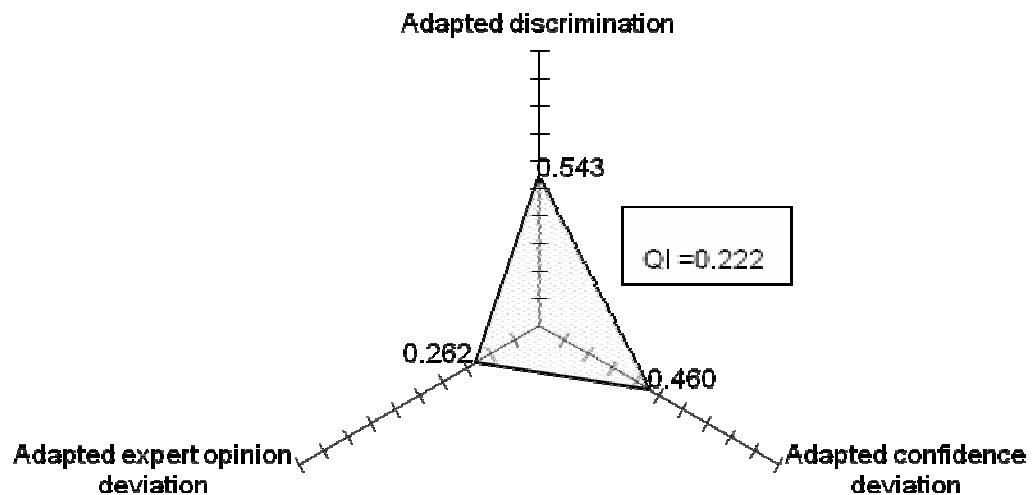
C561ai		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	CRQ	
Item Difficulty	-2.63	Easy
Discrimination	0.543	Discriminates fairly well
Confidence Index	0.460	Small deviation from expected confidence level
Expert Opinion	0.262	Small deviation from expected performance
Quality Index	0.222	Good quality CRQ

A55M07

The Cartesian coordinates (x, y) of the point $(r, \theta) = (2\sqrt{3}, \dfrac{3\pi}{4})$ are:

A. $(-\sqrt{6}, -\sqrt{6})$

B. $(-\sqrt{6}, \sqrt{6})$

C. $(\sqrt{6}, -\sqrt{6})$

D. $(-3, 2)$

A55M07

A55M07		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	PRQ	
Item Difficulty	-0.76	Moderately easy
Discrimination	0.790	Does not discriminate well
Confidence Index	0.294	Small deviation from expected confidence level
Expert Opinion	0.290	Small deviation from expected performance
Quality Index	0.236	Good quality PRQ (moderate)

C364b(i)

Let $[\![x]\!]$ be the greatest integer less than or equal to x.

Show that $\lim_{x \to 2} f(x)$ exists if $f(x) = [\![x]\!] + [\![-x]\!]$.

<div align="right">CRQ, Calculus, March 2006, Q4b(i)</div>
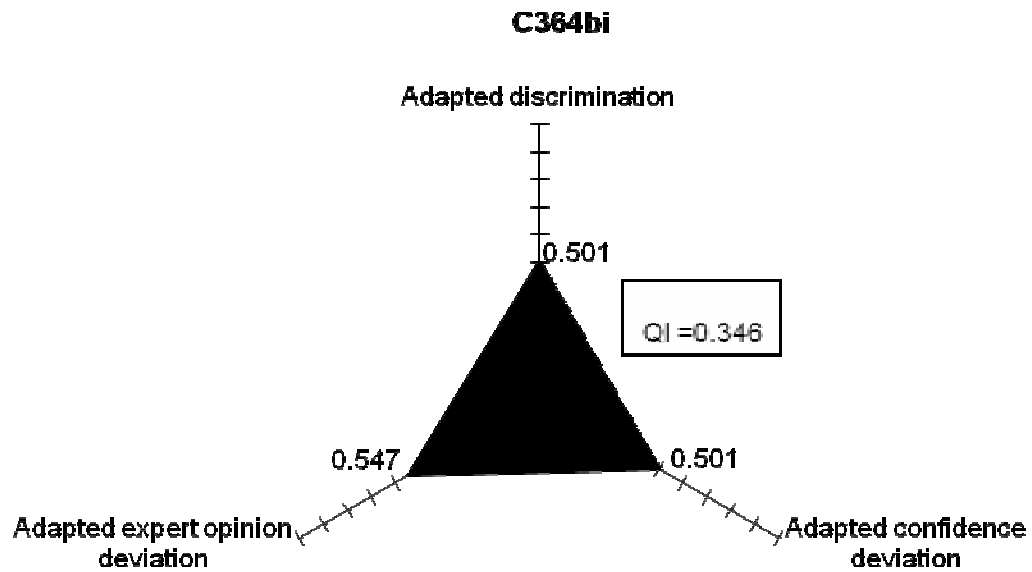
C364bi

C364bi		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	CRQ	
Item Difficulty	4.19	Very difficult
Discrimination	0.501	Discriminates fairly well
Confidence Index	0.501	Average deviation from expected confidence level
Expert Opinion	0.547	Large deviation from expected performance
Quality Index	0.346	Poor quality CRQ

C563a(i)

Consider the following theorem:

Let f be a function that satisfies the following three conditions:

(1) f is continuous on the closed interval $[a,b]$.

(2) f is differentiable on the open interval (a,b).

(3) $f(a) = f(b)$.

Then there exists a number $c \in (a,b)$ such that $f'(c) = 0$.

What is this theorem called?

CRQ, Calculus, May 2006, Q3a(i)

C563ai
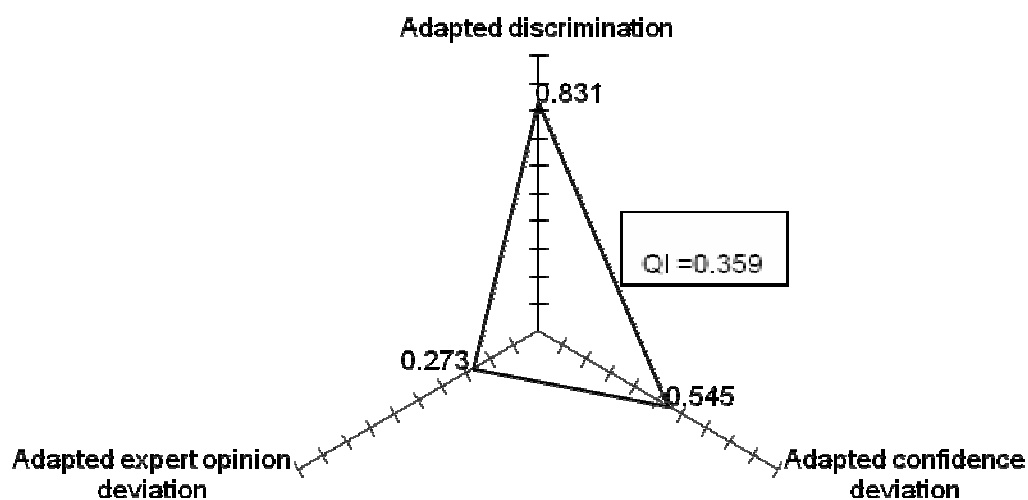
C563ai		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	CRQ	
Item Difficulty	-4.74	Very easy
Discrimination	0.831	Discriminates poorly
Confidence Index	0.545	Large deviation from expected confidence level
Expert Opinion	0.273	Small deviation from expected performance
Quality Index	0.359	Poor quality CRQ

192

C45MB5

If $\displaystyle\lim_{x \to 2} f(x)$ exists, then

A. $f(2)$ is undefined

B. $f(2) = 3$

C. $f(2) = 2$

D. $f(2)$ is unknown

PRQ, Calculus, March 2005, Tut Test 1B, Q5

C45MB5
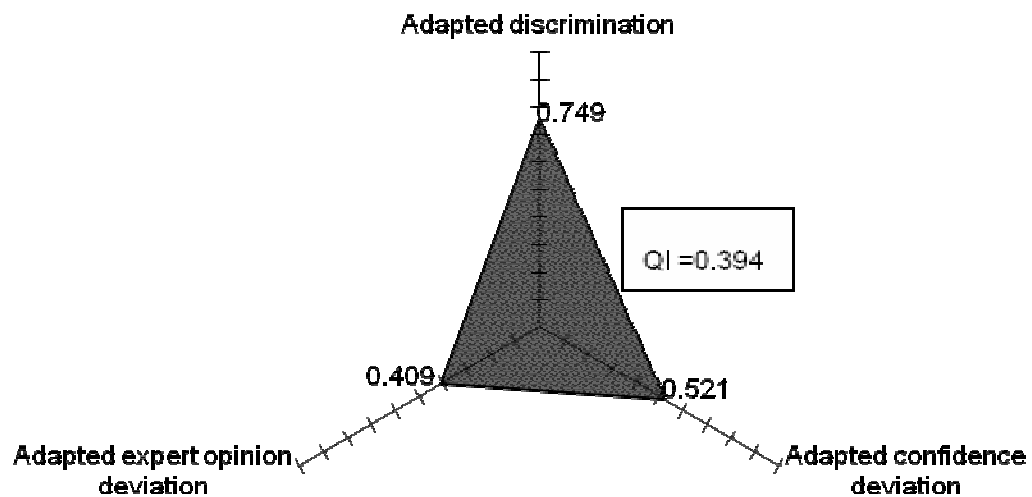
C45MB5		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	PRQ	
Item Difficulty	1.91	Difficult
Discrimination	0.749	Discriminates poorly
Confidence Index	0.521	Large deviation from expected confidence level
Expert Opinion	0.409	Small deviation from expected performance
Quality Index	0.394	Poor quality PRQ

C36M02

Find the following limit: $\lim_{x \to 2} \dfrac{x^2 - 4}{x - 2}$

A. does not exist

B. -2

C. 4

D. 2

E. 1
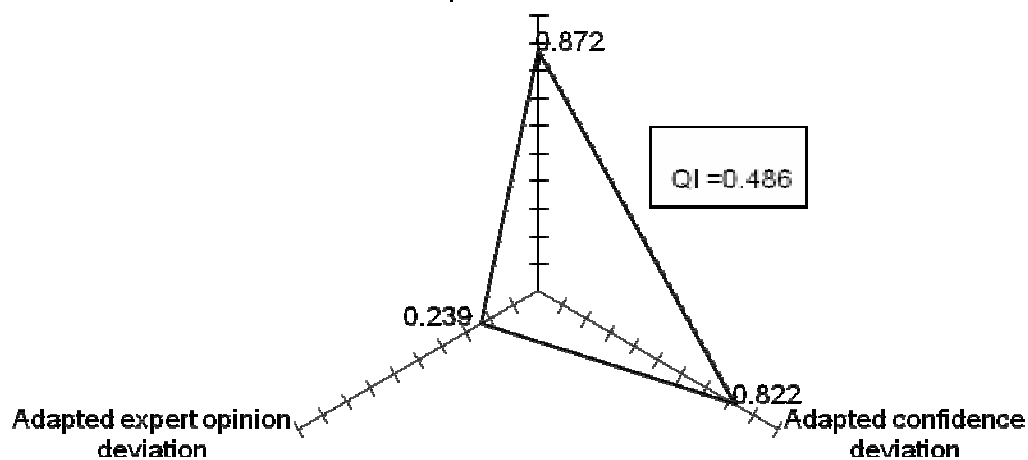
C36M02

C36M02		Comment
Assessment Component	Disciplinary	
PRQ/CRQ	PRQ	
Item Difficulty	-5.05	Very easy
Discrimination	0.872	Discriminates very poorly
Confidence Index	0.822	Very large deviation from expected confidence level
Expert Opinion	0.239	Small deviation from expected performance
Quality Index	0.486	Poor quality PRQ

194

3. Conceptual component

C65M09

Choose the correct statement, given that $\int_0^5 f(x)dx = 9$ and $\int_2^5 f(x)dx = -1$.

A. $\int_0^2 f(x)dx = 10$

B. $\int_2^0 f(x)dx = 10$

C. $\int_5^2 f(x)dx = -1$

D. $\int_0^2 f(x)dx = 8$

E. None of the above
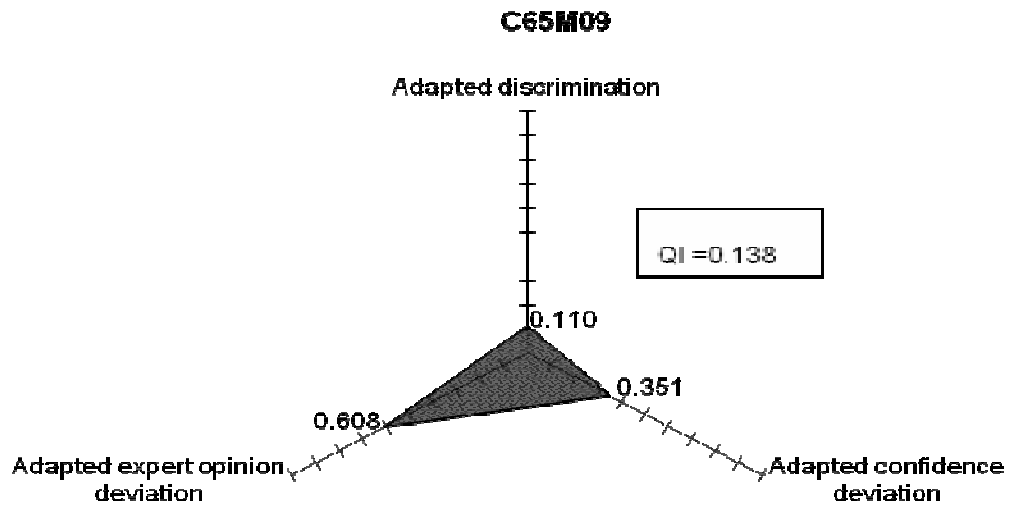
C65M09

C65M09		Comment
Assessment Component	Conceptual	
PRQ/CRQ	PRQ	
Item Difficulty	1.72	Difficult
Discrimination	0.110	Discriminates well
Confidence Index	0.351	Small deviation from expected confidence level
Expert Opinion	0.608	Large deviation from expected performance
Quality Index	0.138	Good quality PRQ

A1152b

Find the equation of the plane which passes through the point $A(2,3,-5)$ and which contains the line $l:(-1,3,-2)+t(-2,1,5)$
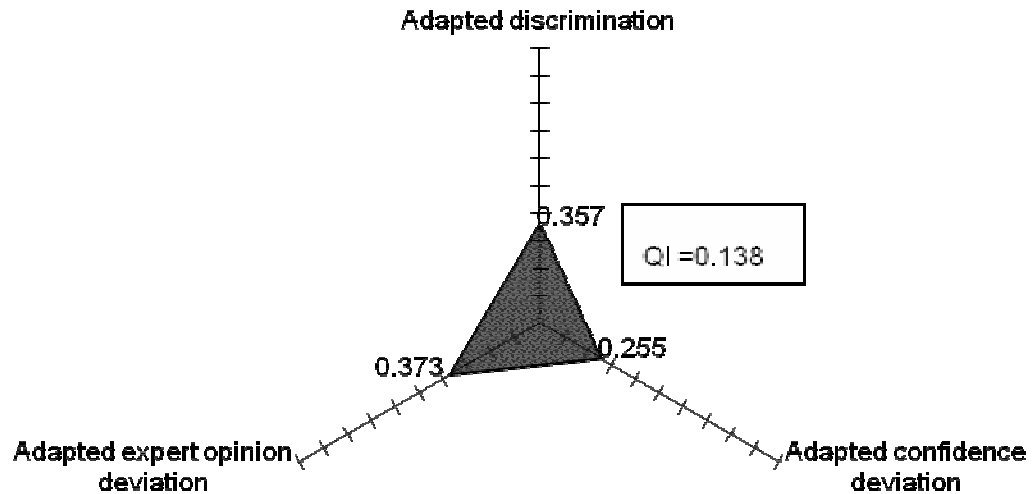
A1152b

A1152b		Comment
Assessment Component	Conceptual	
PRQ/CRQ	CRQ	
Item Difficulty	2.93	Difficult
Discrimination	0.357	Discriminates well
Confidence Index	0.255	Small deviation from expected confidence level
Expert Opinion	0.373	Small deviation from expected performance
Quality Index	0.138	Good quality CRQ (excellent)

C1157a

Find $\int x \cos x \, dx$

C1157a
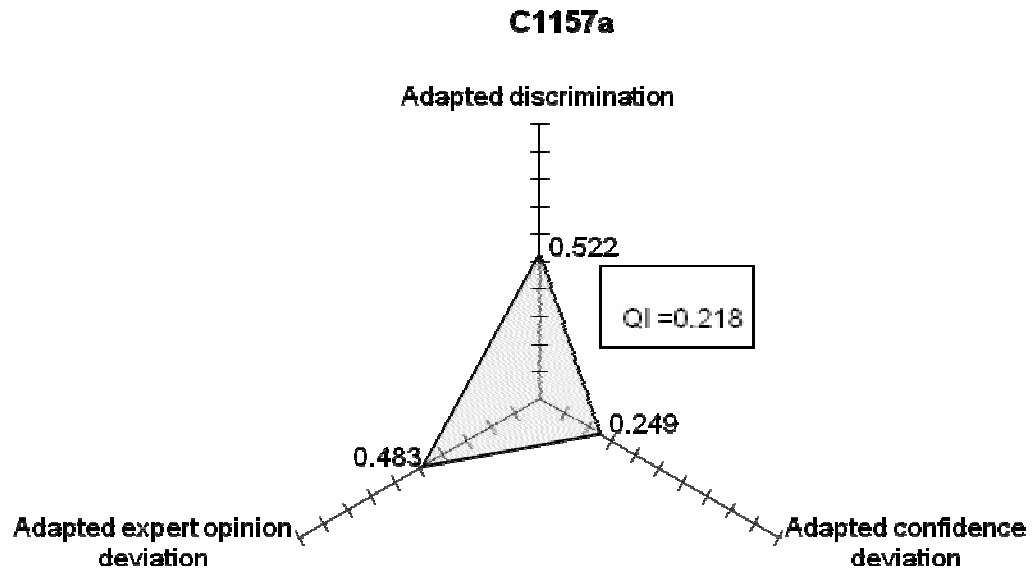
C1157a		Comment
Assessment Component	Conceptual	
PRQ/CRQ	CRQ	
Item Difficulty	-1.45	Moderately easy
Discrimination	0.522	Average discrimination
Confidence Index	0.249	Small deviation from expected confidence level
Expert Opinion	0.483	Small deviation from expected performance
Quality Index	0.218	Good quality CRQ

C45MB8

If $\lim\limits_{x \to a} f(x) = 2$ and $\lim\limits_{x \to a} g(x) = 3$ then $\lim\limits_{x \to a} \dfrac{3f(x) - (g(x))^2}{g(x)} =$

A. $\dfrac{13}{3}$

B. -1

C. $-\dfrac{3}{2}$

D. 1
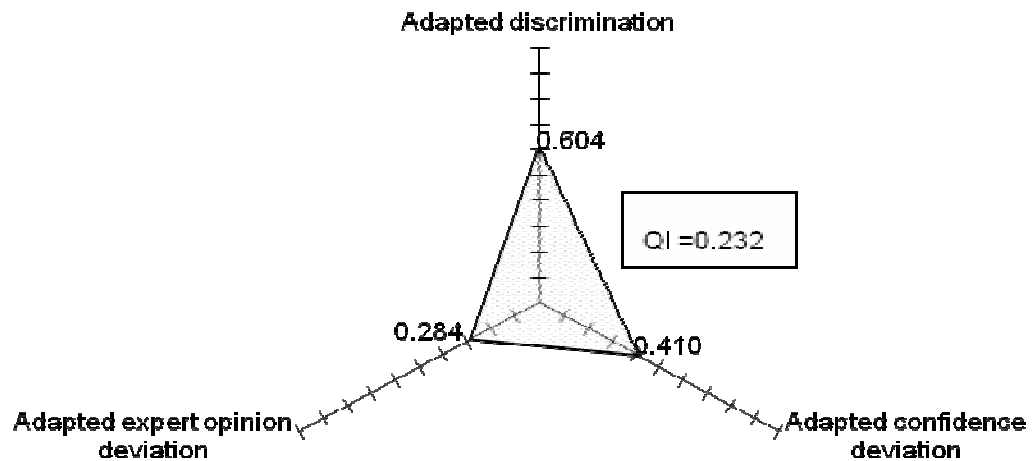
C45MB8

C45MB8		**Comment**
Assessment Component	Conceptual	
PRQ/CRQ	PRQ	
Item Difficulty	-1.94	Easy
Discrimination	0.604	Discriminates poorly
Confidence Index	0.410	Small deviation from expected confidence level
Expert Opinion	0.284	Small deviation from expected performance
Quality Index	0.232	Good quality CRQ (moderate)

A95M02

PQR is a triangle with vertices $P(3,1), Q(5,2)$ and $R(4,3)$. $P\hat{Q}R$ equals

A. $\arccos \dfrac{4}{5}$

B. $\arccos \dfrac{1}{\sqrt{10}}$

C. $\pi - \arccos \dfrac{4}{5} - \arccos \dfrac{1}{\sqrt{10}}$

D. $\arccos \dfrac{-1}{\sqrt{10}}$
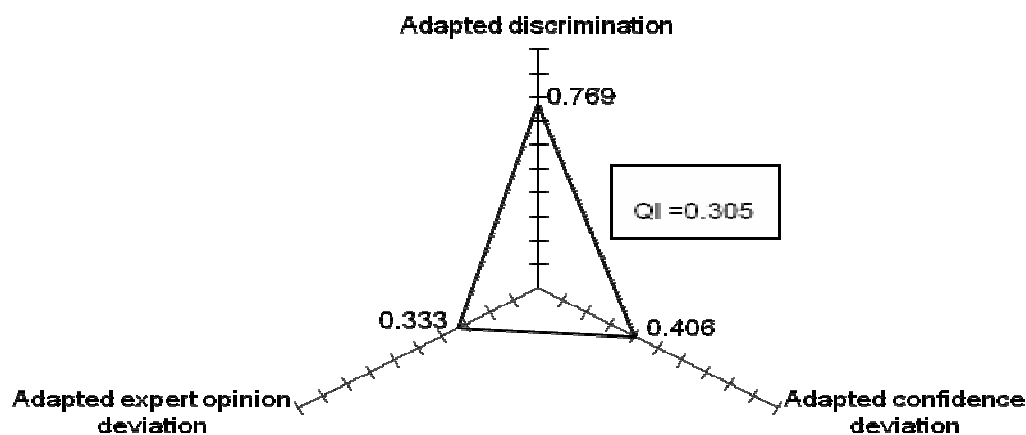
A95M02

A95M02		Comment
Assessment Component	Conceptual	
PRQ/CRQ	PRQ	
Item Difficulty	-3.22	Very easy
Discrimination	0.769	Discriminates poorly
Confidence Index	0.406	Fairly small deviation from expected confidence level
Expert Opinion	0.333	Small deviation from expected performance
Quality Index	0.305	Poor quality PRQ (moderate)

C55M04

The graph below is of the derivative of a function $g(x)$, i.e. the graph of $y = g'(x)$.

y

2

$y = g'(x)$

1

-4   -3   -2   -1   1   2   3   4 x

-1

-2

The critical numbers of $g(x)$ are

A. $-2, 2$

C. $-2, 2, -3, 3$

B. $-3, 3$

D. $-2, -3, 3$

C55M04

Adapted discrimination

0.336

QI = 0.356

0.546

0.723

Adapted expert opinion deviation

Adapted confidence deviation

C55M04		Comment
Assessment Component	Conceptual	
PRQ/CRQ	PRQ	
Item Difficulty	1.50	Moderately difficult
Discrimination	0.336	Discriminates well
Confidence Index	0.723	Large deviation from expected confidence level
Expert Opinion	0.546	Large deviation from expected performance
Quality Index	0.356	Poor quality PRQ

C953a

Consider the following theorem:

Theorem: If a function f is continuous on the closed interval $[a,b]$ and F is an antiderivative

of f on $[a,b]$, then $\int_a^b f(x)dx = F(b) - F(a)$.

What is this theorem called?

C953a

C953a		Comment
Assessment Component	Conceptual	
PRQ/CRQ	CRQ	
Item Difficulty	-5.56	Very easy
Discrimination	1.000	Discriminates very poorly
Confidence Index	0.497	Large deviation from expected confidence level
Expert Opinion	0.434	Fairly small deviation from expected performance
Quality Index	0.562	Poor quality CRQ

C953b

Consider the following theorem:

Theorem: If a function f is continuous on the closed interval $[a,b]$ and F is an antiderivative

of f on $[a,b]$, then $\int_a^b f(x)dx = F(b) - F(a)$.

Consider the proof of this theorem:

Proof: Divide the interval $[a,b]$ into n sub-intervals by the points

$a = x_0 < x_1 < ... < x_{n-1} < x_n = b$.

Show that $F(b) - F(a) = \sum_{i=1}^{n}[F(x_i) - F(x_{i-1})]$.

CRQ, Calculus, August 2005, Q3b

C953b

Adapted discrimination

0.831

QI =0.927

0.865   0.839

Adapted expert opinion   Adapted confidence
deviation   deviation
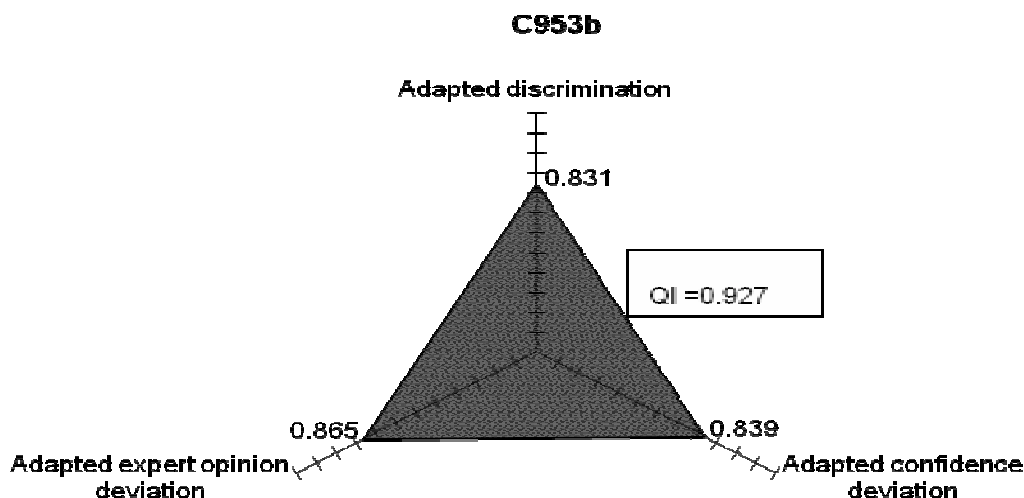
C953b		Comment
Assessment Component	Conceptual	
PRQ/CRQ	CRQ	
Item Difficulty	2.4	Difficult
Discrimination	0.831	Discriminates poorly
Confidence Index	0.839	Large deviation from expected confidence level
Expert Opinion	0.865	Large deviation from expected performance
Quality Index	0.927	Poor quality CRQ

202

4. Logical component

A662.2

Use properties of sigma notation and the fact that

$$\sum_{r=1}^{n} r = \frac{n(n+1)}{2} \text{ to prove that } \sum_{r=1}^{n} r^2 = \frac{n(n+1)(2n+1)}{6} .$$

CRQ, Algebra, June 2006, Q2.2

A6622

Adapted discrimination

QI =0.069

0.048

0.251

0.495

Adapted expert opinion
deviation

Adapted confidence
deviation
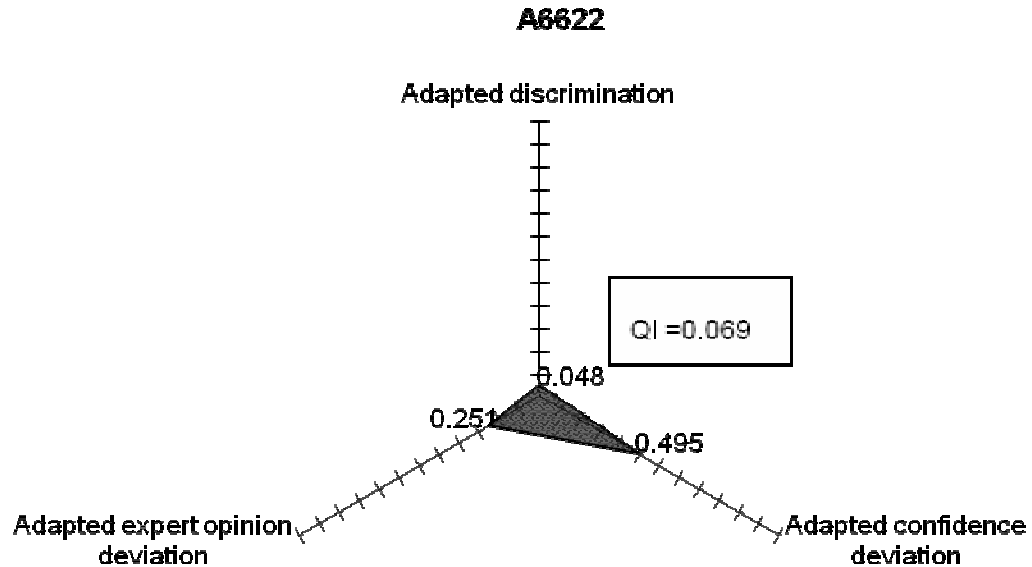
A662.2		Comment
Assessment Component	Logical	
PRQ/CRQ	CRQ	
Item Difficulty	1.52	Difficult
Discrimination	0.048	Discriminates well
Confidence Index	0.495	Average deviation from expected confidence level
Expert Opinion	0.251	Small deviation from expected performance
Quality Index	0.069	Good quality CRQ (excellent)

A55M08

You are given the sector OAB of a circle of radius 2 with $AC = p$.
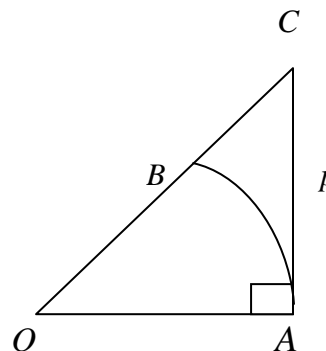
Arc length AB equals:

A. 2

B. $\arcsin 2/p$

C. $\arctan p/2$

D. $2\arctan(p/2)$

A55M08

Adapted discrimination

0.378

QI =0.265

0.504   0.479

Adapted expert opinion   Adapted confidence
deviation   deviation
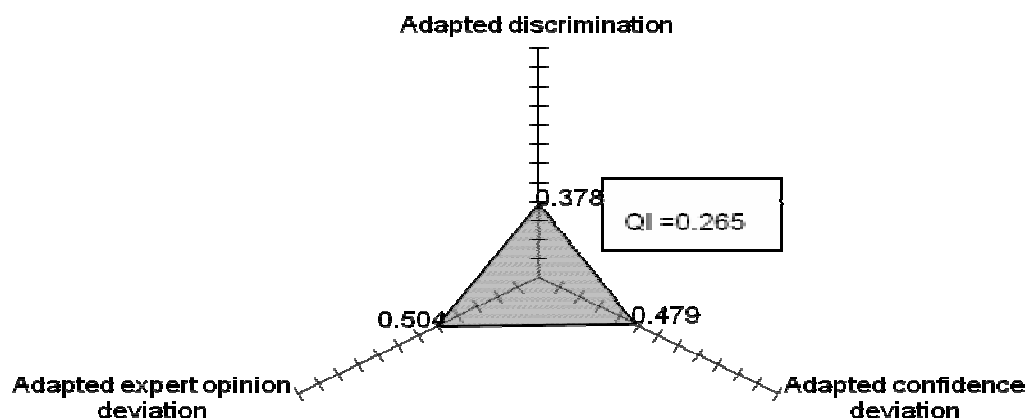
A55M08		Comment
Assessment Component	Logical	
PRQ/CRQ	PRQ	
Item Difficulty	0.15	Moderately difficult
Discrimination	0.378	Discriminates well
Confidence Index	0.479	Small deviation from expected confidence level
Expert Opinion	0.504	Average deviation from expected performance
Quality Index	0.265	Good quality PRQ (moderate)

A562a

A polar graph is defined by the equation $r(\theta) = 5\cos 3\theta$ for $\theta \in [0, 2\pi]$

Is the graph symmetric about the $x-$axis, the $y-$axis, both or neither? Motivate your answer.
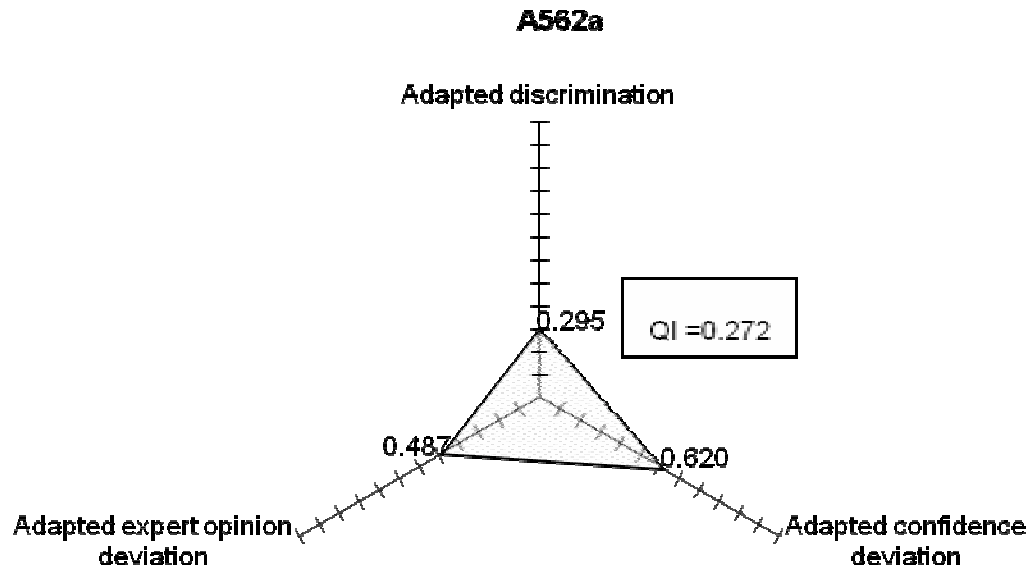
A562a

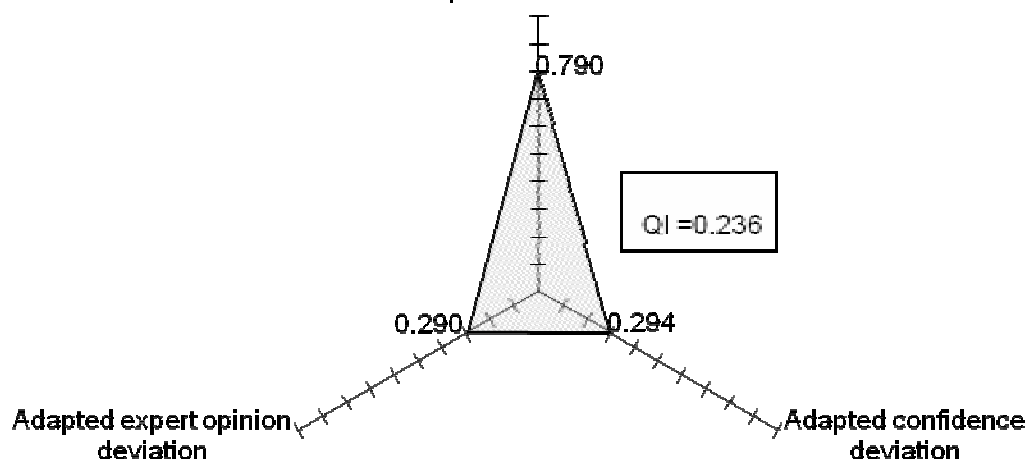| A562a		Comment
Assessment Component	Logical	
PRQ/CRQ	CRQ	
Item Difficulty	-1.62	Easy
Discrimination	0.295	Discriminates well
Confidence Index	0.620	Large deviation from expected confidence level
Expert Opinion	0.487	Small deviation from expected performance
Quality Index	0.272	Good quality CRQ (moderate)

A85M05

If $z = 3 + 2i$ and $w = 1 - 4i$, then in real-imaginary form $\dfrac{z}{w}$ equals:

A. $-\dfrac{5}{17} + \dfrac{14}{17}i$

B. $\dfrac{5}{15} - \dfrac{14}{\sqrt{15}}i$

C. $3 - 4i$

D. $\dfrac{11}{17} + \dfrac{14}{17}i$

PRQ, Algebra, August 2005, Tut Test Q5

A85M05
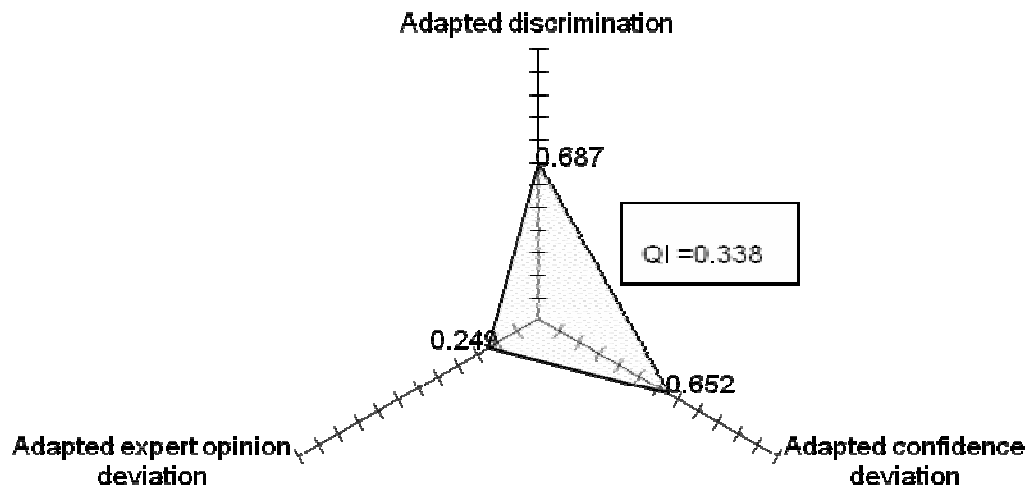
A85M05		Comment
Assessment Component	Logical	
PRQ/CRQ	PRQ	
Item Difficulty	-2.31	Easy
Discrimination	0.687	Discriminates poorly
Confidence Index	0.652	Large deviation from expected confidence level
Expert Opinion	0.249	Small deviation from expected performance
Quality Index	0.338	Poor quality PRQ

206

C46MA5

If $\lim_{x \to a}[f(x) + g(x)]$ exists, then

A. $\lim_{x \to a} f(x) = \lim_{x \to a} g(x)$.

B. neither $\lim_{x \to a} f(x)$ nor $\lim_{x \to a} g(x)$ exists.

C. both $\lim_{x \to a} f(x)$ and $\lim_{x \to a} g(x)$ exist.

D. we cannot tell if $\lim_{x \to a} f(x)$ or $\lim_{x \to a} g(x)$ exists.

PRQ, Calculus, March 2006, Tut Test A,Q5

C46MA5
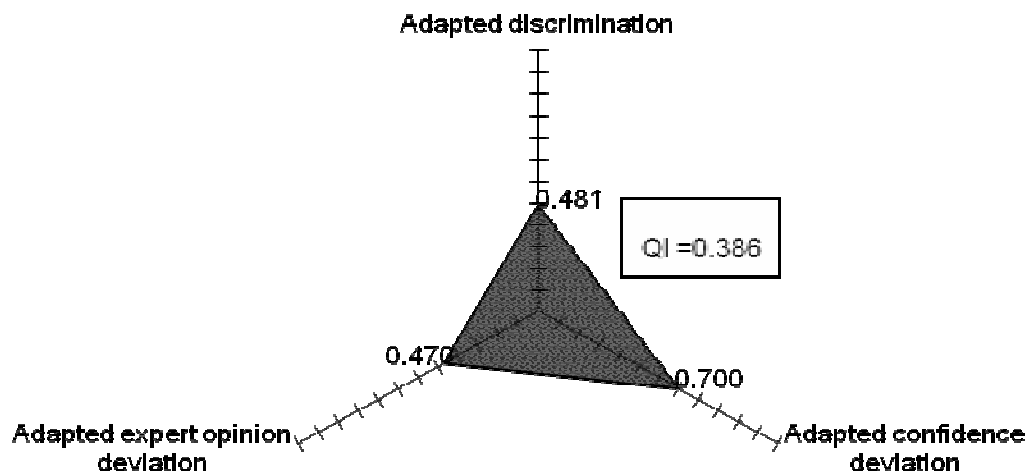
C46MA5		Comment
Assessment Component	Logical	
PRQ/CRQ	PRQ	
Item Difficulty	2.47	Difficult
Discrimination	0.481	Average discrimination
Confidence Index	0.700	Large deviation from expected confidence level
Expert Opinion	0.470	Small deviation from expected performance
Quality Index	0.386	Poor quality PRQ

207

A562d

A polar graph is defined by the equation $r(\theta) = 5\cos 3\theta$ for $\theta \in [0, 2\pi]$.

What is the name of this polar graph?
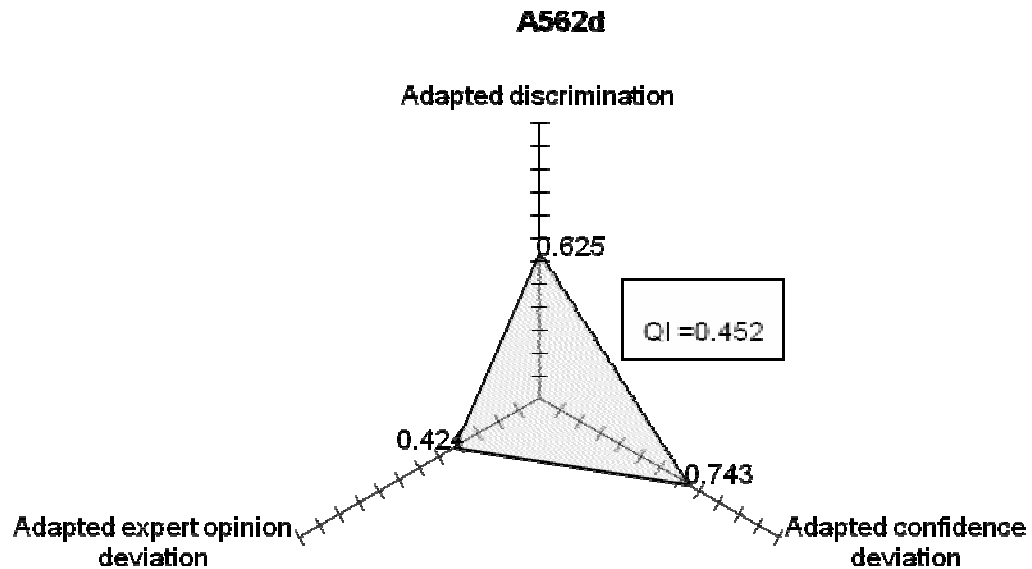
A562d

A562d		Comment
Assessment Component	Logical	
PRQ/CRQ	CRQ	
Item Difficulty	-1.42	Moderately easy
Discrimination	0.625	Discriminates poorly
Confidence Index	0.743	Large deviation from expected confidence level
Expert Opinion	0.424	Small deviation from expected performance
Quality Index	0.452	Poor quality CRQ

C563aii

Consider the following theorem:

Let f be a function that satisfies the following three conditions:

    (1) f is continuous on the closed interval $[a,b]$.

    (2) f is differentiable on the open interval (a,b).

    (3) $f(a) = f(b)$.

Then there exists a number $c \in (a,b)$ such that $f'(c) = 0$.

Let $f(x) > f(a)$ for some $x \in (a,b)$.

    Give a **complete proof** of the theorem in this case.

CRQ, Calculus, May 2006, Q3aii

C563aii
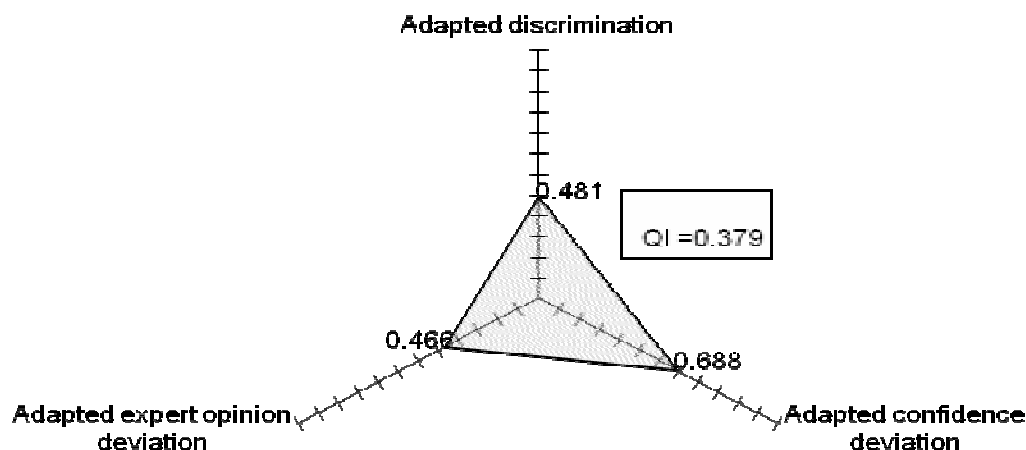
C563aii		Comment
Assessment Component	Logical	
PRQ/CRQ	CRQ	
Item Difficulty	-0.46	Moderately easy
Discrimination	0.481	Average discrimination
Confidence Index	0.688	Large deviation from expected confidence level
Expert Opinion	0.466	Small deviation from expected performance
Quality Index	0.379	Poor quality CRQ

5.   Modelling component

A652b

Solve $-2\cos x + 2\sqrt{3}\sin x = 4\cos^2 x - 4\sin^2 x$
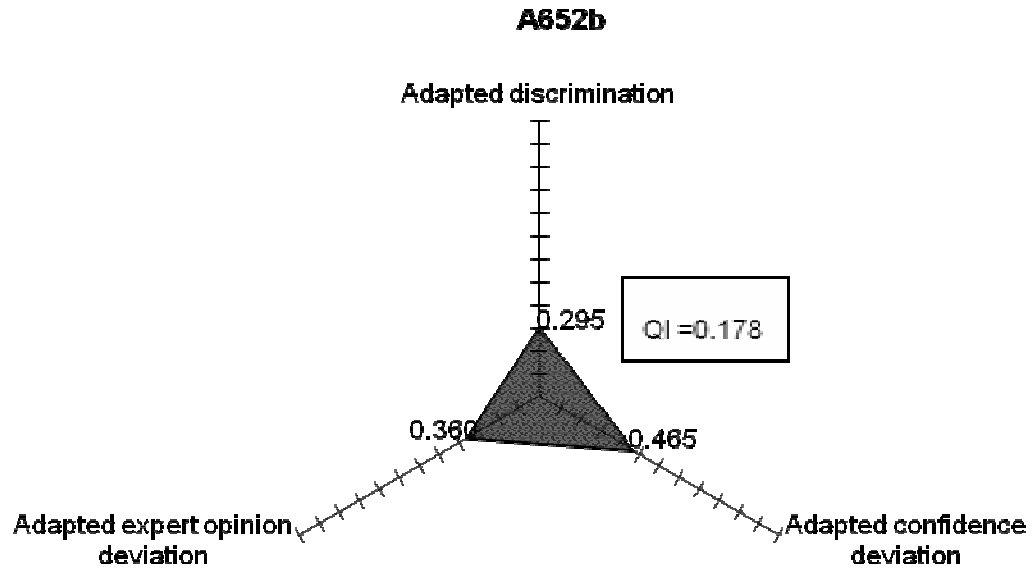
A652b

A652b		Comment
Assessment Component	Modelling	
PRQ/CRQ	CRQ	
Item Difficulty	2.81	Difficult
Discrimination	0.295	Discriminates well
Confidence Index	0.465	Small deviation from expected confidence level
Expert Opinion	0.360	Small deviation from expected performance
Quality Index	0.178	Good quality CRQ (excellent)

A95M03

If $\vec{a} = (1,2), \vec{b} = (-1,3), \vec{c} = (4,-2)$ and $\vec{d} = (3,-3),$ then $(\vec{a} \cdot \vec{d})\vec{b} - (\vec{b} \cdot \vec{c})\vec{d}$ equals

A. $(-54,12)$

B. -4

C. $3(11,-13)$

D. not possible
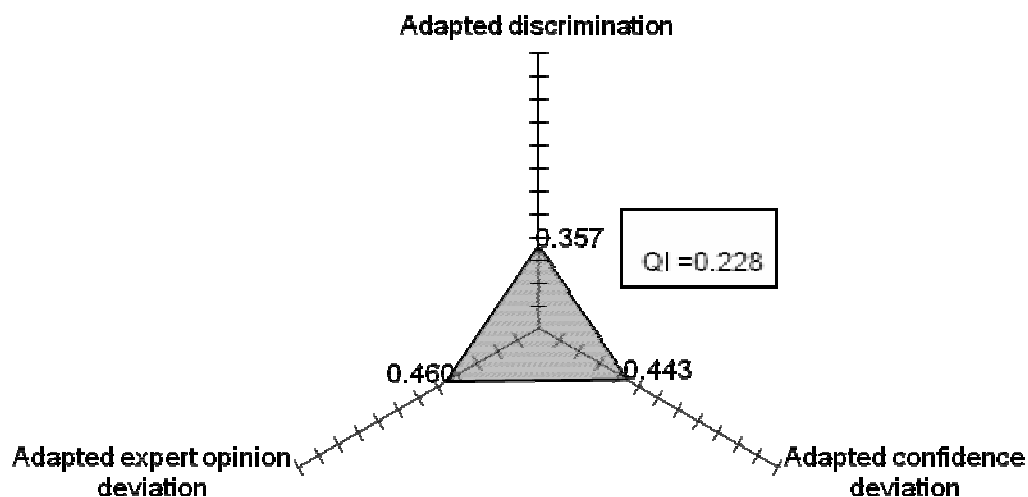
A95M03

A95M03		Comment
Assessment Component	Modelling	
PRQ/CRQ	PRQ	
Item Difficulty	0.84	Moderately difficult
Discrimination	0.357	Discriminates well
Confidence Index	0.443	Small deviation from expected confidence level
Expert Opinion	0.460	Small deviation from expected performance
Quality Index	0.228	Good quality PRQ

C35M01

$$\lim_{h \to 0} \frac{\sqrt{9+h}-3}{h} \text{ is equal to}$$

A. $\lim\limits_{h \to 0} \dfrac{1}{\sqrt{9+h}+3}$

B. The slope of the tangent line to $y = \sqrt{x}$ at the point $P(9,3)$

C. The slope of the tangent line to $y = \sqrt{x}$ at the point $P(9,-3)$

D. Both (A) and (B)

E. All of $(A), (B)$ and (C)

PRQ, Calculus, March 2005, Q1

C35M01
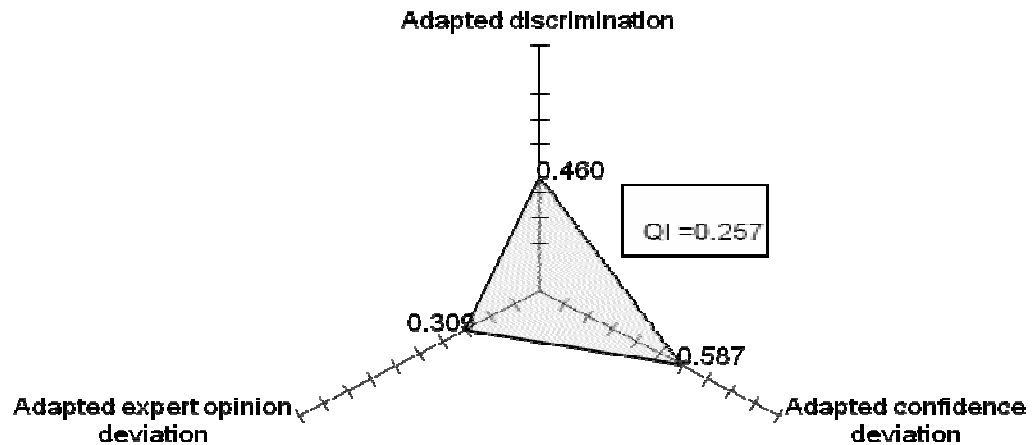
C35M01		Comment
Assessment Component	Modelling	
PRQ/CRQ	PRQ	
Item Difficulty	-0.36	Moderately easy
Discrimination	0.460	Discriminates well
Confidence Index	0.587	Large deviation from expected confidence level
Expert Opinion	0.309	Small deviation from expected performance
Quality Index	0.257	Good quality PRQ (moderate)

212

C1156a

Match each of the differential equations given in Column A with the type listed in Column B.

A. Differential Equation	B. Type
a. $\dfrac{dy}{dx} - \dfrac{y}{x} = \ln x$	1. Variable separable
b. $\dfrac{dy}{dx} = e^x \Big/ e^y$	2. Homogeneous
c. $(x^2 + y^2)dx + 2xy\,dy = 0$	3. Exact
d. $2x + y^3 + (3xy^2 + ye^{2y})\dfrac{dy}{dx} = 0$	4. Linear

CRQ, Calculus, November 2005, Q6a

C1156a
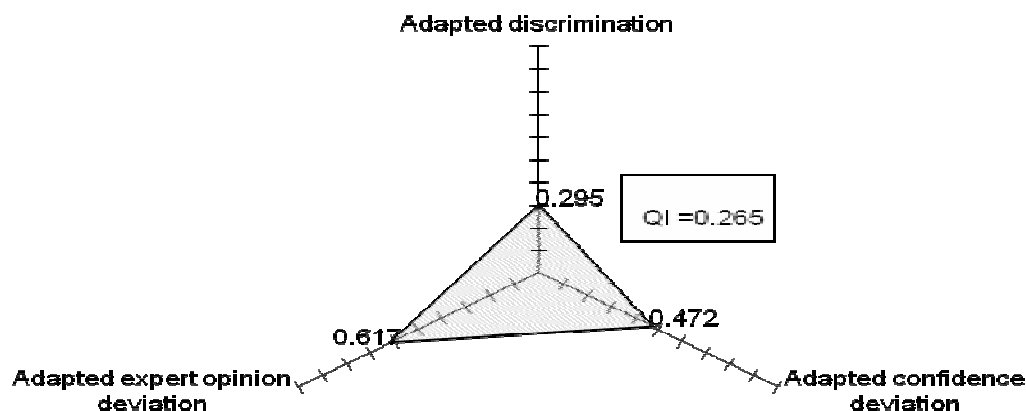
C1156a		Comment
Assessment Component	Modelling	
PRQ/CRQ	CRQ	
Item Difficulty	-0.22	Moderately easy
Discrimination	0.295	Discriminates well
Confidence Index	0.472	Small deviation from expected confidence level
Expert Opinion	0.617	Large deviation from expected performance
Quality Index	0.265	Good quality CRQ (moderate)

C66M06

Let $f(x)$ be a function such that $f(4) = -1$ and $f'(4) = 2$. If $x < 4$, then $f''(x) < 0$ and if $x > 4$, then $f''(x) > 0$. The point $(4, -1)$ is a ☐ of the graph of f.

A. Relative maximum

B. Relative minimum

C. Critical point

D. Point of inflection

E. None of the above

PRQ, Calculus, June 2006, Q6

C66M06

Adapted discrimination
0.687

QI = 0.379

0.496

0.452

Adapted expert opinion deviation

Adapted confidence deviation
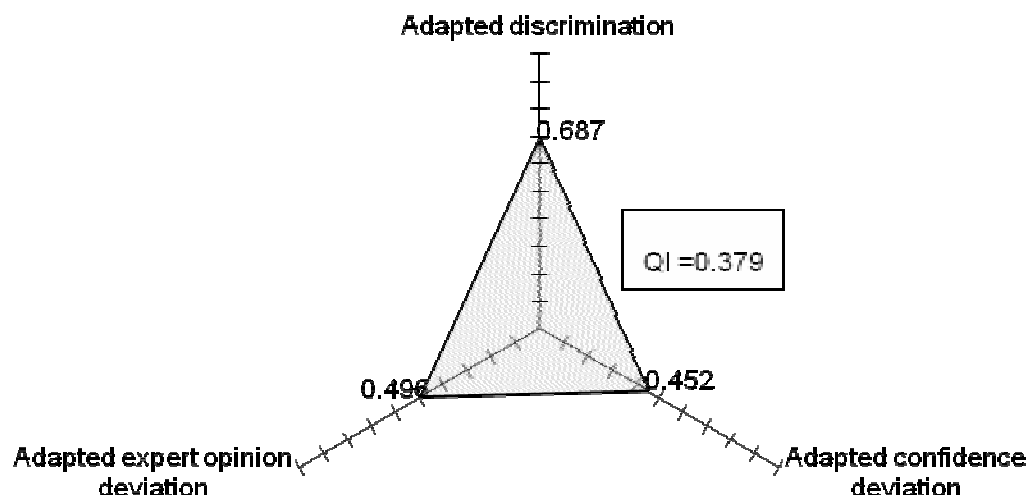
C66M06		Comment
Assessment Component	Modelling	
PRQ/CRQ	PRQ	
Item Difficulty	-1.00	Moderately easy
Discrimination	0.687	Discriminates poorly
Confidence Index	0.452	Small deviation from expected confidence level
Expert Opinion	0.496	Average deviation from expected performance
Quality Index	0.379	Poor quality PRQ

214

C561aii

A bacterial colony is estimated to have a population of $P(t) = \dfrac{24t + 10}{t^2 + 1}$ million, t hours after the

introduction of a toxin.

Is the population increasing or decreasing at this time?

C561aii
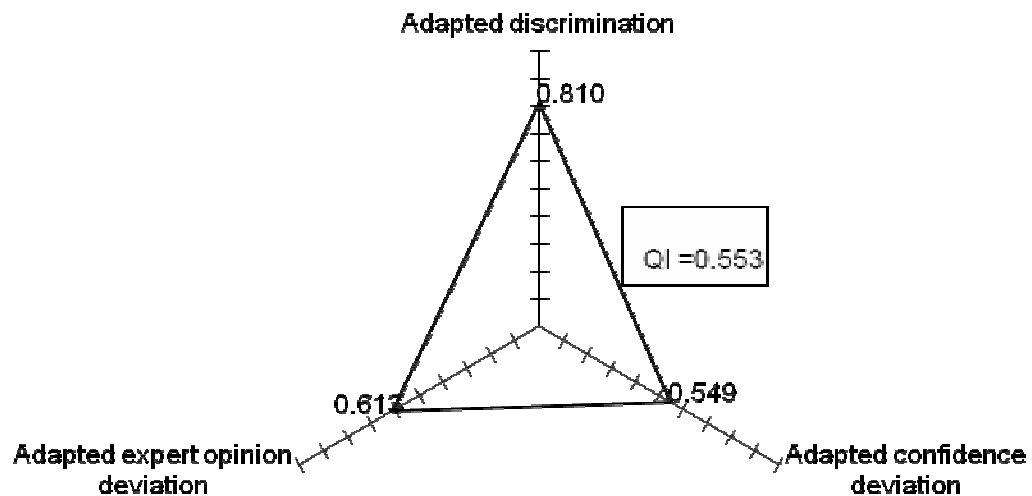
C561aii		Comment
Assessment Component	Modelling	
PRQ/CRQ	CRQ	
Item Difficulty	-4.51	Very easy
Discrimination	0.810	Discriminates poorly
Confidence Index	0.549	Large deviation from expected confidence level
Expert Opinion	0.613	Large deviation from expected performance
Quality Index	0.553	Poor quality CRQ

6.   Problem solving component

C1152a

Split $\dfrac{3}{(x-1)(x^2+x+1)}$ into partial fractions.

C1152a
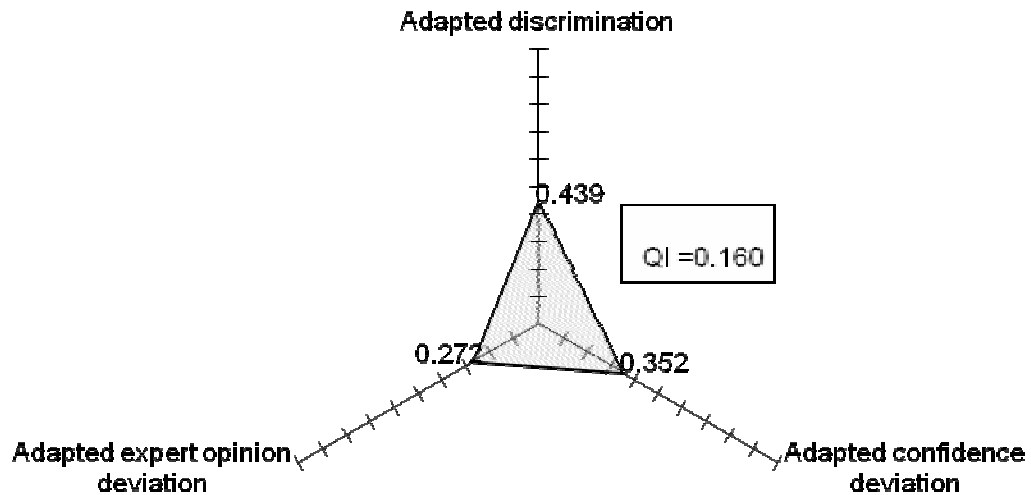
C1152a		Comment
Assessment Component	Problem solving	
PRQ/CRQ	CRQ	
Item Difficulty	-1.37	Moderately easy
Discrimination	0.439	Discriminates well
Confidence Index	0.352	Small deviation from expected confidence level
Expert Opinion	0.272	Small deviation from expected performance
Quality Index	0.160	Good quality CRQ (moderate)

C65M10

The points of inflection for the function $f(x) = 8x + 2 - \sin x$ for $0 < x < 3\pi,$ are

A. $(\pi, 8\pi)$ and $(2\pi, 16\pi + 2)$

B. $(\pi, 2)$ and $(2\pi, 16\pi + 2)$

C. $(\pi, 8\pi)$ and $(2\pi, 16\pi)$

D. $(\pi, 8\pi + 2)$ and $(2\pi, 16\pi + 2)$

E. $(\pi, 8\pi + 2)$ and $(2\pi, 16\pi)$

PRQ, Calculus, June 2005, Q10

C65M10

Adapted discrimination

0.213   QI =0.181

0.352

0.609

Adapted expert opinion deviation

Adapted confidence deviation
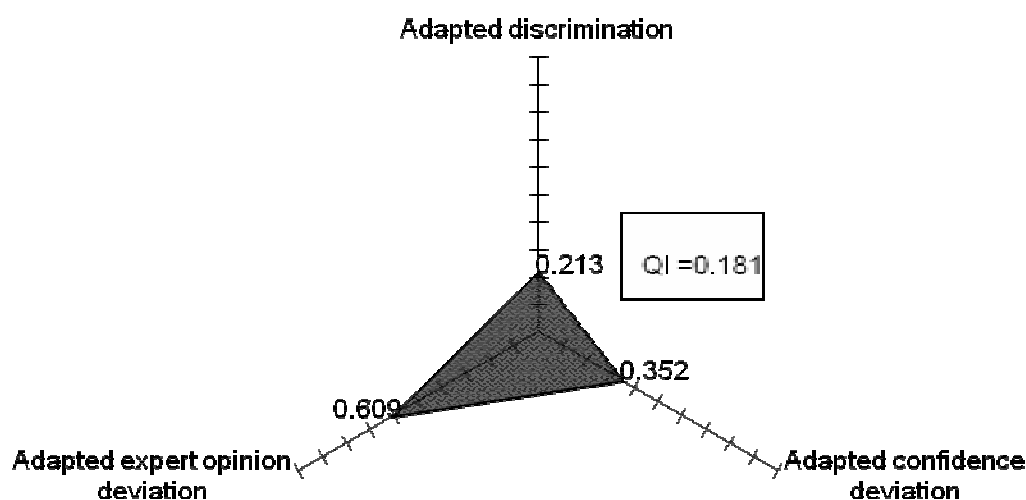
C65M10		Comment
Assessment Component	Problem solving	
PRQ/CRQ	PRQ	
Item Difficulty	1.73	Difficult
Discrimination	0.213	Discriminates well
Confidence Index	0.352	Small deviation from expected confidence level
Expert Opinion	0.609	Large deviation from expected performance
Quality Index	0.181	Good quality PRQ

217

A65M04

If $\dfrac{1}{2}\arccos 2x = \dfrac{\pi}{2}$, then x equals

A. 0

B. -1

C. $\dfrac{1}{2}$

D. $-\dfrac{1}{2}$
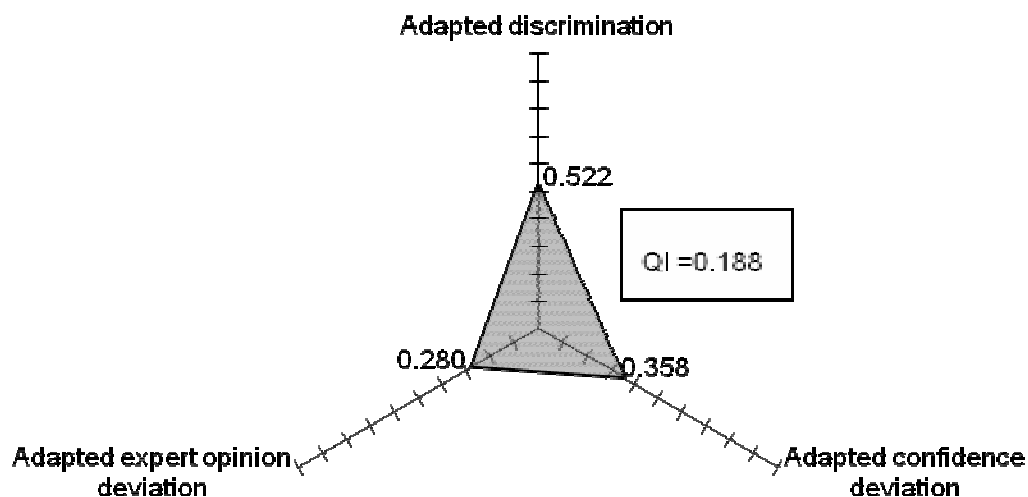
A65M04

A65M04		Comment
Assessment Component	Problem solving	
PRQ/CRQ	PRQ	
Item Difficulty	0.14	Moderately difficult
Discrimination	0.522	Average discrimination
Confidence Index	0.358	Small deviation from expected confidence level
Expert Opinion	0.280	Small deviation from expected performance
Quality Index	0.188	Good quality PRQ

218

A951

Evaluate $\sum_{r=1}^{100}[(r+1)^{r+1} - r^r]$.
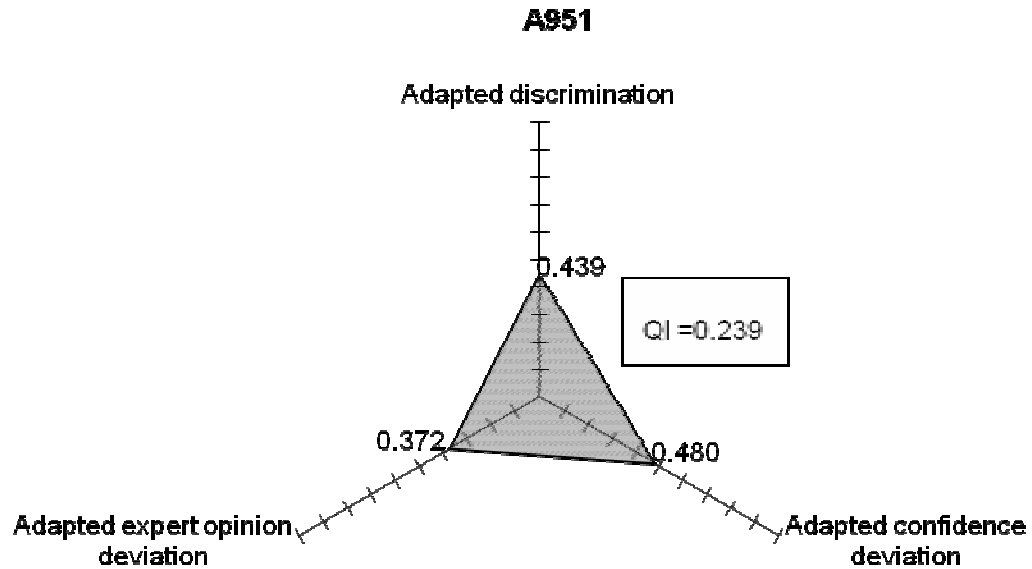
A951

A951		Comment
Assessment Component	Problem solving	
PRQ/CRQ	CRQ	
Item Difficulty	0.67	Moderately difficult
Discrimination	0.439	Discriminates well
Confidence Index	0.480	Small deviation from expected confidence level
Expert Opinion	0.372	Small deviation from expected performance
Quality Index	0.239	Good quality CRQ (moderate)

A65M02

$$\sum_{i=r+1}^{k} \pi =$$

A. $\pi(r+1-k)$

B. $k(r-\pi+1)$

C. $\pi(k-r+2)$

D. $\pi(k-r)$
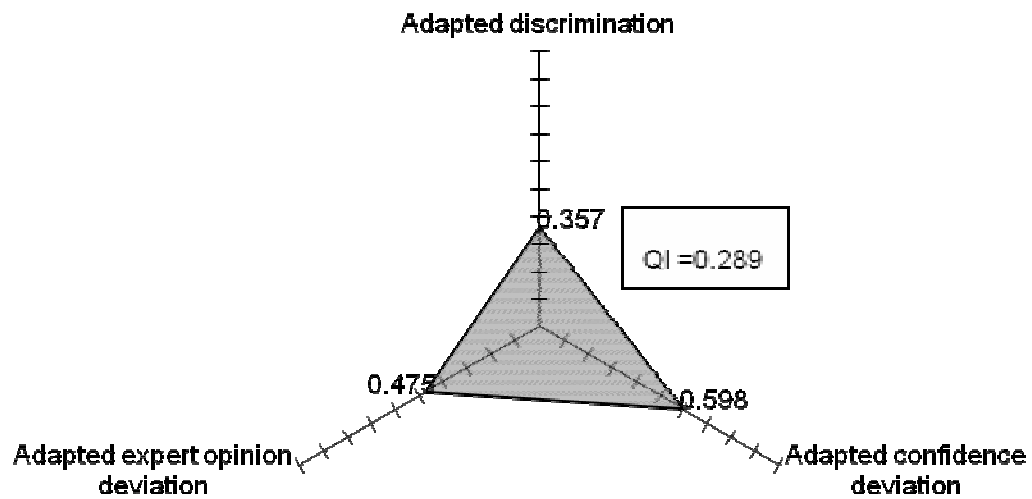
A65M02

A65M02		Comment
Assessment Component	Problem solving	
PRQ/CRQ	PRQ	
Item Difficulty	0.98	Moderately difficult
Discrimination	0.357	Discriminates well
Confidence Index	0.598	Large deviation from expected confidence level
Expert Opinion	0.475	Small deviation from expected performance
Quality Index	0.289	Poor quality PRQ (moderate)

C55M01

Determine from the graph of $y = f(x)$ whether f possesses extrema on the interval $[a,b]$.
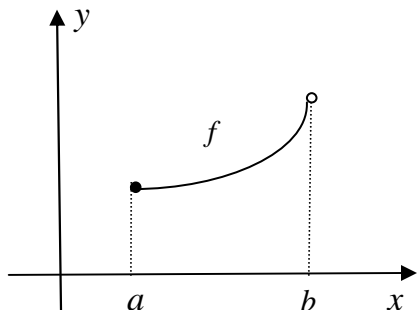
A. Maximum at $x = a$; minimum at $x = b$.

B. Maximum at $x = b$; minimum at $x = a$.

C. No extrema.

D. No maximum; minimum at $x = a$.

PRQ, Calculus, May 2005, Q1

C55M01

QI = 0.349
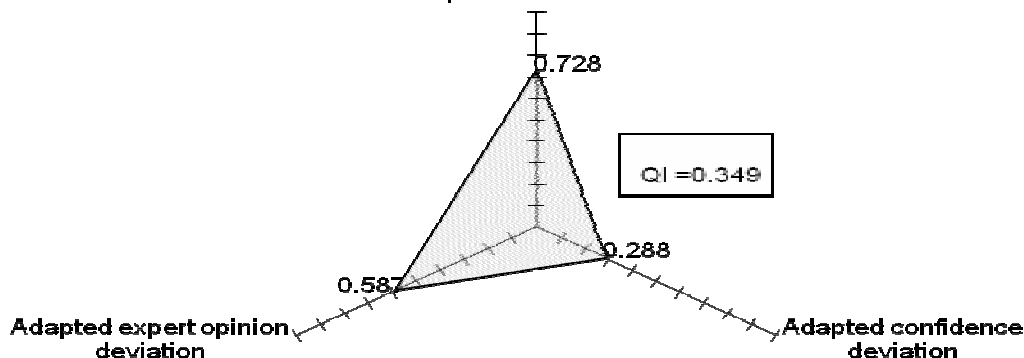
C55M01		Comment
Assessment Component	Problem solving	
PRQ/CRQ	PRQ	
Item Difficulty	-0.50	Moderately easy
Discrimination	0.728	Discriminates poorly
Confidence Index	0.288	Small deviation from expected confidence level
Expert Opinion	0.587	Large deviation from expected performance
Quality Index	0.349	Poor quality PRQ

C663c

In a given semi-circle of radius 2, a rectangle is inscribed as shown in the figure below.
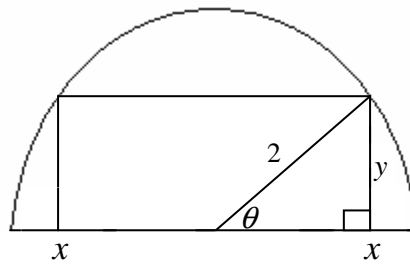
Find the value of θ corresponding to the maximum area, and test whether this value

for θ gives a maximum.

<div align="right">CRQ, Calculus, June 2006, Q3c</div>
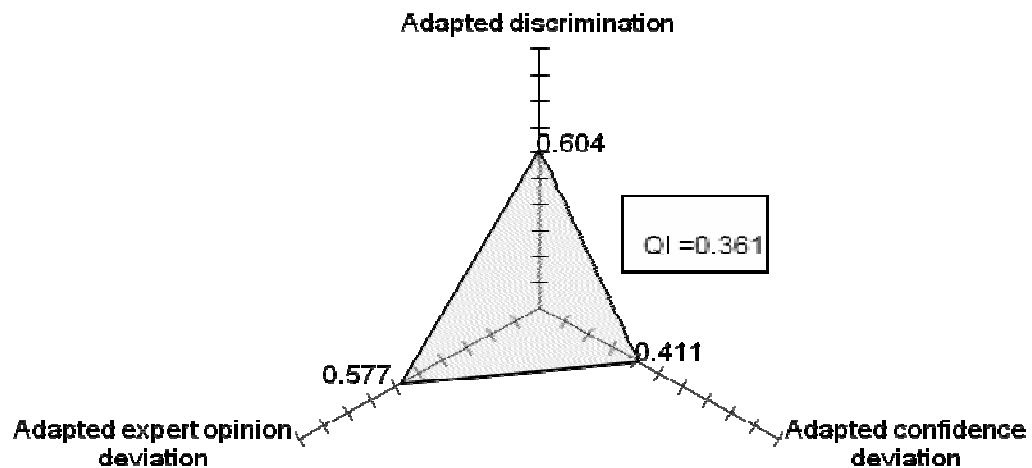
C663c		Comment
Assessment Component	Problem solving	
PRQ/CRQ	CRQ	
Item Difficulty	-0.13	Moderately easy
Discrimination	0.604	Discriminates poorly
Confidence Index	0.411	Small deviation from expected confidence level
Expert Opinion	0.577	Large deviation from expected performance
Quality Index	0.361	Poor quality CRQ

A1154bii

$$M = \begin{pmatrix} 1 & -2 & -3 & : & 3 \\ -1 & 3 & 5 & : & -4 \\ 4 & -5 & k^2-15 & : & k+12 \end{pmatrix}$$

Suppose the system given by M represents three planes, $P_1, P_2, P_3.$ That is, we have:

$$P_1: x - 2y - 3z = 3$$
$$P_2: -x + 3y + 5z = -4$$
$$P_3: 4x - 5y + (k^2-15)z = k+12$$

Find the value(s) of k such that the three planes intersect in a single point. Do not calculate the co-ordinates of that point.
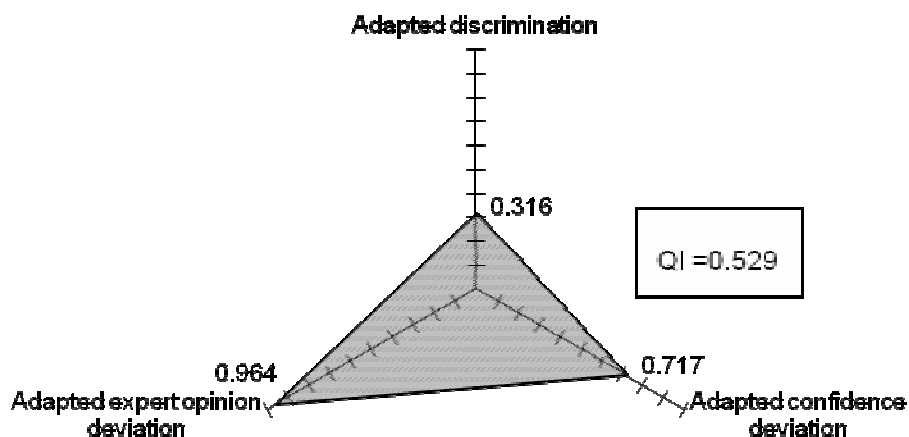
A1154biii

A1154biii		Comment
Assessment Component	Problem solving	
PRQ/CRQ	CRQ	
Item Difficulty	0.35	Moderately difficult
Discrimination	0.316	Discriminates well
Confidence Index	0.717	Large deviation from expected confidence level
Expert Opinion	0.964	Large deviation from expected performance
Quality Index	0.529	Poor quality CRQ

7.   Consolidation component

C951

Rewrite the following integral as the sum of integrals such that there are no absolute values. DO NOT solve the integral. Give full reasons for your answer.

$$\int_{-2}^{5}\left|4x-x^2\right|dx$$

C951
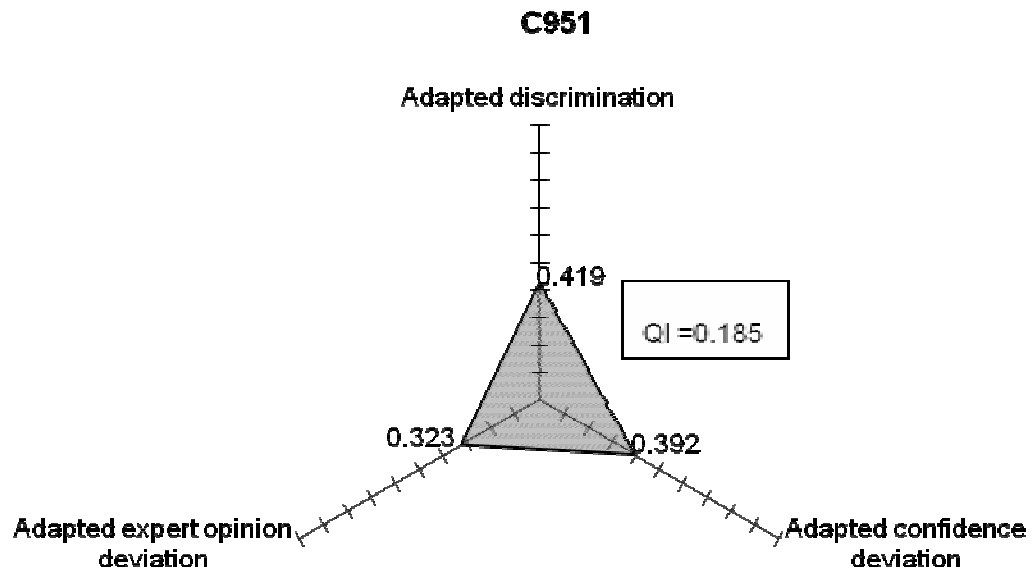
C951		Comment
Assessment Component	Consolidation	
PRQ/CRQ	CRQ	
Item Difficulty	0.86	Moderately difficult
Discrimination	0.419	Discriminates well
Confidence Index	0.392	Small deviation from expected confidence level
Expert Opinion	0.323	Small deviation from expected performance
Quality Index	0.185	Good quality CRQ

A45MA4

If f is an odd function and g is an even function then

A. $f \circ g$ is an even function

B. $f \circ g$ is an odd function

C. f is a one-to-one function

D. g is a one-to-one function

PRQ, Algebra, March 2005, Tut Test A, Q4

A45MA4

Adapted discrimination

0.275

QI =0.207

0.296

0.698

Adapted expert opinion
deviation

Adapted confidence
deviation
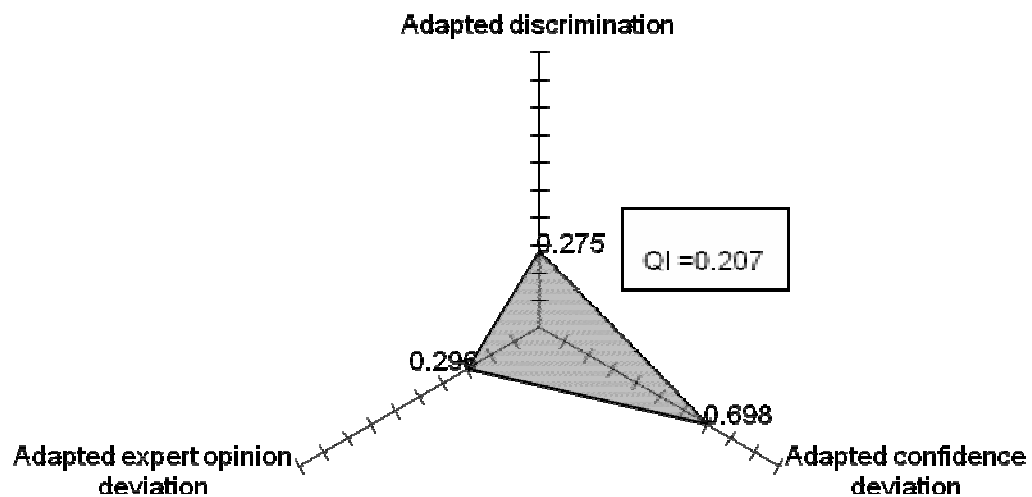
A45MA4		Comment
Assessment Component	Consolidation	
PRQ/CRQ	PRQ	
Item Difficulty	1.11	Moderately difficult
Discrimination	0.275	Discriminates well
Confidence Index	0.698	Large deviation from expected confidence level
Expert Opinion	0.296	Small deviation from expected performance
Quality Index	0.207	Good quality PRQ

A661.2

This question deals with the statement $P(n): n^3 + (n+1)^3 + (n+2)^3$ is divisible by 9.

Use Pascal's triangle to expand and then simplify $(k+3)^3$.
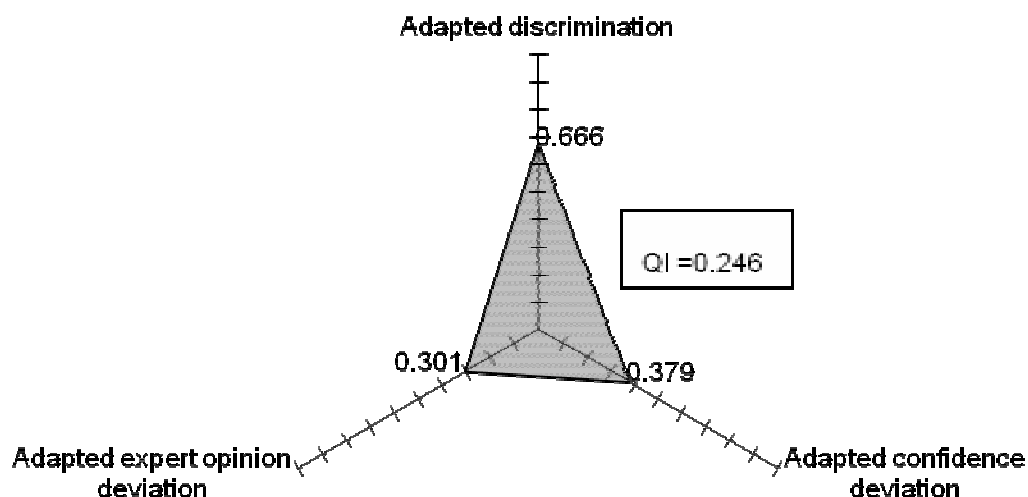
A661.2

A661.2		Comment
Assessment Component	Consolidation	
PRQ/CRQ	CRQ	
Item Difficulty	0.02	Moderately difficult
Discrimination	0.666	Discriminates poorly
Confidence Index	0.379	Small deviation from expected confidence level
Expert Opinion	0.301	Small deviation from expected performance
Quality Index	0.246	Good quality CRQ (moderate)

226

C85M07

On which interval is the function $f(x) = e^{3x} - e^x$ increasing?

A. $(\ln 9, \infty)$

B. $(0, \infty)$

C. $(-\infty, \infty)$

D. $(-\dfrac{1}{2} \ln 3, \infty)$

E. None of the above

C85M07

Adapted discrimination

0.687

QI =0.272

0.230

0.514

Adapted expert opinion deviation

Adapted confidence deviation
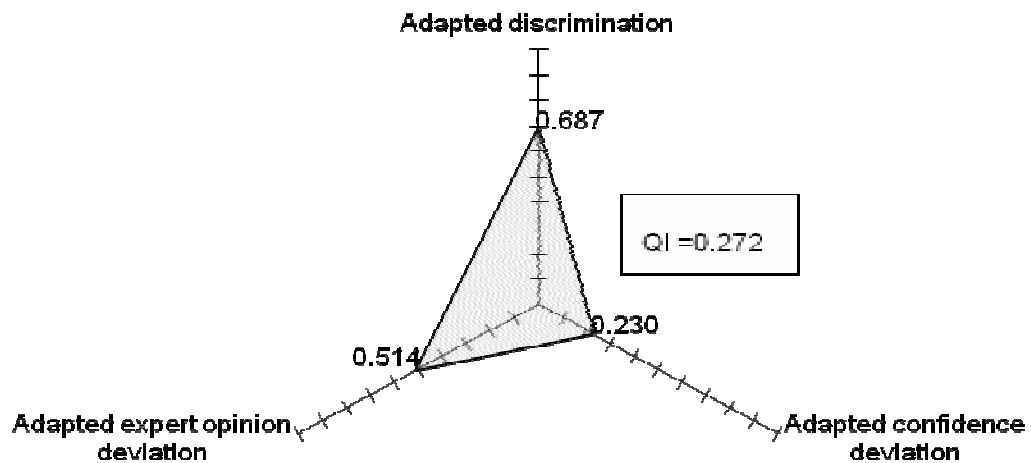
C85M07		Comment
Assessment Component	Consolidation	
PRQ/CRQ	PRQ	
Item Difficulty	-1.17	Moderately easy
Discrimination	0.687	Discriminates poorly
Confidence Index	0.230	Small deviation from expected confidence level
Expert Opinion	0.514	Average deviation from expected performance
Quality Index	0.272	Good quality PRQ (moderate)

C654

State the Fundamental Theorem of Calculus.

C654

Adapted discrimination

0.481

QI =0.310

0.248

0.819

Adapted expert opinion deviation

Adapted confidence deviation
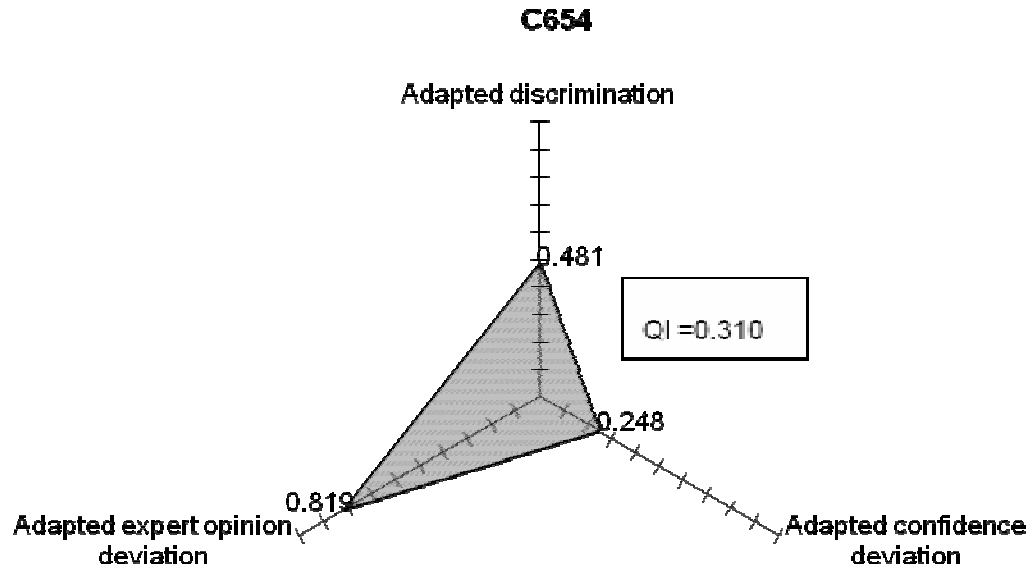
C654		Comment
Assessment Component	Consolidation	
PRQ/CRQ	CRQ	
Item Difficulty	0.29	Moderately difficult
Discrimination	0.481	Average discrimination
Confidence Index	0.248	Small deviation from expected confidence level
Expert Opinion	0.819	Large deviation from expected performance
Quality Index	0.310	Poor quality CRQ (moderate)

A56M01

Let $y = f(x) = \cos(\arcsin x)$. Then the range of f is

A. $\{y \mid 0 \le y \le 1\}$

B. $\{y \mid -1 \le y \le 1\}$

C. $\{y \mid -\dfrac{\pi}{2} < y < \dfrac{\pi}{2}\}$

D. $\{y \mid -\dfrac{\pi}{2} \le y \le \dfrac{\pi}{2}\}$

E. None of the above

PRQ, Algebra, May 2006, Q1
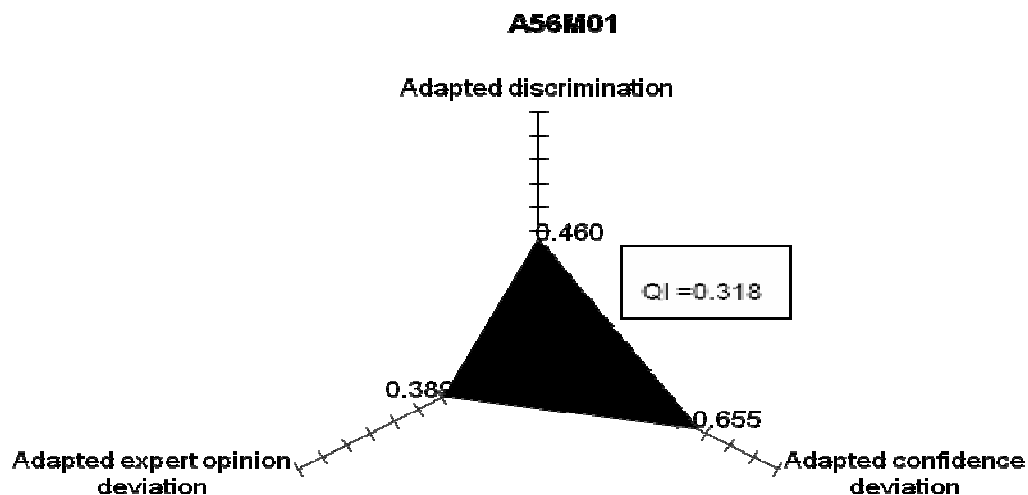
A56M01

A56M01		Comment
Assessment Component	Consolidation	
PRQ/CRQ	PRQ	
Item Difficulty	3.07	Very difficult
Discrimination	0.460	Discriminates fairly well
Confidence Index	0.655	Large deviation from expected confidence level
Expert Opinion	0.389	Small deviation from expected performance
Quality Index	0.318	Poor quality PRQ (moderate)

C662f

Let $f(x) = \dfrac{x^2}{(x-2)^2}$.

You may assume that $f'(x) = \dfrac{-4x}{(x-2)^3}$ and $f''(x) = \dfrac{8x+8}{(x-2)^4}$.

Find the points of inflection of f (if any).

C662f

Adapted discrimination

0.646

QI =0.595

0.609
Adapted expert opinion
deviation

0.783
Adapted confidence
deviation
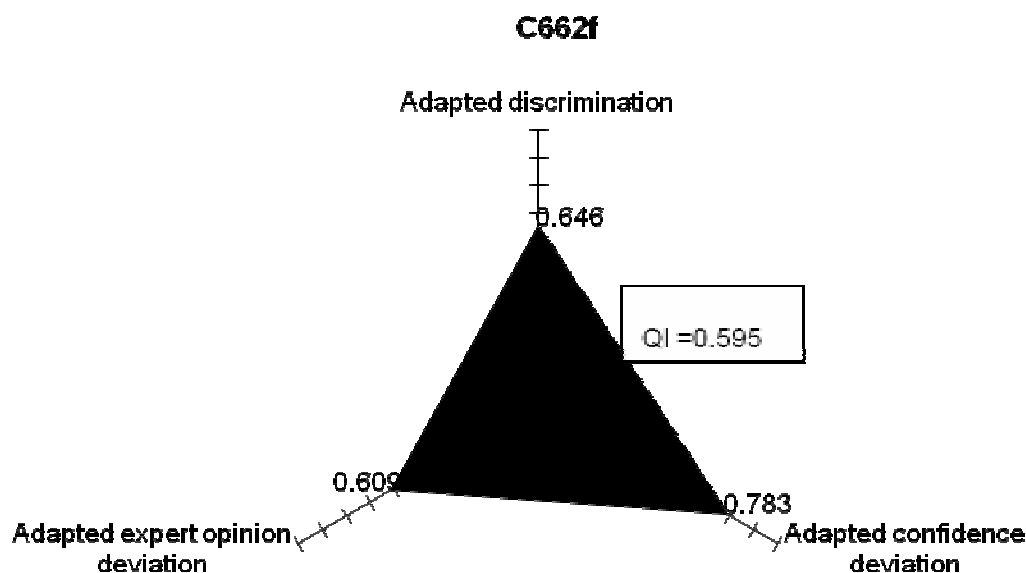
C662f		Comment
Assessment Component	Consolidation	
PRQ/CRQ	CRQ	
Item Difficulty	3.75	Very difficult
Discrimination	0.646	Discriminates poorly
Confidence Index	0.783	Large deviation from expected confidence level
Expert Opinion	0.609	Large deviation from expected performance
Quality Index	0.595	Poor quality CRQ

C46MB6

$$\lim_{x \to -1} \frac{x^2 + 4x + 3}{x^2 - 1} =$$

A. -1

B. 0

C. undefined

D. 4
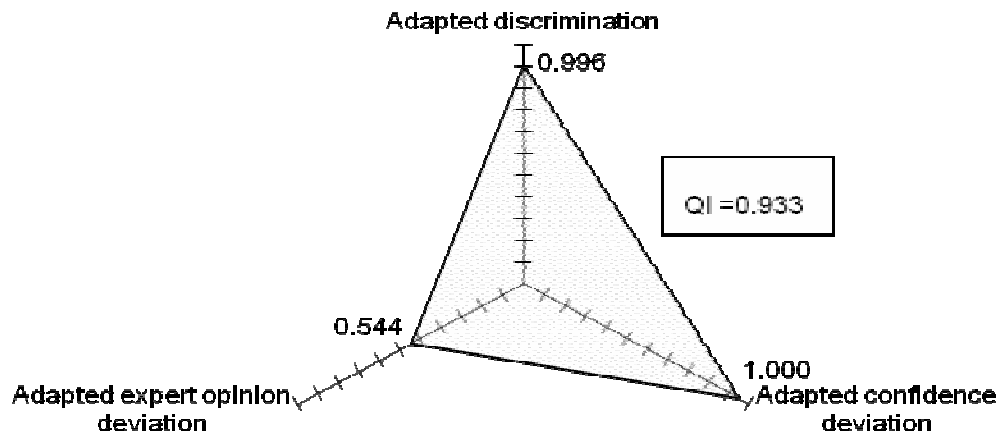
C46MB6

C46MB6		Comment
Assessment Component	Consolidation	
PRQ/CRQ	PRQ	
Item Difficulty	-2.24	Easy
Discrimination	0.996	Discriminates poorly
Confidence Index	1.000	Large deviation from expected confidence level
Expert Opinion	0.544	Large deviation from expected performance
Quality Index	0.933	Poor quality PRQ

231

6.4   RESULTS

6.4.1   Comparison of PRQs and CRQs within each assessment component

Table 6.3 summarises the quality of the item, both PRQs and CRQs, within each assessment component. Within each component the number of good and poor quality items are given, both for the PRQ and CRQ formats. The numbers are also given as percentages of the total number of items.

Table 6.3:   Component analysis – trends.

COMPONENT	No. of PRQs	No. of CRQs	Total no. of items	Good quality items	Poor quality items	Good PRQs	Good CRQs	Poor PRQs	Poor CRQs
1.Technical	11	22	33	17 [52%]	16 [48%]	8 [73%]	9 [41%]	3 [27%]	13 [59%]
2.Disciplinary	24	34	58	28 [48%]	30 [52%]	12 [50%]	16 [47%]	12 [50%]	18 [53%]
3.Conceptual	26	30	56	28 [50%]	28 [50%]	14 [54%]	14 [47%]	12 [46%]	16 [53%]
4.Logical	7	6	13	5 [39%]	8 [61%]	1 [14%]	4 [67%]	6 [86%]	2 [33%]
5.Modelling	3	10	13	8 [62%]	5 [38%]	2 [67%]	6 [60%]	1 [33%]	4 [40%]
6.Problem solving	7	4	11	6 [55%]	5 [45%]	4 [57%]	2 [50%]	3 [43%]	2 [50%]
7.Consolidation	16	7	23	12 [52%]	11 [48%]	7 [44%]	5 [71%]	9 [56%]	2 [29%]

1.   Technical

In the technical assessment component, there is a higher percentage (73%) of good PRQs than good CRQs (41%). 73% good PRQs compared to good 41% CRQs shows us that PRQs are more successful than CRQs as an assessment format in the technical component. There is also a much higher percentage (73%) of good PRQs than poor PRQs (27%). CRQs, however, are not that successful in this component, with the results showing 59% poor CRQs compared to 41% good CRQs. The conclusion is that the technical assessment component lends itself better to PRQs than to CRQs.

2.   Disciplinary

In this study, the disciplinary component is the assessment component with the most items (58), of which 34 were CRQs and 24 were PRQs. In this component it is interesting to note that the percentages of good PRQs (50%) and good CRQs (47%) are almost equal. In addition, there is no difference between the good PRQs (50%) and the poor PRQs (50%), with very little difference between the good CRQs (47%) and poor CRQs (53%). PRQs and CRQs can be considered as equally successful assessment formats in the disciplinary component.

3.   Conceptual

The conceptual component also contained many items (56), with an almost equal number of PRQs and CRQs (26 PRQs versus 30 CRQs). 50% of the items are of good quality and 50% are of poor quality. In this component, there is no clear trend that PRQs are better than CRQs or vice versa. There is a slight leaning towards good PRQ assessment (47% good CRQs compared to 54% good PRQs). Therefore, in the conceptual assessment component, PRQs could be used as successfully as CRQs as a format of assessment.

4.   Logical

In this study, it is interesting to note that the majority of questions within the logical component were of a poor quality mainly due to the large percentage of poor PRQs. There are noticeably more good quality CRQs (67%) than good quality PRQs (14%), and noticeably more poor quality PRQs (86%) than poor quality CRQs (33%). A very high percentage of the PRQs (86%) in the logical component were of a poor quality.   The conclusion is that the logical assessment component lends itself better to CRQs than to PRQs.

5.   Modelling

In the modelling component, very few PRQs were used as assessment items in comparison to CRQs, 3 PRQs versus 10 CRQs, probably because it is difficult to set PRQs in this component. Despite the small number of PRQs, it was encouraging to note that the good PRQs (67%) far outweighed the poor PRQs

(33%). So in terms of quality, the PRQs were highly successful in the modelling component. There are also more good CRQs (60%) than poor CRQs (40%). It appears that although more difficult to set in the modelling component, PRQs could be used as successfully in the modelling assessment component as CRQs.

6.   Problem solving

Although the problem solving component had the least number of items (11), it is interesting to note that there are more PRQs (7) than CRQs (4). There is a slightly higher percentage (57%) of good PRQs than good CRQs (50%). Although the sample is too small to make definite conclusions, there is no reason to disregard the use of PRQs in this assessment component. In fact, PRQs seem to be slightly more successful than CRQs, and the conclusion is that PRQ assessment format can add value to the assessment of the problem solving component.

7.   Consolidation

It was somewhat surprising to note that corresponding to the highest level of conceptual difficulty, the consolidation component displayed an unusually higher proportion of PRQs (16) to CRQs (7). This supports the earlier claim that PRQs are not only appropriate for testing lower level cognitive skills (Adkins, 1974; Aiken, 1987; Haladyna, 1999; Isaacs, 1994; Johnson, 1989; Oosterhof, 1994; Thorndike, 1997; Williams, 2006). In the consolidation component there is a significant higher percentage (71%) of good CRQs than good PRQs (44%). In addition, there is a higher percentage of poor PRQs (56%) than good PRQs (44%). The high percentage of good CRQs (71%) in comparison to poor CRQs (29%) indicates that the consolidation assessment component lends itself better to CRQs than to PRQs.

CHAPTER 7:   DISCUSSION AND CONCLUSIONS

In Chapter 7, I set about discussing my research results. The discussion in this chapter will include the interpretation of the results and the implications for future research. I intend to discuss how the research results could have implications for assessment practices in undergraduate mathematics.

Using the Quality Index model, as developed in section 5.3, I will illustrate which items can be classified as good or poor quality mathematics questions.   A comparison of good and poor quality mathematics questions in each of the PRQ and CRQ assessment formats will be made.   Furthermore, I draw conclusions from my research about which of the mathematics assessment components, as defined in section 5.1, can be successfully assessed with respect to each of the two assessment formats, PRQ and CRQ.

In this way, I endeavour to probe and clarify the first two research subquestions as stated in section 3.2 i.e. How do we measure the quality of a good mathematics question? and; 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?

7.1   GOOD AND POOR QUALITY MATHEMATICS QUESTIONS

Section 7.1 summarises the development and features of the QI model for the sake of completeness of this chapter.

In section 5.3, the Quality Index (QI) was defined in terms of the three measuring criteria: discrimination, confidence deviation and expert opinion deviation. Each of these three criteria represented the three arms of a radar plot. In the proposed QI model, all three criteria were considered to be equally important in their contribution to the overall quality of a question.

The QI model can be used both to quantify and visualise how good or how poor the quality of a mathematics question is. The following three features of the radar plots could assist us to visualise the quality and the difficulty of the item:

(1) the shape of the radar plot;

(2) the area of the radar plot;

(3) the shading of the radar plot.

1. Shape of the radar plot

When comparing the radar plots for the good quality items with those of the poor quality items, it is evident that the shapes of these radar plots are also very different. For the good mathematics questions, the shape seems to resemble a small equilateral triangle. This ideal shape is achieved when all three arms of the radar plot are shorter than the average length of 0.5 on each axis i.e. are all very close to 0, as well as all three arms being almost equal in magnitude. Such a situation would be ideal for a mathematics question of good quality, since all three measuring criteria would be close to zero which indicates a small deviation from the expected confidence level as well as a small deviation from the expected student performance, and would also indicate an item that discriminates well. In contrast, those radar plots corresponding to items of a poor quality did not display this small equilateral triangular shape. One notices that these radar plots are skewed in the direction of one or more of the three axes. This skewness in the shape of the radar plot reflects that the three measuring criteria do not balance each other out. The axis towards which the shape is skewed reflects which of the criteria contribute to the overall poor quality of the question. However, there are poor quality items which have radar plots resembling the shape of a large equilateral triangle. The difference is that although the plot has three arms equal in magnitude, all three arms are longer than the average length of 0.5 and are in fact all very close to 1 (i.e. very far from 0).

2. Area of the radar plot

Another visual feature of the radar plot is its area. In this study, the area of the radar plot represents the Quality Index (QI) of the item. By defining the QI as the area, a balance is obtained between the three measuring criteria. If the QI value is less than 0.282 (the median QI), then the question is classified as a good quality mathematics question. If the QI value is greater than or equal to 0.282, the question is considered to be of a poor quality. When investigating the area of the good quality items, it is evident that such items have a small area i.e. a QI value close to zero. In such radar plots, the three arms are all shorter than the average length of 0.5 on each axis, and are all close to 0. For the poor quality items, the corresponding radar plot has a large area with QI values far from 0 (i.e. close to 1). In such radar plots, the three arms are generally longer than the average length of 0.5 on each axis, and are all far away from 0. The closer the QI value is to 0, the better the quality of the question.

We can conclude that both the area and the shape of the radar plot assist us to form an opinion on the quality of a question.

In Figure 7.1, both the shape and the area of the radar plot indicate a good quality assessment item. The shape resembles an equilateral triangle and the area is small.

Figure 7.2 visually illustrates an assessment item of poor quality. The shape is skewed in the direction of both the discrimination and confidence axes and the radar plot has a large area. The poor performance of all three measuring criteria contributes to this item being a poor quality item. The item does not discriminate well and both students and experts misjudged the difficulty of the question. The large, skewed shape of the radar plot indicates an item of poor quality.

Figure 7.1: A good quality item.   **Figure 7.2:** A poor quality item.
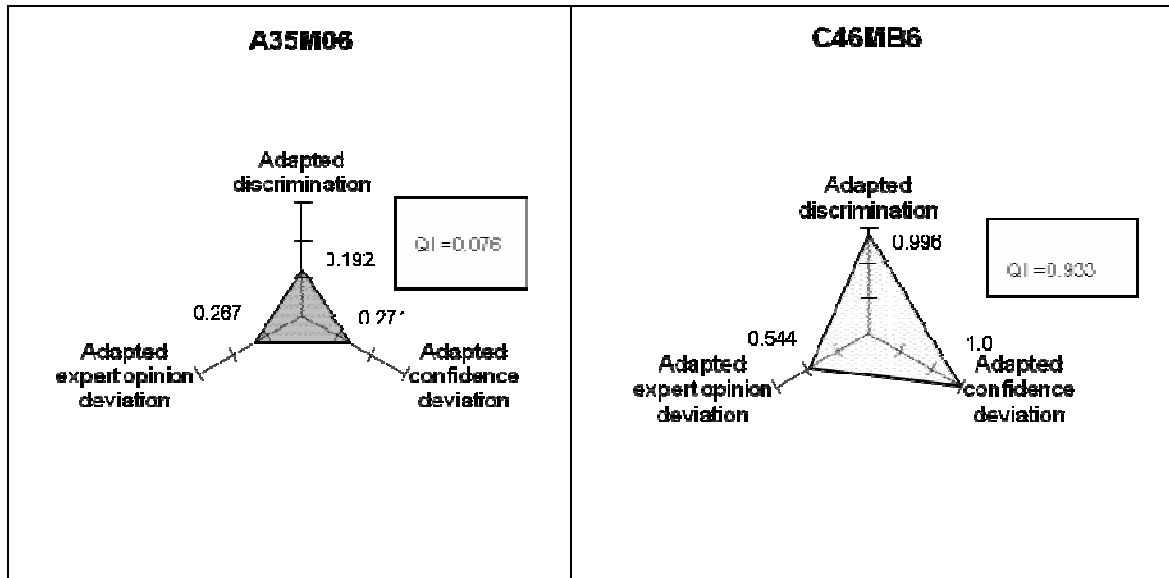
3.   Shading of the radar plot

In this study, the shading of the radar plot helped us to visualise the difficulty level of the question. Six shades of grey, ranging from white through to black (as shown in Table 5.4), represented the six corresponding difficulty levels chosen in this study ranging from very easy through to very difficult. Difficulty level is an important parameter, but does not contribute to classifying a question as good or not. Both easy questions and difficult questions can be classified as good or poor. Not all difficult questions are of a good quality, and not all easy questions are of a poor quality. For example, in Figure 7.3, the dark grey shading of the radar plot represents a difficult item. The large area and skew shape of the plot represents a poor quality item. So Figure 7.3 visually represents a difficult, poor quality item. In Figure 7.4, the very light shading of the radar plot represents an easy item. The small area and shape of the radar plot represents a good quality item. So Figure 7.4 visually represents an easy, good quality item.

Figure 7.3: A difficult, poor quality item.   **Figure 7.4:** An easy, good quality item.
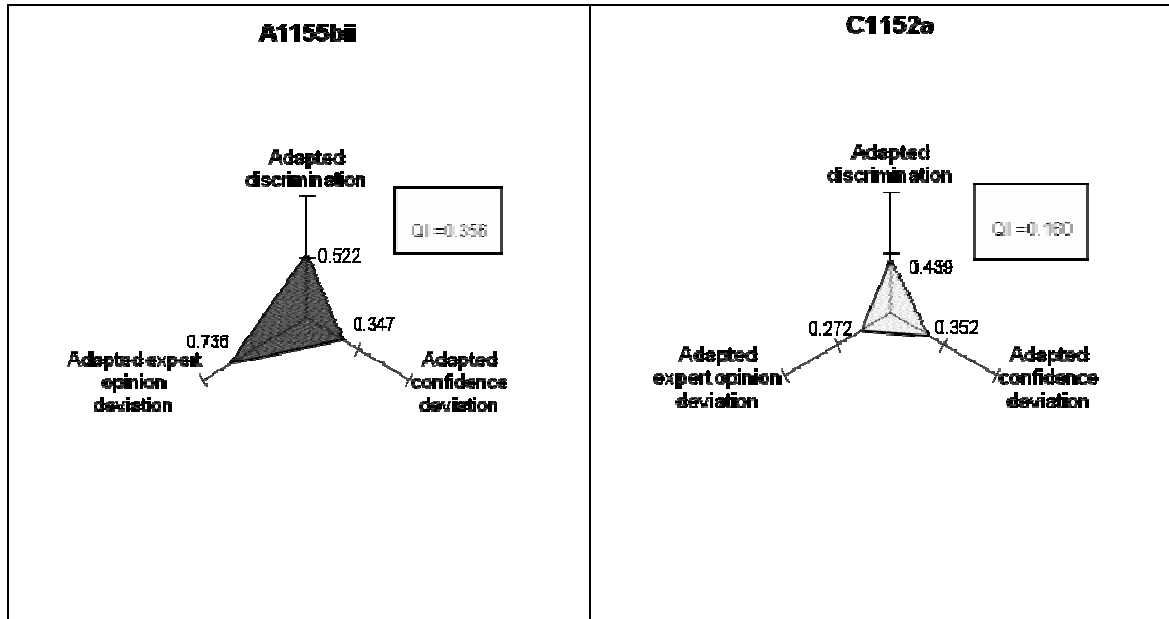
7.2   A COMPARISON OF PRQs AND CRQs IN THE MATHEMATICS ASSESSMENT COMPONENTS

In section 6.3, Table 6.3 summarised the quality of both PRQs and CRQs within each assessment component. It was noted that certain assessment components lend themselves better to PRQs than to CRQs. For example, in the **technical assessment component**, there were almost twice as many good quality PRQs than good quality CRQs.   For the assessor, this means that the PRQ assessment format can be successfully used to assess mathematics content which requires students to adopt a routine, surface learning approach.   In this component, PRQs can successfully assess content which students will have been given in lectures or will have practised extensively in tutorials. In addition there were more than twice as many poor quality CRQs than poor quality PRQs. The conclusion is that the PRQ format successfully assesses cognitive skills such as manipulation and calculation, associated with the technical assessment component.

Another component in which PRQs can be used successfully is the **disciplinary assessment component**. In this component, there was no difference between the good quality PRQs and the poor quality PRQs, with very little difference between the good quality CRQs and the poor quality CRQs. The PRQ format can be used to assess cognitive skills involving recall (memory) and knowledge (facts) equally successfully as the CRQ format. Thus in the disciplinary assessment component, results show that it is easy to set PRQs of a good quality, thus saving time in both the setting and marking of questions involving knowledge and recall.

As we proceed to the higher order **conceptual assessment component**, it is once again encouraging that the results indicate that PRQs can hold more than their own against CRQs. PRQs could be used successfully as a format of assessment for tasks involving comprehension skills whereby students are required to apply their learning to new situations or to present information in a new or different way. The results challenge the viewpoint of Berg and Smith (1994) that PRQs cannot successfully assess graphing abilities. The shift away from a surface approach to learning to a deeper approach, as mentioned by Smith *et al.* (1996), can be just as successfully assessed with PRQs as with the more traditional open-ended CRQs. The conclusion is that the PRQ assessment format can be successfully used in the conceptual assessment component.

The **modelling assessment component** tasks, requiring higher order cognitive skills of translating words into mathematical symbols, have traditionally been assessed using the CRQ format. The results from this study show that although there are few PRQs corresponding to this component, probably due to the fact that it is more difficult to set PRQs than CRQs of a modelling nature, the PRQs were highly successful. The perhaps somewhat surprising conclusion is that PRQs can be used very successfully in the modelling component. This result disproves the claim made by Gibbs (1992) that one of the main disadvantages of PRQs is that they do not measure the depth of student thinking. It also puts to rest the concern expressed by Black (1998) and Resnick & Resnick (1992) that the PRQ assessment format encourages students to adopt a surface

learning approach. Although PRQs are more difficult and time consuming to set in the modelling assessment component (Andresen *et al.*, 1993), these results encourage assessors to think more about our attempts at constructing PRQs which require words to be translated into mathematical symbols. The results show that there is no reason why PRQs cannot be authentic and characteristic of the real world, the very objections made by Bork (1984) and Fuhrman (1996) against the whole principle of the PRQ assessment format.

Another very encouraging result was the high percentage of good quality PRQs as opposed to poor quality PRQs in the **problem solving assessment component**. This component encompasses tasks requiring the identification and application of a mathematical method to arrive at a solution. It appears that PRQs are slightly more successful than CRQs in this assessment component which encourages a deep approach to learning. Greater care is required when setting problem-solving questions, whether PRQs or CRQs, but the results show that PRQ assessment can add value to the assessment of the problem solving component. Once again this result shows that PRQs do not have to be restricted to the lower order cognitive skills so typical of a surface approach to learning (Wood & Smith, 2002).

The results indicate that PRQs were not as successful in the logical and consolidation assessment components. In the **logical assessment component**, there were noticeably more poor quality PRQs than poor quality CRQs. The nature of the tasks involving ordering and proofs lends itself better to the CRQ assessment format. There were very few good PRQs in the logical assessment component. The high percentage of the poor quality PRQs in the logical assessment component leads to the conclusion that this component lends itself better to CRQs than to PRQs.

In the **consolidation assessment component**, involving cognitive skills of analysis, synthesis and evaluation, there were noticeably more good quality CRQs than good quality PRQs. This trend towards more successful CRQs than PRQs indicates that CRQs add more value to the assessment of this

241

component. This is not an unexpected result, as at this highest level of conceptual difficulty, assessment tasks require students to display skills such as justification, interpretation and evaluation. Such skills would be more difficult to assess using the PRQ format. However, as shown by many authors (Gronlund, 1988; Johnson, 1989; Tamir, 1990), the 'best answer' variety in contrast to the 'correct answer' variety of PRQs does cater for a wide range of cognitive abilities. In these alternative types of PRQs the student is faced with the task of carefully analysing the various options and of making a judgement to select the answer which best fits the context and the data given. The conclusion is that the consolidation assessment component encourages the educator or assessor to think more about their attempts at constructing suitable assessment tasks. According to Wood and Smith (2002), assessment tasks corresponding to a high level of conceptual difficulty should provide a useful check on whether we have tested all the skills, knowledge and abilities that we wish our students to demonstrate. As the results have shown, PRQs can be used as successfully as CRQs as an assessment method for those mathematics assessment components which require a deeper learning approach for their successful completion.

7.3   CONCLUSIONS

The mathematics assessment component taxonomy, proposed by the author in section 5.1, is hierarchical in nature, with cognitive skills that need a surface approach to learning at one end, while those requiring a deeper approach appear at the other end of the taxonomy. The results of this research study have shown that it is not necessary to restrict the PRQ assessment format to the lower cognitive tasks requiring a surface approach. The PRQ assessment format can, and does add value to the assessment of those components involving higher cognitive skills requiring a deeper approach to learning. According to Smith *et al.* (1996), many students enter tertiary institutions with a surface approach to learning mathematics and this affects their results at university. The results of this research study have addressed the research question of whether we can successfully use PRQs as an assessment format in

undergraduate mathematics and the mathematics assessment component taxonomy was proposed to encourage a deep approach to learning. In certain assessment components, PRQs are more difficult to set than CRQs, but this should not deter the assessor from including the PRQ assessment format within these assessment components. As the discussion of the results has shown, good quality PRQs can be set within most of the assessment components in the taxonomy which do promote a deeper approach to learning.

In the Niss (1993) model, discussed in section 2.3, the first three content objects require knowledge of facts, mastery of standard methods and techniques and performance of standard applications of mathematics, all in typical, familiar situations. Results of this study have shown that PRQs are highly successful as an assessment format for Niss's first three content objects. As we proceed towards the content objects in the higher levels of Niss's assessment model, students are assessed according to their abilities to activate or even create methods of proofs; to solve open-ended, complex problems; to perform mathematical modelling of open-ended real situations and to explore situations and generate hypotheses. Results of this study again show that even though PRQs are more difficult to set at these higher cognitive levels, they can add value to the assessment at these levels.

Results of this study show that the more cognitively demanding conceptual and problem solving assessment components are better for CRQs. Traditional assessment formats such as the CRQ assessment format have in many cases been responsible for hindering or slowing down curriculum reform (Webb & Romberg, 1992). The PRQ assessment format can successfully assess in a valid and reliable way, the knowledge, insights, abilities and skills related to the understanding and mastering of mathematics in its essential aspects. As shown by the qualitative results, PRQs can provide assistance to the learner in monitoring and improving his/her acquisition of mathematical insight and power, while also improving their confidence levels. Furthermore, PRQs can assist the educator to improve his/her teaching, guidance, supervision and counselling, while also saving time. The PRQ assessment format can reduce marking loads

for mathematical educators, without compromising the value of instruction in any way. Inclusion of the PRQ assessment format into the higher cognitive levels would bring new dimensions of validity into the assessment of mathematics.

Table 7.1 presents a comparison of the success of PRQs and CRQs in the mathematics assessment components.

Table 7.1: A comparison of the success of PRQs and CRQs in the mathematics assessment components.

Mathematics assessment Component	Comparison of success
1. Technical	PRQs can be used successfully
2. Disciplinary	No difference
3. Conceptual	PRQs can be used successfully
4. Logical	CRQs more successful
5. Modelling	PRQs can be used successfully
6. Problem solving	PRQs can be used successfully
7. Consolidation	CRQs more successful

As Table 7.1 illustrates, the enlightening conclusion is that there are only two components where CRQs outperform PRQs, namely the logical and consolidation assessment components. In two other components, PRQs are observed to slightly outperform CRQs, namely the conceptual and problem solving assessment components. The PRQs outperform the CRQs substantially in the technical and modelling assessment components. In one component there is no observable difference, the disciplinary assessment component.

7.4   ADDRESSING THE RESEARCH QUESTIONS

In this study, a model has been developed to measure the quality of a mathematics question. This model, referred to as the Quality Index (QI) model, was used to address the research question and subquestions as follows:

Research question:

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

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?

• Addressing the first subquestion:

There is no single way of measuring the quality of a good question. I, as author of the thesis, have proposed one model as a measure of the quality of a question. I have illustrated the use of this model and found it to be an effective and quantifiable measure.

The QI model can assist mathematics educators and assessors to judge the quality of the mathematics questions in their assessment programmes, thereby deciding which of their questions are good or poor. Retaining unsatisfactory questions is contrary to the goal of good mathematics assessment (Kerr, 1991). Mathematics educators should optimise both the quantity and the quality of their assessment, and thereby optimise the learning of their students (Romberg, 1992).

The QI model for judging how good a mathematics question is has a number of apparent benefits. The model is visually satisfying; whether a question is of good or poor quality can be witnessed at a single glance. Visualising the difficulty level in terms of shades of grey adds convenience to the model. Another visual advantage of this model is that shortcomings in different aspects of an item, such as that experts completely under estimate the expected level of student performance in the particular item, can also be instantly visualised. In addition, the model provides a quantifiable measure of the quality of a question, an aspect that makes the model useful for comparison purposes. The fact that the model can be applied to judge the level of difficulty of both PRQs and CRQs makes it useful for both traditional "long question" environments, as well as the increasingly popular online, computer centred environments.

- ## Addressing the second subquestion:

In terms of the mathematics assessment components, it was noted that certain assessment components lend themselves better to PRQs than to CRQs. In particular, the PRQ format proved to be more successful in the technical, conceptual, modelling and problem solving assessment components, with very little difference in the disciplinary component, thus representing a range of assessment levels from the lower cognitive levels to the higher cognitive levels. Although CRQs proved to be more successful than PRQs in the logical and consolidation assessment components, PRQs can add value to the assessment of these higher cognitive component levels. Greater care is needed when setting PRQs in the logical and consolidation assessment components. The inclusion of the PRQ format in all seven assessment components can reduce marking loads for mathematics educators, without compromising the validity of the assessment. The PRQ assessment format can successfully assess in a valid and reliable way. The results have shown, both quantitatively and qualitatively, that PRQs can improve students' acquisition of mathematical insight and knowledge, while also improving their confidence levels. The PRQ assessment format can be used as successfully as the CRQ format to encourage students to adopt a deeper approach to the learning of mathematics.

- Addressing the third subquestion:

With respect to the student preferences regarding different mathematics assessment formats, the results from the qualitative investigation seemed to indicate that there were two distinct camps; those in favour of PRQs and those in favour of CRQs. Those in favour of PRQs expressed their opinion that this assessment format did promote a higher conceptual level of understanding and greater accuracy; required good reading and comprehension skills and was very successful for diagnostic purposes. Those in favour of CRQs were of the opinion that this assessment format promoted a deeper learning approach to mathematics; required good reading and comprehension skills; partial marks could be awarded for method and students felt more confident with this more traditional approach. Furthermore, from the students' responses, it also seemed as if the weaker ability students preferred the CRQ assessment format above the PRQ assessment format. The reasons for this preference were varied: CRQs provide for partial credit; there was a greater confidence with CRQs than with PRQs; PRQs require good reading and comprehension skills; PRQs encourage guessing and the distracters cause confusion.

- Addressing the main research question:

As this study aimed to show, PRQs can be constructed to evaluate higher order levels of thinking and learning, such as integrating material from several sources, critically evaluating data and contrasting and comparing information. The conclusion is that PRQs can be successfully used as an assessment format in undergraduate mathematics, more so in some assessment components than in others.

7.5   LIMITATIONS OF STUDY

The tests used in this study were conducted with tertiary students in their first year of study at the University of the Witwatersrand, Johannesburg, enrolled for the mainstream Mathematics I Major course. The study could be extended to other tertiary institutions and to mathematics courses beyond the first year level.

The judgement of how good or poor a mathematics question is, is modulo the QI model developed in this study. In the proposed QI model, I assumed that the three arms of the radar plot contribute equally to the overall quality of the mathematics question. This assumption needs to be investigated.

The qualitative component of this study was not the most important part of the research. The small sample of students interviewed was carefully selected to include differences in mathematical ability, from different racial backgrounds and different gender classes. Consequently, I regarded their responses as being indicative of the opinions of the Mathematics I Major cohort of students. The third research subquestion, dealing with student preferences regarding the different assessment formats, was included as a small subsection of the study and was not the main focus of this study. The qualitative component could be expanded in future by increasing the sample size of interviewees and by using questionnaires in which all the students in the first year mathematics major course could be asked to express their feelings and opinions regarding different mathematics assessment formats.

7.6 IMPLICATIONS FOR FURTHER RESEARCH

Collection of confidence-level data in conceptual mathematics tests provides valuable information about the quality of a mathematics question. The analysis suggests that confidence of responses should be collected, but also that it is critical to consider not only students' overall confidence but to consider separately confidence in both correct and incorrect answers. The prevalence of overconfidence in the calibration of performance presents a paradox of educational practice.

> On the one hand, we want students to have a healthy sense of academic self-concept and persist in their educational endeavours. On the other hand, we hope that a more realistic understanding of their limitations will be the impetus for educational development. The challenge for educators is to implement constructive interventions that lead to improved calibration and performance

without destroying students' self-esteem and confidence (Bol & Hacker, 2008, p2).

In this study, three parameters were identified to measure the quality of a mathematics question: discrimination index, confidence index and expert opinion. Further work needs to be carried out to investigate whether more contributing measuring criteria can be identified to measure the overall quality of a good mathematics question, and how this would affect the calculation of the Quality Index (QI) as discussed in section 5.3.2. As the assumption was made that the three parameters contributed equally to the quality of a mathematics question, the QI was defined as the area of the radar plot. The QI model could be adjusted or refined using other formulae.

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 in a subject. According to Planinic *et al.* (2006), misleading and even incorrect results can stem from an erroneous assumption that raw scores are in fact linear measures. Rasch analysis, the statistical method used in this research, is a technique that enables researchers to look objectively at data. The Rasch model (1960), can provide linear measures of item difficulties and students' confidence levels. Often, analysis of raw test score data or attitudinal data is carried out, but it is not always the case that such raw scores can be immediately assumed to be linear measures, and linear measures facilitate objective comparison of students and items (Planinic *et al.* 2006). According to Wright and Stone (1979), the Rasch model is a more precise and moral technique that can be used to comment on a person's ability and that the introduction thereof is long overdue. The Rasch method of data analysis could be valuable for other researchers in the fields of mathematics and science education research.

It might be important for mathematics educators and researchers to further explore the QI model with questions not limited to Calculus and Linear Algebra topics of many traditional first year tertiary mathematics courses. In doing so, mathematics educators and assessors can be provided with an important model

to improve the overall quality of their assessment programmes and enhance student learning in mathematics.

This research study could be expanded to other universities. Tertiary mathematics educators need to use models of the type developed in this study to quantify the quality of the mathematics questions in their undergraduate mathematics assessment programmes. The QI model can also be used by tertiary mathematics educators to design different formats of assessment tasks which will be significant learning experiences in themselves and will provide the kind of feedback that leads to success for the individual student, thus reinforcing positive attitudes and confidence levels in the students' performance in undergraduate mathematics.

> The way students are assessed influences what and how they learn more than any other teaching practice (Nightingale *et al.*, 1996, p7).

Good quality assessment of students' knowledge, skills and abilities is crucial to the process of learning. In this research study, I have shown that the more traditional CRQ format is not always the only and best way to assess our students in undergraduate mathematics. PRQs can be constructed to evaluate higher order levels of thinking and learning. The research study conclusively shows that the PRQ format can be successfully used as an assessment format in undergraduate mathematics.

As mathematics educators and assessors, we need to radically review our assessment strategies to cope with changing conditions we have to face in South African higher education.

> The possibility that innovative assessment encourages students to take a deep approach to their learning and foster intrinsic interest in their studies is widely welcomed (Brown & Knight, 1994, p24).

REFERENCES

Adkins, D.C. (1974). *Test construction: development and interpretation of achievement tests* (2nd ed.). Columbus, Otl: Charles Merrill Publishing.

Adler, J. (2001). *Teaching mathematics in multilingual classrooms.* Dordrecht: Kluwer Academic Publishers.

Aiken, L.R. (1987). Testing with multiple-choice items. *Journal of Research and Development in Education, 20*, 44-58.

American Psychological Association (1963). Ethical standards of psychologists. *American Psychologist, 23,* 357-361.

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andersen, E.B. & Olsen, L.W. (1982). The life of Georg Rasch as a mathematician and as a statistician. In A. Boomsma, M.A.J. van Duijn & T.A.B. Sniders (Eds.), *Essays in item response theory.* New York: Springer.

Anderson, J.R. (1995). *Cognitive psychology and its implications* (4th ed.). W.H. Freeman Publishers.

Andresen, L., Nightingale, P., Boud, D. & Magin, D. (1993). *Strategies for assessing students.* Birmingham: SCED.

Andrich, D. (1982). An index of person separation in latent trait theory the traditional KR.20 index, and the Guttman scale response pattern. *Educational Research and Perspectives, UWA, 9(1),* 95-104.

Andrich, D. (1988). *Rasch models for measurements*. USA: Sage Publications, Inc.

Andrich, D. & Marais, I. (2006). EDU435/635. *Instrument Design with Rasch IRT and Data Analysis 1, Unit Materials - Semester 2.* Perth, Western Australia: Murdoch University.

Angel, S.A. & LaLonde, D.E.   (1998). Science success strategies: An interdisciplinary course for improving science and mathematics education. *Journal of Chemical Education, 75(11)*, 1437-41.

Angrosino, M.V. & Mays de Pérez, K.A. (2000). Rethinking observation: From method to context. In N.K. Denzin & Y.S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed.) (pp. 673-702). Thousand Oaks, CA: Sage.

Anguelov, R., Engelbrecht J. & Harding, A. (2001). Use of technology in undergraduate mathematics teaching in South African universities. *Quaestiones Mathematicae*, *Suppl.* 1, 183-191.

Astin, A.W. (1991). *Assessment for excellence.* New York: Macmillan.

Aubrecht II, G.J. & Aubrecht, J.D. (1983). Constructing objective tests. *Am. J. Phys., 51(7),* 613-620.

Baker, L. & Brown, A. (1984). Metacognitive skills and reading. In P.D. Pearson, M. Kamil, R. Barr & P. Rosenthal (Eds.), *Handbook of reading research* (pp. 353-394). New York: Longman.

Ball, G., Stephenson, B., Smith, G.H., Wood, L.N., Coupland, M. & Crawford, K. (1998). Creating a diversity of experiences for tertiary students. *Int. J. Math. Educ. Sci. Technol.*, *29(6)*, 827-841.

Baron, M.A. & Boschee, F. (1995). Outcome-based education: Providing direction for performance-based objectives. *Educational Planning, 10(2),* 25-36.

Barak, M. & Rafaeli, S. (2004). On-line question-posing and peer-assessment as means for web-based knowledge sharing in learning. *Int. J. Human – Computer Studies, 61,* 84-103.

Begle, E.G. & Wilson, J.W. (1970). Evaluation of mathematics programs. In E.G. Begle (Ed.), *Mathematics Education* (69th Yearbook of the National Society for the study of Education, Part I, 376-404). Chicago: University of Chicago Press.

Beichner, R. (1994). Testing student interpretation of kinematics graphs. *American Journal of Physics, 62,* 750-762.

Berg, C.A. & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education, 78,* 527-554.

Biggs, J. & Collis, N.F. (1982). Mathematics Profile Series Operations Test. In J.B. Biggs (Ed.), *Evaluating the quality of learning: the SOLO Taxonomy* (pp. 82-89). New York: Academic Press.

Biggs, J. (1991). Student learning in the context of school. In J. Biggs (Ed.), *Teaching for learning: the view from cognitive psychology* (pp. 7-20). Hawthorn, Victoria: Australian Council for Educational Research.

Biggs, J. (1994). Learning outcomes: competence or expertise? *Australian and New Zealand Research*, *2(1)*, 1-18.

Biggs, J. (2000). *Teaching for quality learning at university*. Buckingham: Open University Press.

Birenbaum, M. & Dochy, F. (1996). *Alternatives in assessment of achievements, learning processes and prior knowledge.* Boston: Kluwer Academic Publishers.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Black, P. (1998). *Testing: friend or foe? Theory and practice of assessment and testing.* London: Falmer Press.

Blanton, H., Buunk, B.P., Gibbons, F.X. & Kuyper, H. (1999). When better-than-others compare upward: Choice of comparison and comparative evaluation as independent predictors of academic performance. *Journal of Personality and social Psychology 76,* 420-430.

Bless, C. & Higson-Smith, C. (1995). *Fundamentals of social research methods: An African perspective.* Boston: Allan & Bacon.

Bloom, B.S. (Ed.) (1956). *Taxonomy of educational objectives. The classification of educational goals. Handbook 1:* The cognitive domain. New York: David McKay.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Bol, L. & Hacker, D.J. (2008). Focus on research: Understanding and improving calibration accuracy.
Retrieved on 1 March, 2007 from http://uhaweb.hartford.edu/ssrl/research.htm

Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah N J: Erlbaum Assoc.

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

Bork, A. (1984). "Letter to the Editor". *Am. J. Phys., 52,* 873-874.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in higher education, 15(11),* 101-111.

Boud, D. (1995). *Enhancing learning through self-assessment.* London: Kogan Page.

Braswell, J.S. & Jackson, C.A. (1995). *An introduction of a new free-response item type in mathematics.* Paper presented at the Annual meeting of the National Council on Measurement in Education. San Francisco: CA.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice format. *Journal of Educational Measurement, 29,* 253-271.

Brown, G., Bull, J. & Pendlebury, M. (1997). *Assessing student learning in higher education.* New York: Routledge.

Brown, S. & Knight, P. (1994). *Assessing learners in higher education.* London: Kogan Page.

Brown, S. (1999). Institutional strategies for assessment. In S. Brown & A. Glasner (Eds.), *Assessment matter in higher education. Choosing and using diverse approaches* (pp. 3-13). Buckingham: Open University Press.

Burns, N. & Grove, S.K. (2003). *Understanding nursing research* (3rd ed.). Philadelphia: W.B. Saunders Company.

California Mathematics Council (CMC) and EQUALS. (1989). *Assessment alternatives in mathematics: An overview of assessment techniques that promote learning.* University of California, Berkeley: CMC and EQUALS.

Campione, J.C., Brown, A.L. & Connell, M.L. (1988). Metacognition: On the importance of understanding what you are doing. In R.I. Charles & E.A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 93-114). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carvalho, M.K. (2007). Confidence judgments in real classroom settings: Monitoring performance in different types of tests. *International Journal of Pyschology,* 1-16.

Case, S.M. & Swanson, D.B. (1989). Strategies for student assessment. In Boud, D. & Feletti, G. (Eds.), *The challenge of problem-based learning* (pp 269-283). London: Kogan Page.

Collis, K.F. (1987). Levels of reasoning and the assessment of mathematical performance. In T.A. Romberg & D.M. Stewart (Eds.), *The monitoring of school mathematics: Background papers.* Madison: Wisconsin Center for Education Research.

Corcoran, M. & Gibb, E.G. (1961). Appraising attitudes in the learning of mathematics. In *Yearbook (1961) – National Council of Teachers of Mathematics.* Reston, VA: NCTM.

Cresswell, J.W. (1998). *Qualitative inquiry and research design: Choosing among five traditions.* Thousand Oaks, CA: Sage.

Cresswell, J.W. (2002). *Educational Research: Planning, conducting and evaluating quantitative and qualitative research.* Upper Saddle River, New Jersey: Pearson Education, Inc.

Cretchley, P.C. (1999). An argument for more diversity in early undergraduate mathematics assessment. *Delta: 1999. The Challenge of Diversity*, 17-80.

Cretchley, P.C. & Harman, C.J. (2001). Balancing the scales of confidence – computers in early undergraduate mathematics learning. *Quaestiones Mathematicae*, *Suppl. 1*, 17-25.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58(4),* 43-81.

Cumming, J.J. & Maxwell, G.S. (1999). Contextualising authentic assessment. *Assessment in Education, 6(2),* 177-194.

Dahlgren, L. (1984). Outcomes of learning. In F. Marton, D. Hounsell & N. Entwistle (Eds.), *The experience of learning.* Edinburgh: Scottish Academic Press.

De Lange, J. (1994). Assessment: No change without problems. In T.A Romberg (Ed.), *Reform in School Mathematics and authentic assessment* (pp. 87-172). Albany NY: SUNY Press.

Dison, L. & Pinto, D. (2000). Example of curriculum development under the South African National Qualifications Framework. In S. Makoni (Ed.), *Improving teaching and learning in higher education. A handbook for Southern Africa* (pp. 201-202). Johannesburg, South Africa: Wits University Press.

Ebel, R. (1965). Confidence weighting and test reliability. *Journal of Educational Mesurement, 2,* 49-57.

Ebel, R. (1972). *Essentials of educational measurement.* New York: Prentice Hall.

Ebel, R. & Frisbie, D.A. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Ehrlinger, J. (2008). Skill level, self-views and self-theories as sources of error in self-assessment. *Social and Personality Psychology Compass, 2(1),* 382-398.

Eisenberg, T. (1975). Behaviorism: The bane of school mathematics. *Journal of Mathematical Education, Science and Technology, 6(2),* 163-171.

Elton, L. (1987). *Teaching in higher education:* Appraisal and training. London: Kogan Page.

Engelbrecht, J. & Harding, A. (2002). Is mathematics running out of numbers? *South African Journal of Science, 99(1/2),* 17-20.

Engelbrecht, J. & Harding, A. (2003). Online assessment in mathematics: multiple assessment formats. *New Zealand Journal of Mathematics, 32 (Supp.)*, 57-66.

Engelbrecht, J. & Harding, A. (2004). Combining online and paper assessment in a web-based course in undergraduate mathematics. *Journal of computers in Mathematics and Science Teaching, 23(3),* 217-231.

Engelbrecht, J., Harding, A. & Potgieter, M. (2005). Undergraduate students' performance and confidence in procedural and conceptual mathematics. *Int. J. Math. Educ. Sci. Technol., 36(7),* 701-712.

Engelbrecht, J. & Harding, A. (2006). Impact of web-based undergraduate mathematics teaching on developing academic maturity: A qualitative investigation. *Proceedings of the 8th Annual Conference on WWW Applications.* Bloemfontein, South Africa.

Entwistle, N. (1992). *The impact of teaching on learning outcomes in higher education: A literature review.* Sheffield: Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom, Universities' Staff Development Unit.

Erwin, T.D. (1991). *Assessing student learning and development: A guide to the principles, goals and methods of determining college outcomes.* San Francisco: Jossey-Bass.

Freeman, J. & Byrne, P. (1976). *The assessment of postgraduate training in general practice* (2nd ed.). Surrey: SRHE.

Freeman, R. & Lewis, R. (1998). *Planning and implementing assessment.* London: Kogan Page.

Friel, S. & Johnstone, A.H. (1978). Scoring systems which allow for partial knowledge. *Journal of Chemical Education, 55,* 717-719.

Fuhrman, M. (1996). Developing good multiple-choice tests and test questions. *Journal of Geoscience Education, 44,* 379-384.

Gall, M.D., Gall, J.P. & Borg, W.R. (2003). *Educational Research: an introduction* (7th ed.). USA: Pearson Education Inc.

Gay, S. & Thomas, M. (1993). Just because they got it right, does not mean they know it? In N.L. Webb and A.F. Coxford (Eds.), *Assessment in the mathematics classroom.* Reston, VA: NCTM.

Geyser, H. (2004). Learning from assessment. In S. Gravett & H. Geyser. (Eds.), *Teaching and learning in higher education* (pp. 90-110). Pretoria, South Africa: Van Schaik.

Gibbs, G. (1992). *Assessing more students.* Oxford: The Oxford Centre for Staff Development.

Gibbs, G., Habeshaw, S. & Habeshaw, T. (1988). *53 interesting ways to assess your students* (2nd ed.). Bristol: Technical and Educational Services Ltd.

Gifford, B.R. & O'Connor, M.C. (1992). *Changing assessments: Alternative views of aptitude, achievement and instruction.* Boston and Dordrecht: Kluwer.

Glaser, R. (1988). Cognitive and environmental perspectives on assessing achievement. In E. Freeman (Ed.), *Assessment in the service of learning: Proceedings of the 1987 ETS Invitational Conference* (pp. 40-42). Princeton, N.J.: Educational Testing Service.

Glass, G.V. & Stanley, J.C. (1970). Measurement, scales and statistics. *Statistical methods in education and psychology,* (pp. 7-25). New Jersey: Prentice Hall.

Greenwood, L., McBride, F., Morrison, H., Cowan, P. & Lee, M. (2000). Can the same results be obtained using computer-mediated tests as for paper-based tests for National Curriculum assessment? *Proceedings of the International Conference in Mathematics/Science Education and Technology, 2000(1),* 179-184.

Groen, L. (2006) Enhancing learning and measuring learning outcomes in mathematics using online assessment. *UniServe Science Assessment Symposium Proceedings,* 56-61.

Gronlund, N.E. (1976). *Measurement and evaluation in teaching* (3rd ed.). New York: Macmillan.

Gronlund, N.E. (1988). *How to construct achievement tests.* Englewood Cliffs, NJ: Prentice Hall.

Haladyna, T.M. (1999). *Developing and validating multiple choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Hamilton, L.S. (2000). Assessment as a policy tool. *Review of Research in Education, 27(1),* 25-68.

Harlen, W. & James, M.J. (1977). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education, 4(3),* 365-380.

Harper, R. (2003). Correcting computer-based assessments for guessing. *Journal of Computer Assisted Learning, 19(1),* 208.

Harper, R. (2003). Multiple choice questions – a reprieve. *Bioscience Education e-Journal, 2.*
Retrieved on 18 May, 2004 from http://bio.ltsn.ac.uk/journal/vol1/beej-2-6.htm

Harvey, J.G. (1992). Mathematics testing with calculators: ransoming the hostages. In T.A. Romberg (Ed.). *Mathematics assessment and evaluation: Imperatives for mathematics education* (pp. 139-168)*.* Albany, NY: Suny Press.

Harvey, L. (1993). An integrated approach to student assessment. Paper presented to *Measure for Measure,* Act III conference, Warwick.

Hasan, S., Bagayoko, D. & Kelley, E.L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education, 34(5)*, 294-299.

Heywood, J. (1989). *Assessment in higher education.* London: Kogan Page.

Hibberd, S. (1996). The mathematical assessment of students entering university engineering courses. *Studies in Educational Evaluation*, *22(4)*, 375-384.

Hiebert, J. & Carpenter, T.P. (1992). Learning and teaching with understanding. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 97-111). New York: Macmillan.

Hoffman, B. (1962). *The tyranny of testing.* New York: Greenwood Press.

Hounsell, D., McCulloch, M. & Scott, M. (Eds.) (1996). *The ASSHE Inventory: Changing assessment practices in Scottish higher education.* Sheffield: UCOSDA.

Hubbard, R. (1995). *53 ways to ask questions in mathematics and statistics.* Bristol: Technical and Educational Services.

Hubbard, R. (1997). Assessment and the process of learning statistics. *Journal of Statistics Education, 5(1).*
Retrieved on 17 June, 2007 from
http://www.amstat.org/publications/jse/v5n1/hubbard.html

Hubbard, R. (2001). The why and how of getting rid of conventional examinations. *Quaestiones Mathematicae, Suppl. 1,* 57-64.

Hughes, C. & Magin, D. (1996). Demonstrating knowledge and understanding. In P. Nightingale (Ed.), *Assessing learning in universities* (pp. 127-161) Sydney: University of New South Wales Press.

Huysamen, G.K. (1983). *Introductory statistics and research design for the behavioural sciences,* Volume 1. Bloemfontein: Department of Psychology, UOFS.

Isaacs, G. (1994). *Multiple choice testing: A guide to the writing of multiple choice tests and to their analysis*. Campbelltown, NSW: HEROSA.

Isaacson, R.M. & Fujita, F. (2006). Metacognitive knowledge monitoring and self-regulated learning: Academic success and reflections on learning. *Journal of the Scholarship of Teaching and Learning, 6,* 39-55.

Jessup, G. (1991). *Outcomes: NVQs and the emerging model of education and training.* London: Falmer Press.

Johnson, J.K. (1989). …Or none of the above. *The Science Teacher, 56(2),* 57-61.

Johnstone, A.H. & Ambusaidi, A. (2001). Fixed-response questions with a difference. *Chemistry Education: Research and Practice in Europe*, *2(3)*, 313-327.

Kehoe, J. (1995). Writing multiple choice tests items. *Practical Assessment, Research and Evaluation, 4(9).*
Retrieved on 5 December, 2005 from http://PAREonline.net/getvn.

Kenney, P.A. & Silver, E.A. (1993). An examination of relationships between 1990 NAEP mathematics items for grade 8 and selected themes from NCTM Standards. *Journal for Research in Mathematics Education, 24(2)*, 159-167.

Kerr, S.T. (1991). Lever and fulcrum: educational technology in teachers' thought and practice. *Teachers College Record, 93(1)*, 114-136.

Kilpatrick, J. (1993). The chain and the arrow: From the history of mathematics assessment. In M. Niss (Ed.), *Investigations into assessment in mathematics education: An ICMI study* (pp. 31-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Knight, P. (1995). *Assessment for learning in higher education*. Published in association with the Staff and Educational Development Association. London: Kogan Page.

Krutetskii, V.A. (1976). *The psychology of mathematical abilities in school children*. Chicago: University of Chicago Press.

Lajoie, S. (1991). A framework for authentic assessment in mathematics. *NCRMSE Research Review: The teaching and learning of Mathematics, 1(1),* 6-12.

Larisey, M.M. (1994). Student self assessment: a tool for learning. *Adult learning, 5(6),* 9-10.

Lawson, D. (1999). Formative assessment using computer-aided assessment. *Teaching Mathematics and its applications*, *18(4)*, 155-158.

Linacre, J.M. (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions*, *7(2)*, 328.
Retrieved on 13 February, 2006 from http://www.rasch.org/rmt/rmt74m.htm

Linacre, J.M. & Wright, B.D. (1999). *Winsteps Rasch model program*. Chicago: MESA Press.

Linacre, J.M. (2002). Optimizing rating scale effectiveness. *Journal of Outcome Measurement*, *3*, 85-106.

Linacre, J.M. (2005). *WINSTEPS Rasch measurement computer program*. Chicago: Winsteps.com.

Linacre, J.M. (2007). *Practical Rasch measurement, Lesson 2.*
Retrieved on 7 August, 2007 from www.statistics.com

Linn, R.L. (1989). *Educational measurement* (3rd ed.). New York: Macmillan.

Luckett, K. & Sutherland, L. (2000). Assessment practices that improve teaching and learning. In S. Makoni (Ed.), *Improving teaching and learning in higher education. A handbook for Southern Africa* (pp. 98-130). Johannesburg, South Africa: Wits University Press.

Makoni, S.(Ed.) (2000). *Improving teaching and learning in higher education. A handbook for Southern Africa* (pp. 98-130). Johannesburg, South Africa: Wits University Press.

Martinez, M. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28,* 131-145.

Marton, F. & Saljö, R. (1984). Approaches to learning. In F. Marton, D. Hounsell & N. Entwistle (Eds.), *The experience of learning* (pp. 36-55). Edinburgh: Scottish Academic Press.

Massachusetts Department of Education. (1987). *The 1987 Massachusetts Educational Assessment Program*. Quincy: Massachusetts Department of Education.

Mathematical Sciences Education Board (MSEB). (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.

Mathematical Sciences Education Board (MSEB). (1993). *Measuring what counts: A conceptual guide for mathematics assessment*. Washington, DC: National Academy Press.

McDonald, M. (2002). *Systematic assessment of learning outcomes: Developing multiple-choice exams.* Massachusetts, USA: Jones and Bartlett Publishers.

McFate, C. & Olmsted, J. (1999). Assessing student preparation through placement tests. *Journal of Chemical Education, 76(4)*, 562-565.

McIntosh, H. (Ed.) (1974). *Techniques and problems of assessment.* London: Edward Arnold.

McMillan, J.H. & Schumacher, S. (2001). Research in Education: *A conceptual introduction* (5th ed.). New York: Addison Wesley Longman, Inc.

Merriam, S.B. (1998). *Qualitative research and case study applications in education.* San Francisco: Jossey-Bass Publishers.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan Publishing Company.

Minick, N., Stone, C.A. & Forman, E.A. (1993). *Contexts for learning: Sociocultural dynamics in children's development.* New York: Oxford University Press.

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics (NCTM). (1995). *Assessment standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics.* Reston, VA: NCTM.
Retrieved on 7 September, 2006 from
http://standards.nctm.org/previous/currevstds/9-12sb.htm

Nightingale, P., Te Wiata, I., Toohey, S., Ryan, G., Hughes, C. & Magin, D. (1996). *Assessing learning in universities.* Sydney: University of New South Wales Press.

Niss, M. (1993). *Investigations into assessment in mathematics education. An ICMI Study.* Netherlands: Kluwer Academic Publishers.

Ochse, C. (2003). Are positive self-perceptions and expectancies really beneficial in an academic context? *South African Journal of Higher Education, 17(1),* 6-73.

Oosterhof, A. (1994). *Classroom applications of educational measurement.* Englewood Cliffs, NJ: Macmillan.

Ormell, C.P. (1974). Bloom's taxonomy and the objectives of education. *Educational Research, 17,* 3-18.

Osterlind, S.J. (1998). *Constructing test items: Multiple choice, constructed-response, performance and other formats* (2nd ed.). Boston: Kluwer Academic Publications.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Robertsw, R. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology*, *129*, 257-299.

Planinic, M., Boone, W.J., Krsnik, R. & 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., Rogan, J.M. & Howie, S. (2005). Chemical concepts inventory of Grade 12 learners and UP foundation year students. *African Journal of Research in SMT Education, 9(2)*, 121-134.

Pressley, M., Ghatala, E.S., Woloshyn, V. & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realise it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25(3),* 232-249.

Ramsden, P. (1984). The context of learning. In F. Marton, D. Hounsell & N. Entwistle (Eds.), *The experience of learning.* Edinburgh: Scottish Academic Press.

Ramsden, P. (1992). *Learning to teach in higher education.* London: Routledge.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institute.

Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements. In Blegvad, M. (Ed.), *The Danish Yearbook of Philosophy* (pp. 58-94). Copenhagen: The Danish Institute of Educational Research.

Rasch, G. (1980). Foreword and introduction. *Probabilistic models for some intelligence and attainment tests* (pp. 3-12, pp. ix-xix). Chicago: The University of Chicago Press.

Resnick, L.B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Resnick, L.R. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford and M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston and Dordrecht: Kluwer.

Robins, R.W. & Beer, J.S. (2001). Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology, 80,* 340-352.

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher, 94(1),* 31-37.

Romberg, T.A., Zarinnia, E.A. & Collis, K.F. (1990). A new world view of assessment in mathematics. In G.Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 21-38). Washington, DC: American Association for the advancement of Science.

Romberg, T.A. (1992). *Mathematics assessment and evaluation. Imperatives for mathematics educators.* Albany: State University of New York Press.

Rowntree, D. (1987). *Assessing students: How shall we know them?* (2nd ed.). London: Kogan Page.

Schoenfeld, A.H. (Ed.)(1987). *Cognitive science and mathematics education.* Hillsdale, N.J: Lawrence Erlbaum Associates.

Schoenfeld, A.H. (2002). Making mathematics work for all children: Issues of standards, testing and equity. *Educational Researcher, 31(1),* 13-25.

Schumacher, S. & McMillan, J.H. (1993). *Research in education: A conceptual introduction.* New York: Harper Collins.

Scouller, K. & Prosser, M. (1994). Students' experiences in studying for multiple-choice examinations. *Studies in Higher Education*, *19(3)*, 267-279.

Scriven, M. (1991). *Evaluation thesaurus,* 4th ed. London: Sage.

Senk, S.L., Beckmann, C.E. & Thompson, D.R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for Research in Mathematics Education, 28(2),* 187-215.

Sinkavich, F.J. (1995). Performance and metamemory: Do students know what they don't know? *Journal of Instructional Psychology, 22(1),* 77-87.

Sluijsmans, D., Moerkerke, G., van-Merrienboer, J. & Dochy, F. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation, 27,* 153-173.

Smith, G.H., Wood, L.N., Crawford, K., Coupland, M., Ball, G. & Stephenson, B. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *Int. J. Math. Educ. Sci. Technol.*, *27(1)*, 65-77.

Smith, G.H. & Wood, L.N. (2000). Assessment of learning in university mathematics. *Int. J. Math. Educ. Sci. Technol.*, *31(1)*, 125-132.

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

South African Qualifications Authority (SAQA). (2001). *Criteria and guidelines for the assessment of NQF registered unit standards and qualifications: Policy document.* Pretoria: SAQA.

Steen, L.A. (1999). Assessing assessment. In B. Gold (Ed.), *Assessment practices in undergraduate mathematics* (pp. 1-8). Washington, DC: Mathematical Association of America.

Stenmark, J.K. (1991) *Mathematics assessment: myths, models, good questions and practical suggestions.* Reston, VA: NCTM.

Stewart, J. (2000). *Calculus International Student Edition* (5th ed.). United States of America: Thomson Learning, Inc.

Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education*, *12(5)*, 563-573.

Tang, H. (1996). What is Rasch? *Rasch Measurement Transactions, 10(2)*, 507.

Thorndike, R.M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Tobias, S. & Everson, H. (2002). Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring. *College Board Report No. 2002-3.* New York: College Board.

Traub, R.E. & Fisher, C.W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1,* 355-369.

Traub, R.E. & Rowley, G.L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 19(1),* 37-45.

Treagust, D.F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in Science. *International Journal of Science Education*, *10,* 159-169.

Tyler, R.W. (1931). A generalized technique for constructing achievement tests. *Educational Research Bulletin, 8,* 199-208.

Wagner, E.P, Sasser, H. & DiBiase, W.J. (2002). Predicting students at risk in general chemistry using pre-semester assessments and demographic information. *Journal of Chemical Education, 79(6),* 749-755.

Webb, J.H. (1989). Multiple-choice questions in mathematics. *S.-Afr. Tydskr. Opvoedk., 9(1),* 216-218.

Webb, N. & Romberg, T.A. (1992) Implications of the NCTM standards for mathematics assessment. In T.A. Romberg (Ed.), *Mathematics Assessment and Evaluation: Imperatives for Mathematics Educators* (pp. 37-60). Albany: State University of New York Press.

Webb, J.M. (1994). The effects of feedback timing on learning facts: the role of response confidence. *Contemporary Educational Psychology*, *19*, 251-265.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.), *Educational measurement.* Washington DC: American Council of Education.

Wiggins, G. (1989). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan,* 703-713.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17,* 45-58.

Williams, J.B. (2006). Assertion – reason multiple-choice testing as a tool for deep learning: a qualitative analysis. *Assessment in Higher Education, 31(3),* 287-301.

Wood, L.N. & Smith, G.H. (1999). Flexible assessment. In W. Spunde, P. Cretchley, & R. Hubbard (Eds.), *The Challenge of Diversity* (pp. 229-233). Laguna Quays: University of Southern Queensland Press.

Wood, L.N. & Smith, G.H. (2001). Survey of the use of flexible assessment. *Quaestiones Mathematicae*, *Suppl. 1,* 73-82.

Wood, L.N. & Smith, G.H. (2002). Students' perceptions of difficulty in mathematical tasks. In M. Boezi (Ed.), *2nd International Conference on the Teaching of Mathematics,* Crete, Greece, July. New Jersey, USA: John Wiley & Sons.

Wood, L.N., Smith, G.H., Petocz, P., Reid, A. (2002). Correlations between students' performance in assessment and categories of a taxonomy. In M. Boezi (Ed.), *2ⁿᵈ International Conference on the Teaching of Mathematics*, Crete, Greece, July. New Jersey, USA: John Wiley & Sons.

World Book Dictionary (1990). Chicago, London, Sydney Toronto: World Book. Inc.

Wright, B.D.(1992) Point-biserials and item fits. *Rasch Measurement Transactions, 5(4),* 174.

Wright, B.D. & Linacre, J.M. (1989). *Observations are always ordinal: measurements, however, must be interval.* Chicago, IL: MESA Psychometric Laboratory.

Wright, B.D. & Stone, M.H. (1979). *The measurement model. Best Test Design.* Chicago: MESA Press.
Retrieved on 15 April, 2006 from http://www.rasch.org/books.htm

Yorke, M. (1988). The management of assessment in higher education. *Assessment and evaluation in higher education, 23,* 101-116.

Zohar, A. & Dori, Y.J. (2002). Higher order thinking skills and low achieving students: are they mutually exclusive? *The Journal of the Learning Sciences, 12(2),* 145-182.

Academic Information Systems Unit

Private Bag 3, WITS 2050 South Africa
Tel +27 11 717 1211/2/4 or 1061 Fax +27 11 717 1229

29 January 2007

I , Belinda Huntley, Staff Number 08901381, hereby declare that I will not use the information furnished to me by the University of the Witwatersrand in a manner that will bring the University in disrepute or in a way that it could be traced back to the University. I further agree that my research may be used by the University if it so desired. The Registrar has approved the use of this e-mail contact because of the importance the University attaches to the survey. Permission was granted on the understanding that you are not obliged to respond and that you may curtail your involvement at any time in the process

Signature*B.Huntley*.…………………………………

Date:2007/01/28…………………….

Table 1.2: Exit level outcomes (ELOs) of the undergraduate curriculum*

Exit Level Outcomes (ELOs)

The qualifying learner:

1.   generates, explores and considers options and makes decisions about ways of seeing systems and situations, and considers different ways of applying and integrating scientific knowledge to solve theoretical, applied or real life problems *specifically through research and the production of a research project*

2.   demonstrates an *advanced* understanding of key aspects of specified scientific systems and situations

3.   demonstrates an *advanced* understanding of specified bodies of content and their inter-connectedness in chosen disciplines

4.   demonstrates an *advanced* understanding of the boundaries, inter-connections, value and knowledge creation systems of chosen disciplines within the sciences

5.   reflects on possible implications for self and system of different ways of seeing and intervening in systems and situations

6.   demonstrates an ability to reflect with self and others, critical of own and other peoples' thoughts and actions, and capable of self-organisation and working in groups in the face of continual challenge from the environment

7.   demonstrates consciousness of, and engagement with own learning processes and the nature of knowledge, and how new knowledge can be acquired

8.   demonstrates an ability to conduct oneself as an independent learner and practitioner.

9.   *demonstrates an ability to reflect on the importance of scientific paradigms and methods in understanding scientific concepts and their changing nature*

(Source: Executive Information System, School of Mathematics, Academic Review 2000-2004, University of the Witwatersrand)
italicised text refers to the BScHons degree only; other text is common to the BSc and BScHons degrees

Table 1.3: Associated assessment criteria (AAC)*

   A. The learner should demonstrate an ability to consider a range of options and make decision about:

A.1   ways of seeing systems and situations, and to consider different ways of applying and integrating scientific knowledge to solve theoretical, applied or real life problems
A.2   methods for integrating information to solve complex problems
A.3   appropriate methods to carry out investigations to solve problems
A.4   appropriate use of quantitative techniques in the chosen discipline
A.5   selecting and appropriate method for communicating a set of data
A.6   the most appropriate personal learning strategies and organisation of work.
A.7   awareness of quality control, scientific standards and ethical norms as they pertain to the application of their chosen discipline in scientific investigations and the work place
A.8   awareness of the career path and professional responsibilities that accompany their chosen discipline.

   B. The learner should demonstrate an understanding of:

B.1   the use of critical thinking and logic in analysing situations
B.2   information storage and retrieval systems
B.3   basic computing skills; *effective communication and competent application of the relevant techniques including numerical and computer skills*
B.4   how to prepare a written scientific document; *how to design, execute and present scientific investigations such as through a small scale scientific report/research project*
B.5   modes of communicating, interpreting and translating data
B.6   relevant uses of quantitative methods to analyse and check for the plausibility of data
B.7   how to design and carry out scientific investigations
B.8   fundamental/*advances* techniques in the discipline

   C. The learner should demonstrate an ability to reflect on and critically evaluate:

C.1   the use of *advanced* investigative techniques and their strengths and weaknesses
C.2   the appropriateness of own interventions including strengths and weaknesses and possible future improvement of these
C.3   the relative merits of issues raised by science and technology and the relevance of science to everyday life and global issues
C.4   successes, strengths and weaknesses and possible improvement of personal learning strategies
C.5   own and other peoples' participation in a culturally and racially diverse learning situations and society.
C.6   scientific paradigms and methods in understanding scientific concepts and their changing nature
C.7   the practice and application of knowledge and understanding they have acquired of their chosen discipline in the workplace

---

(Source: Executive Information System, School of Mathematics, Academic Review 2000-2004, University of the Witwatersrand)
italicised text refers to the BScHons degree only; underlined text refers to the BSc degree only; other text is common to the BSc and BScHons degrees

Table 1.4: Critical cross-field outcomes (CCFOs)

CCFO (a)   identifying and solving problems in which responses display that responsible

decisions using critical and creative thinking have been made.

CCFO (b)   working with others as a member of a team, group, organisation, community.

CCFO (c)   organising and managing oneself and one's activities responsibly and effectively.

CCFO (d)   collecting, analysing, organising and critically evaluating information.

CCFO (e)   communicating effectively using visual, mathematical and/or language skills in the

modes of oral and/or written persuasion.

CCFO (f)   using science and technology effectively and critically, showing responsibility

towards the environment and health of others.

CCFO (g)   demonstrating an understanding of the world as a set of related systems by

recognising that problem-solving contexts do not exist in isolation.

CCFO (h)   contributing to the full personal development of each learner and the social and

economic development of society at large, by making it the underlying intention of

any programme of learning to make an individual aware of the importance of:

1. reflecting on and exploring a variety of strategies to learn more effectively;

2. participating as responsible citizens in the life of local, national and global communities;

3. being culturally and aesthetically sensitive across a range of social contexts;

4. exploring education and career opportunities;

5. developing entrepreneurial opportunities.

(Source: Executive Information System, School of Mathematics, Academic Review
2000-2004, University of the Witwatersrand)

Table 6.2: Misfitting and discarded test items

Item	Item difficulty	Model SE	INFIT		OUTFIT		PTMEA CORR
			MnSQ	ZSTD	MnSQ	ZSTD	
C45MB7	-3.94	0.47	0.83	-0.3	0.25	-1.5	0.26
C561B	-3.47	0.62	0.74	-0.4	0.29	-1.2	0.44
C46MA6	1.72	0.23	1.21	2.0	1.67	3.0	0.33
I036M04	-2.71	0.22	0.91	-0.6	0.45	-2.3	0.50
C361B	-3.31	0.36	0.86	-0.4	0.49	-1.4	0.32
C35M02	-3.61	0.47	1.11	0.4	1.61	1.1	0.08
C45MB6	-2.1	0.17	1.19	2.0	1.64	2.8	0.36

Test items Rasch statistics

ITEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MNSQ	ZSTD	MNSQ	ZSTD	
C35M01	216	295	-0.36	0.15	1.02	0.3	1.02	0.2	0.49
C35M02	174	179	-3.94	0.47	0.83	-0.3	0.25	-1.5	0.26
C35M03	242	297	-0.97	0.17	0.99	0	1.06	0.4	0.44
C35M04	276	298	-2.27	0.24	1.02	0.2	0.75	-0.7	0.33
C35M05	214	295	-0.32	0.15	1.19	2.5	1.25	2.1	0.41
A35M06	185	296	0.26	0.14	0.87	-2.2	0.82	-2.3	0.62
A35M07	238	297	-0.89	0.16	0.95	-0.5	0.95	-0.2	0.48
A35M08	73	278	2.25	0.15	1.03	0.5	1.02	0.2	0.68
A45MA146	253	418	0.2	0.11	1.01	0.2	0.98	-0.2	0.54
A45MA246	300	415	-0.5	0.12	0.95	-0.8	0.91	-0.8	0.53
A45MA346	323	417	-0.85	0.13	0.96	-0.5	0.87	-1	0.5
A45MA4	80	197	1.11	0.16	1.04	0.6	1.1	1	0.58
C45MA5	148	200	-0.7	0.18	1	0.1	1.03	0.3	0.48
C45MA6	189	200	-2.84	0.33	0.98	0	0.69	-0.6	0.3
C45MA7	119	199	0.13	0.16	0.93	-1	0.93	-0.8	0.58
C45MA8	118	127	-2.98	0.36	1.14	0.6	1.2	0.6	0.2
A45MB146	115	215	0.34	0.16	0.88	-1.9	0.8	-2.1	0.58
A45MB246	118	215	0.25	0.16	0.91	-1.5	0.83	-1.8	0.56
A45MB346	171	216	-1.18	0.19	1.05	0.5	0.88	-0.6	0.39
A45MB4	43	116	1.56	0.22	1.02	0.2	1.2	1.2	0.46
C45MB5	36	117	1.91	0.23	1.18	1.6	1.24	1.2	0.35
C45MB6	46	49	-3.47	0.62	0.74	-0.4	0.29	-1.2	0.44
C45MB7	37	108	1.72	0.23	1.21	2	1.67	3	0.33
C45MB8	88	100	-1.94	0.34	0.94	-0.2	0.67	-0.8	0.42
C55M01	257	327	-0.5	0.15	1.1	1.3	1.06	0.4	0.36
C55M02	240	328	-0.13	0.14	0.95	-0.7	1.06	0.5	0.46
C55M03	179	322	0.9	0.13	1.16	2.8	1.28	2.8	0.44
C55M04	145	328	1.5	0.13	1.02	0.3	1.03	0.4	0.55
C55M05	227	328	0.12	0.14	0.91	-1.5	0.85	-1.1	0.51
A55M06	21	251	4.56	0.24	0.91	-0.5	0.66	-1.1	0.73
A55M07	226	284	-0.76	0.16	1.05	0.6	1.13	0.9	0.33
A55M08	223	324	0.15	0.14	0.86	-2.2	0.74	-2.2	0.53
I65M0166	396	664	0.27	0.09	1.2	4.9	1.34	5.2	0.37
I65M0266	303	652	0.98	0.09	0.99	-0.1	0.98	-0.4	0.54
I65M0366	516	638	-1.1	0.11	0.95	-0.9	0.88	-1	0.41
I65M0466	416	669	0.14	0.09	1.04	1.1	1.04	0.7	0.46
I65M0566	342	662	0.7	0.09	1.03	0.9	1.01	0.3	0.5
I65M06	279	324	-1.36	0.17	0.99	-0.1	1.1	0.6	0.32
I65M0766	546	675	-1.04	0.11	0.93	-1.1	1.01	0.1	0.41
I65M08	271	328	-1.04	0.16	0.98	-0.2	0.95	-0.3	0.35
I65M09	127	349	1.72	0.12	0.81	-3.7	0.77	-2.9	0.66
I65M10	125	343	1.73	0.13	0.91	-1.7	0.9	-1.2	0.61
I65M1166	395	644	0.18	0.09	0.99	-0.2	0.93	-1.1	0.5
I65M1266	218	631	1.62	0.09	1.13	2.9	1.23	3	0.49
A651A663	394	686	1.1	0.09	0.98	-0.6	0.87	-1.8	0.57
A651B	87	353	2.97	0.14	1.01	0.1	0.93	-0.5	0.61
A652A	283	369	-0.33	0.14	1	0	1.05	0.3	0.47
A652B561B	95	353	2.81	0.14	1.09	1.2	1.16	1.2	0.57
A653	274	369	-0.15	0.14	1.09	1.3	1.15	0.9	0.45
C651A662A	749	957	-0.9	0.09	0.87	-2.7	0.75	-2	0.54
C651B662B	512	652	-0.33	0.11	0.98	-0.3	1.06	0.5	0.45
C651C	250	369	0.27	0.13	0.99	-0.2	0.91	-0.7	0.53
C651D662E	506	686	0.1	0.1	1.01	0.2	0.97	-0.2	0.48
C651E662G	430	686	0.8	0.09	1	-0.1	1.03	0.3	0.53
C652A	273	335	-0.84	0.16	1.07	0.8	0.96	-0.2	0.41
C652B	254	369	0.2	0.13	0.99	-0.1	0.8	-1.5	0.53
C652C	260	369	0.1	0.13	1.01	0.2	0.83	-1.2	0.51
C652D	95	353	2.81	0.14	1.03	0.4	0.92	-0.6	0.6

ITEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.
C653A	229	256	-1.93	0.22	1.06	0.4	1.14	0.6	0.31
C653B	282	335	-1.07	0.16	1.02	0.3	1.15	0.8	0.39
C654	249	369	0.29	0.13	1.08	1.3	1.22	1.6	0.48
A85M0184	279	771	1.22	0.08	0.97	-0.8	0.92	-1.3	0.52
A85M0284	427	773	0.24	0.08	1.17	5	1.19	3.7	0.36
A85M0384	472	771	-0.08	0.08	0.91	-2.6	0.86	-2.6	0.52
A85M0484	400	772	0.41	0.08	0.92	-2.6	0.88	-2.6	0.53
A85M0584	572	640	-2.31	0.14	0.93	-0.7	0.73	-2	0.38
C85M0684	182	754	1.96	0.09	1.15	2.9	1.32	3.4	0.38
C85M0784	565	724	-1.17	0.1	1	0.1	1.03	0.3	0.38
C85M0884	301	775	1.08	0.08	0.93	-2.1	0.98	-0.3	0.53
C85M0984	472	770	-0.08	0.08	1.04	1.1	1.05	0.9	0.44
C85M1084	382	772	0.53	0.08	0.98	-0.7	0.98	-0.4	0.49
I95M01	225	352	-0.61	0.13	0.97	-0.5	0.89	-1	0.54
I95M02	197	220	-3.22	0.24	0.95	-0.2	0.75	-0.9	0.34
I95M03	133	350	0.84	0.13	0.99	-0.2	0.99	-0.1	0.54
I95M04	208	355	-0.3	0.13	1.1	1.7	1.27	2.7	0.46
I95M05	104	346	1.3	0.13	1	-0.1	1.09	0.7	0.52
I95M06	197	351	-0.16	0.13	1	0	1.08	0.9	0.52
I95M07	94	348	1.49	0.14	1.07	1	1.17	1.1	0.48
I95M08	92	346	1.52	0.14	0.86	-2.1	0.74	-1.7	0.6
A951	185	363	0.67	0.12	1.02	0.5	1.02	0.2	0.5
A952A	188	363	0.63	0.12	0.99	-0.2	0.92	-0.8	0.52
A952B	270	341	-1.15	0.15	1.23	2.6	1.22	1.3	0.3
A952C	189	363	0.61	0.12	0.96	-0.8	0.97	-0.2	0.53
A952D	112	355	1.8	0.13	1.07	1.2	1.08	0.7	0.46
A953A	265	341	-1.04	0.15	1.02	0.3	1.1	0.7	0.42
A953B	273	341	-1.22	0.15	0.86	-1.7	0.68	-2.2	0.53
A953C	101	355	2	0.13	0.89	-1.7	0.83	-1.4	0.57
C951	172	359	0.86	0.12	1	0	0.96	-0.4	0.51
C952	183	363	0.7	0.12	1.03	0.5	1.01	0.2	0.5
C953A	28	29	-5.56	1.03	0.94	0.2	0.41	-0.3	0.15
C953B	80	345	2.4	0.14	1.31	3.6	1.36	2.3	0.31
C953CI	273	318	-1.83	0.18	0.91	-0.8	0.84	-0.7	0.44
C953CII	224	363	0.08	0.13	0.93	-1.2	0.84	-1.5	0.54
C953D	221	363	0.13	0.12	0.92	-1.6	0.85	-1.5	0.55
C954	272	341	-1.2	0.15	0.93	-0.8	0.95	-0.3	0.46
C955	251	288	-2.09	0.19	1.06	0.5	0.94	-0.2	0.34
I115M01	162	359	0.67	0.12	0.96	-0.8	0.96	-0.6	0.48
I115M02	142	368	1	0.12	0.86	-3	0.83	-2.3	0.56
I115M03	140	360	0.98	0.12	1.01	0.1	1	0	0.46
I115M04	133	356	1.07	0.12	1.07	1.4	1.13	1.6	0.41
I115M05	205	361	0.03	0.12	1.03	0.6	1.05	0.8	0.39
I115M06	142	370	1.01	0.12	1.04	0.8	1.03	0.5	0.43
I115M07	270	350	-1.12	0.14	0.96	-0.5	0.93	-0.6	0.39
I115M08	220	359	-0.19	0.12	0.97	-0.6	0.96	-0.5	0.43
I115M09	168	367	0.63	0.12	0.95	-1.1	0.95	-0.8	0.49
I115M10	134	364	1.1	0.12	0.88	-2.4	0.84	-2.1	0.55
I115M11	263	346	-1.03	0.14	1.07	1	1.09	0.8	0.3
I115M12	87	356	1.85	0.14	0.99	-0.2	0.98	-0.2	0.47
I115M13	188	362	0.34	0.12	1.07	1.6	1.07	1	0.4
I115M14	178	364	0.5	0.12	0.97	-0.8	0.96	-0.6	0.47
I115M15	116	355	1.33	0.13	1.19	3.2	1.27	2.8	0.33
A1151I	182	205	-2.92	0.25	1.04	0.3	1.17	0.7	0.38
A1151II	222	265	-2.08	0.19	1.1	0.9	1.1	0.5	0.4
A1152A	233	339	-0.58	0.14	1.02	0.3	0.94	-0.5	0.5
A1152B	55	325	2.93	0.17	0.9	-1	0.78	-1.1	0.54
A1152C	29	289	3.83	0.21	1.06	0.4	1.09	0.4	0.43
A1153A	211	348	-0.03	0.13	1.16	2.7	1.38	3.1	0.42
A1153B	188	344	0.34	0.13	1.15	2.7	1.22	2.2	0.43
A1154A	235	317	-1.05	0.15	1.04	0.5	0.98	-0.1	0.47
A1154BI	225	339	-0.43	0.13	0.89	-1.8	0.73	-2.5	0.57
A1154BII	65	330	2.66	0.16	0.85	-1.6	0.66	-2.1	0.57

271

ITEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MNSQ	ZSTD	MNSQ	ZSTD	
A1154BIII	187	344	0.35	0.13	0.91	-1.7	0.85	-1.6	0.56
A1155AI	218	339	-0.3	0.13	0.93	-1.2	0.89	-1	0.54
A1155AII	199	348	0.16	0.13	0.92	-1.4	0.87	-1.2	0.55
A1155BI	215	339	-0.25	0.13	1.13	2.1	1.19	1.7	0.44
A1155BII	84	342	2.23	0.15	1.09	1.2	1.06	0.4	0.46
A1155BIII	179	349	0.56	0.13	1.2	3.6	1.27	2.4	0.41
A1156A	139	349	1.2	0.13	0.98	-0.4	0.93	-0.6	0.53
A1156B	188	349	0.42	0.13	1.09	1.6	1.07	0.7	0.47
C1151A	217	348	-0.14	0.13	0.92	-1.4	0.92	-0.7	0.55
C1151B	164	349	0.8	0.13	0.97	-0.6	0.98	-0.1	0.53
C1152A	238	306	-1.37	0.16	0.96	-0.4	0.95	-0.3	0.5
C1152B	66	330	2.64	0.16	0.92	-0.8	0.75	-1.4	0.54
C1153A	166	349	0.76	0.13	0.92	-1.5	0.82	-1.8	0.56
C1153B	107	347	1.78	0.14	1.01	0.1	0.91	-0.6	0.51
C1154A	185	344	0.39	0.13	0.94	-1.2	0.88	-1.2	0.54
C1154B	157	349	0.91	0.13	1.05	1	1.05	0.5	0.49
C1154CI	190	344	0.31	0.13	0.92	-1.5	0.82	-1.9	0.55
C1154CII	129	345	1.36	0.13	1.18	3	1.34	2.8	0.4
C1155	240	306	-1.42	0.16	1.16	1.6	1.36	2.1	0.38
C1156A	213	339	-0.22	0.13	0.88	-2	0.8	-1.9	0.57
C1156B	125	347	1.46	0.13	0.93	-1.1	0.83	-1.5	0.55
C1157A	241	306	-1.45	0.16	1	0	1.15	1	0.46
C1157B	192	348	0.28	0.13	0.89	-2.2	0.84	-1.6	0.57
I036M01	74	285	1.85	0.15	1.1	1.3	1.14	1.1	0.43
I036M02	73	77	-5.05	0.54	0.96	0	0.95	0.1	0.29
I036M03	196	316	-0.38	0.14	1.1	1.6	1.1	0.9	0.49
I036M04	246	277	-2.71	0.22	0.91	-0.6	0.45	-2.3	0.5
I036M05	196	321	-0.31	0.13	1	0.1	0.95	-0.4	0.54
I036M06	205	313	-0.57	0.14	0.92	-1.1	0.87	-1.1	0.58
I036M07	109	313	1.19	0.14	1.04	0.6	1.03	0.3	0.51
I036M08	121	313	0.98	0.13	0.95	-0.8	1.03	0.3	0.55
A36A	239	275	-1.7	0.2	1	0.1	0.95	-0.1	0.38
A36B	243	310	-0.79	0.16	0.98	-0.2	0.75	-1.2	0.48
A36C	207	310	0.02	0.14	0.8	-3.1	0.66	-2.8	0.62
A36D	153	323	1.27	0.14	1.06	0.9	1.1	0.9	0.53
A36E	100	316	2.28	0.14	0.95	-0.7	0.83	-1.2	0.59
C361A	239	276	-1.68	0.19	0.98	-0.1	1.15	0.6	0.37
C361B	138	147	-3.31	0.36	0.86	-0.4	0.49	-1.4	0.32
C361C	252	310	-1.02	0.17	0.89	-1.2	0.83	-0.7	0.49
C362A	168	323	0.99	0.13	1.07	1.2	1.23	1.9	0.51
C362B	210	237	-2.09	0.22	1.04	0.3	1.29	1.1	0.27
C363A	226	310	-0.39	0.15	1.05	0.7	0.98	-0.1	0.46
C363B	38	264	3.94	0.2	0.89	-0.9	0.64	-1.6	0.57
C364A	207	310	0.02	0.14	0.95	-0.7	0.96	-0.3	0.53
C364BI	32	263	4.19	0.21	1.05	0.4	0.92	-0.2	0.47
C364BII	196	323	0.48	0.14	1.32	4.6	1.32	2.2	0.39
A46MA4	89	217	1.41	0.16	1.1	1.5	1.23	1.7	0.47
C46MA5	50	193	2.47	0.18	1.05	0.6	1.03	0.3	0.48
C46MA6	94	99	-3.62	0.47	1.11	0.4	1.61	1.1	0.08
C46MA7	152	218	-0.23	0.17	0.97	-0.3	0.9	-0.7	0.53
C46MA8	150	158	-3.18	0.38	1	0.1	0.81	-0.2	0.23
A46MB4	43	98	0.45	0.23	0.99	-0.1	0.97	-0.2	0.48
C46MB5	60	97	-0.41	0.23	1.03	0.3	1.04	0.4	0.4
C46MB6	72	83	-2.24	0.34	1.01	0.1	1.09	0.4	0.23
C46MB7	37	96	0.73	0.23	1.09	0.9	1.05	0.4	0.42
C46MB8	77	83	-2.96	0.44	1.04	0.2	0.78	-0.3	0.2
I56M01	42	328	3.07	0.18	0.86	-1.2	0.65	-1.8	0.49
I56M02	163	336	0.77	0.12	1.03	0.7	1.07	0.9	0.44
I56M03	241	322	-0.71	0.14	1.08	1.1	1.09	0.8	0.36
I56M04	263	323	-1.2	0.16	1	0	1.05	0.4	0.39
I56M05	251	322	-0.94	0.15	0.99	-0.1	1.01	0.1	0.42
I56M06	158	327	0.79	0.12	0.96	-0.8	0.96	-0.5	0.49
I56M07	80	330	2.13	0.14	1.13	1.7	1.21	1.5	0.33

272

ITEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MNSQ	ZSTD	MNSQ	ZSTD	
I56M08	189	329	0.33	0.13	0.92	-1.6	0.89	-1.5	0.52
A561A	222	304	-1.51	0.15	0.84	-2.2	0.65	-2.7	0.59
A562A	227	305	-1.62	0.15	0.86	-1.9	0.76	-1.6	0.57
A562B	166	298	-0.41	0.14	0.91	-1.5	0.95	-0.5	0.6
A562C	183	304	-0.72	0.14	0.92	-1.3	0.85	-1.5	0.6
A562D	218	304	-1.42	0.15	1.19	2.5	1.44	2.8	0.41
C561AI	263	305	-2.63	0.19	0.96	-0.3	0.78	-0.8	0.45
C561AII	149	159	-4.51	0.36	0.9	-0.3	0.53	-1.1	0.32
C561AIII	116	295	0.5	0.14	1.14	2.2	1.21	2.1	0.52
C561B	246	305	-2.1	0.17	1.19	2	1.64	2.8	0.36
C562	161	298	-0.31	0.13	1.08	1.4	1.09	1.1	0.52
C563AI	120	128	-4.74	0.4	0.86	-0.4	0.59	-0.8	0.31
C563AII	169	298	-0.46	0.14	1.16	2.6	1.17	2	0.48
C563C	213	304	-1.31	0.15	0.97	-0.3	0.9	-0.7	0.54
I66M06	242	315	-1	0.15	1.03	0.4	1.04	0.3	0.38
I66M08	243	278	-2.02	0.19	0.95	-0.4	0.74	-1.2	0.34
I66M09	194	309	-0.14	0.13	0.84	-3.1	0.73	-3	0.58
I66M10	132	284	0.73	0.14	0.88	-2	0.86	-1.7	0.6
A6611	161	171	-2.35	0.33	0.93	-0.2	0.49	-1.5	0.24
A6612	249	317	0.02	0.16	1.06	0.8	1.19	1	0.39
A6613	182	317	1.36	0.13	1.07	1.1	1.02	0.2	0.51
A6614	175	317	1.49	0.13	1.08	1.4	1.04	0.5	0.51
A6621	243	317	0.16	0.15	0.8	-2.8	0.63	-2.3	0.56
A6622	173	317	1.52	0.13	0.72	-5.3	0.59	-4.9	0.69
C661A	205	317	0.94	0.14	0.87	-2.2	0.88	-1	0.58
C661B	246	317	0.09	0.15	1	0	1.07	0.4	0.44
C662C	234	283	-0.47	0.17	0.78	-2.4	0.57	-2.4	0.5
C662D	181	317	1.38	0.13	1.04	0.8	1.02	0.3	0.52
C662F	60	277	3.75	0.16	1.3	3.2	1.44	2.4	0.4
C663A	209	317	0.86	0.14	0.99	-0.1	0.97	-0.2	0.51
C663B	250	317	0	0.16	1.22	2.5	1.16	0.8	0.33
C663C	255	317	-0.13	0.16	1.02	0.2	0.86	-0.6	0.42
C663D	225	317	0.55	0.14	0.97	-0.4	0.89	-0.8	0.51
C664A	212	317	0.81	0.14	1.07	1.1	1	0	0.48
C664B	204	317	0.96	0.14	1	0	0.97	-0.2	0.52
C664C	201	221	-1.61	0.25	1.03	0.2	1.23	0.8	0.2
C665	227	283	-0.27	0.17	0.96	-0.4	1.07	0.5	0.41

Appendix A7

Confidence level items Rasch statistics

TEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MnSQ	ZSTD	MnSQ	ZSTD	
CC35M01	412	264	0.59	0.1	1.05	0.5	1.05	0.5	0.55
CC35M02	168	130	1.99	0.18	0.98	0	0.81	-1	0.53
CC35M03	301	221	1.33	0.12	1.08	0.7	0.89	-0.8	0.53
CC35M04	299	220	1.35	0.13	0.91	-0.7	0.81	-1.4	0.55
CC35M05	440	257	0.25	0.09	0.76	-2.8	0.76	-2.6	0.65
CA35M06	538	294	-0.13	0.08	0.79	-2.6	0.81	-2.3	0.68
CA35M07	431	259	0.34	0.09	0.87	-1.4	0.87	-1.3	0.62
CA35M08	748	288	-1.41	0.07	0.83	-2.4	0.82	-2.4	0.75
CA45MA146	829	392	-0.49	0.07	0.86	-2.1	0.91	-1.3	0.67
CA45MA246	748	387	-0.18	0.07	1.22	2.9	1.18	2.2	0.61
CA45MA346	556	357	0.73	0.08	0.84	-2	0.78	-2.4	0.6
CA45MA4	520	214	-0.93	0.09	0.82	-2.2	0.81	-2.1	0.7
CC45MA5	409	215	-0.04	0.09	0.93	-0.7	0.97	-0.3	0.61
CC45MA6	209	158	1.77	0.16	0.92	-0.4	0.86	-0.8	0.49
CC45MA7	357	212	0.38	0.1	0.85	-1.5	0.79	-1.8	0.61
CC45MA8	358	216	0.47	0.1	0.93	-0.6	0.93	-0.5	0.57
CA45MB146	327	154	-0.35	0.1	0.84	-1.5	0.9	-0.8	0.67
CA45MB246	321	155	-0.26	0.11	1.42	3.5	1.41	3.1	0.55
CA45MB346	250	153	0.6	0.12	0.74	-2.2	0.69	-2.2	0.66
CA45MB4	187	81	-0.73	0.14	0.68	-2.5	0.72	-2	0.72
CC45MB5	153	80	-0.06	0.15	0.7	-2.1	0.69	-1.9	0.69
CC45MB6	163	82	-0.2	0.15	0.94	-0.4	0.96	-0.2	0.64
CC45MB7	165	74	-0.67	0.15	1.18	1.2	1.12	0.8	0.64
CC45MB8	141	80	0.22	0.16	0.83	-1.1	0.83	-0.9	0.66
CC55M01	464	262	0.21	0.09	0.88	-1.4	0.96	-0.3	0.64
CC55M02	393	244	0.67	0.1	0.79	-2.2	0.82	-1.6	0.63
CC55M03	536	253	-0.43	0.08	1.25	2.8	1.21	2.2	0.65
CC55M04	445	259	0.32	0.09	0.8	-2.3	0.76	-2.4	0.68
CC55M05	386	237	0.62	0.1	0.95	-0.4	0.92	-0.6	0.62
CA55M06	571	254	-0.69	0.08	0.93	-0.9	0.94	-0.7	0.7
CA55M07	467	255	0.09	0.09	1.03	0.4	0.91	-0.8	0.67
CA55M08	524	251	-0.39	0.08	1.24	2.6	1.26	2.6	0.64
CI65M0166	768	338	-0.7	0.07	1.05	0.7	1.16	1.9	0.64
CI65M0266	773	334	-0.76	0.07	1.11	1.6	1.16	1.9	0.65
CI65M0366	502	320	0.76	0.09	1.54	5.2	1.45	3.8	0.51
CI65M0466	578	320	0.15	0.08	0.97	-0.3	1.07	0.8	0.59
CI65M0566	654	329	-0.2	0.07	1.06	0.8	1.04	0.5	0.63
CI65M06	280	187	1.03	0.12	0.95	-0.4	0.85	-1	0.59
CI65M0766	518	321	0.62	0.09	0.76	-3	0.76	-2.5	0.64
CI65M08	324	194	0.55	0.11	0.9	-0.9	0.9	-0.8	0.62
CI65M09	433	193	-0.6	0.09	1.08	0.9	1.07	0.7	0.64
CI65M10	396	192	-0.25	0.1	1.06	0.6	1.12	1.1	0.62
CI65M1166	649	312	-0.34	0.07	1.24	3	1.14	1.6	0.64
CI65M1266	746	302	-1.03	0.07	1.34	4.2	1.3	3.4	0.66
CA651A663	350	186	-0.05	0.1	1.09	0.9	1.1	0.9	0.59
CA651B	267	118	-0.64	0.12	1.34	2.6	1.28	2	0.59
CA652A	230	128	0.21	0.13	1.1	0.8	1.1	0.7	0.56
CA652B561B	465	224	-0.36	0.09	0.91	-1.1	0.85	-1.5	0.65
CA653	235	131	0.21	0.12	0.92	-0.6	1.01	0.1	0.57

274

TEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MnSQ	ZSTD	MnSQ	ZSTD	
CC651A662A	334	205	0.53	0.11	0.7	-3.1	0.76	-2.1	0.6
CC651B662B	331	189	0.26	0.11	0.7	-3.1	0.69	-2.8	0.61
CC651C	233	127	0.12	0.12	0.68	-2.8	0.65	-2.6	0.65
CC651D662E	337	181	0.02	0.1	0.81	-1.9	0.79	-1.9	0.62
CC651E662G	345	176	-0.18	0.1	0.68	-3.5	0.69	-2.9	0.67
CC652A	196	119	0.57	0.14	0.68	-2.5	0.65	-2.4	0.62
CC652B	216	122	0.31	0.13	0.75	-2	0.71	-2	0.63
CC652C	214	120	0.28	0.13	0.72	-2.3	0.7	-2.1	0.63
CC652D	249	107	-0.7	0.13	0.85	-1.2	0.86	-1	0.68
CC653A	175	115	0.94	0.15	0.87	-0.8	0.76	-1.4	0.57
CC653B	230	118	-0.04	0.13	1.02	0.2	1.11	0.8	0.59
CC654	208	107	-0.04	0.13	1.28	1.9	1.26	1.6	0.54
CA85M0184	1373	572	-0.71	0.05	1.09	1.7	1.2	3.2	0.62
CA85M0284	1344	570	-0.65	0.05	1.12	2.1	1.08	1.3	0.66
CA85M0384	1256	564	-0.43	0.05	1.2	3.5	1.14	2.2	0.66
CA85M0484	1119	568	0.01	0.06	1.11	1.9	1.07	1	0.63
CA85M0584	807	546	1.16	0.07	1.44	5.3	1.13	1.4	0.56
CC85M0684	1409	567	-0.83	0.05	1.22	3.9	1.32	4.9	0.58
CC85M0784	1043	567	0.28	0.06	1.01	0.2	0.97	-0.4	0.64
CC85M0884	1196	568	-0.22	0.06	1.06	1.1	1.07	1	0.64
CC85M0984	1037	562	0.25	0.06	1.08	1.2	1.03	0.4	0.63
CC85M1084	1355	562	-0.73	0.05	1.2	3.5	1.14	2.3	0.67
CI95M01	420	205	-0.11	0.09	1.6	5.5	1.55	4.6	0.54
CI95M02	353	206	0.54	0.1	1.19	1.8	1.08	0.7	0.58
CI95M03	469	206	-0.51	0.09	0.8	-2.4	0.86	-1.5	0.67
CI95M04	385	205	0.19	0.1	1.09	0.9	1.01	0.2	0.61
CI95M05	511	196	-1.02	0.09	1.34	3.4	1.36	3.3	0.6
CI95M06	469	203	-0.56	0.09	1.27	2.8	1.25	2.3	0.6
CI95M07	510	203	-0.87	0.09	1	0	1.02	0.3	0.64
CI95M08	489	199	-0.79	0.09	1.22	2.4	1.21	2.1	0.61
CA951	327	145	-0.52	0.11	1.06	0.6	1.13	1.1	0.64
CA952A	359	157	-0.6	0.1	0.8	-2.1	0.78	-2	0.67
CA952B	364	156	-0.65	0.1	0.86	-1.4	0.92	-0.7	0.65
CA952C	354	142	-0.87	0.11	0.92	-0.7	0.91	-0.7	0.65
CA952D	344	137	-0.9	0.11	1.05	0.5	1.05	0.5	0.64
CA953A	279	148	0.13	0.11	1.01	0.2	0.93	-0.5	0.64
CA953B	270	147	0.24	0.12	0.81	-1.7	0.74	-2.1	0.68
CA953C	307	138	-0.46	0.11	0.9	-0.9	0.86	-1.1	0.67
CC951	298	152	0.02	0.11	0.74	-2.5	0.89	-0.8	0.67
CC952	321	154	-0.21	0.11	0.68	-3.3	0.66	-3.1	0.7
CC953A	230	151	0.99	0.13	1.11	0.8	1.02	0.2	0.61
CC953B	270	146	0.26	0.12	1.01	0.2	0.92	-0.5	0.66
CC953CI	243	148	0.68	0.13	1.02	0.2	0.91	-0.5	0.64
CC953CII	268	134	-0.08	0.12	0.97	-0.2	0.92	-0.6	0.65
CC953D	267	139	0.09	0.12	0.98	-0.2	0.91	-0.6	0.66
CC954	278	152	0.24	0.11	0.85	-1.3	0.79	-1.6	0.67
CC955	204	134	0.97	0.14	1.16	1.1	0.94	-0.3	0.63
CI115M01	346	174	0.01	0.1	1.38	3.3	1.28	2.3	0.52
CI115M02	320	172	0.25	0.1	0.99	0	1.17	1.4	0.52
CI115M03	358	169	-0.21	0.1	1.3	2.7	1.28	2.4	0.51
CI115M04	431	163	-1.02	0.1	1.36	3.3	1.37	3.1	0.55
CI115M05	350	172	-0.09	0.1	1	0	0.96	-0.3	0.59
CI115M06	401	175	-0.52	0.09	1.05	0.6	1.17	1.6	0.52

275

TEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MnSQ	ZSTD	MnSQ	ZSTD	
CI115M07	335	175	0.11	0.1	1.02	0.2	1.02	0.2	0.56
CI115M08	345	172	-0.05	0.1	1.18	1.7	1.2	1.7	0.53
CI115M09	386	171	-0.47	0.1	1.14	1.4	1.08	0.8	0.57
CI115M10	352	166	-0.22	0.1	1.04	0.4	1.01	0.1	0.58
CI115M11	327	171	0.14	0.1	1.35	3	1.47	3.6	0.5
CI115M12	380	166	-0.51	0.1	1.3	2.9	1.24	2.2	0.53
CI115M13	308	163	0.19	0.11	1.26	2.2	1.15	1.2	0.55
CI115M14	342	162	-0.21	0.1	1.17	1.6	1.13	1.1	0.54
CI115M15	425	161	-1.04	0.1	1.22	2.1	1.24	2.2	0.54
CA1151I	231	131	0.38	0.12	1.15	1.1	1.14	1	0.55
CA1151II	248	131	0.12	0.12	0.76	-2.1	0.77	-1.8	0.63
CA1152A	241	122	-0.01	0.12	1.2	1.6	1.14	1	0.57
CA1152B	271	115	-0.68	0.12	1.11	1	1.13	1	0.59
CA1152C	277	114	-0.84	0.12	0.78	-2	0.79	-1.8	0.65
CA1153A	237	116	-0.16	0.12	0.91	-0.7	0.91	-0.7	0.6
CA1153B	245	112	-0.41	0.12	0.81	-1.6	0.82	-1.4	0.63
CA1154A	236	119	-0.05	0.12	0.89	-0.9	0.87	-0.9	0.61
CA1154BI	240	107	-0.5	0.12	0.73	-2.4	0.75	-2	0.66
CA1154BII	237	101	-0.65	0.12	1.01	0.1	1	0.1	0.62
CA1154BIII	242	100	-0.77	0.13	0.98	-0.1	0.92	-0.6	0.64
CA1155AI	227	111	-0.17	0.12	1.45	3.2	1.38	2.5	0.54
CA1155AII	188	98	0.07	0.14	1.25	1.7	1.24	1.5	0.59
CA1155BI	213	103	-0.21	0.13	0.87	-1	0.82	-1.3	0.64
CA1155BII	235	99	-0.72	0.13	0.79	-1.7	0.76	-1.9	0.68
CA1155BIII	208	97	-0.35	0.13	0.99	0	0.88	-0.8	0.64
CA1156A	245	103	-0.69	0.12	1.02	0.2	0.97	-0.1	0.63
CA1156B	210	100	-0.26	0.13	0.69	-2.6	0.66	-2.5	0.68
CC1151A	227	116	0.06	0.12	0.9	-0.8	0.96	-0.3	0.61
CC1151B	243	118	-0.14	0.12	0.87	-1	1.08	0.6	0.59
CC1152A	226	120	0.16	0.12	0.88	-0.9	0.86	-1	0.63
CC1152B	267	114	-0.62	0.12	0.99	0	0.97	-0.2	0.6
CC1153A	233	110	-0.21	0.12	0.91	-0.7	0.9	-0.7	0.62
CC1153B	255	102	-0.78	0.12	1.09	0.8	1.19	1.4	0.58
CC1154A	229	108	-0.26	0.12	0.97	-0.2	0.89	-0.7	0.62
CC1154B	230	109	-0.26	0.12	0.95	-0.3	0.93	-0.5	0.63
CC1154CI	263	113	-0.6	0.12	0.72	-2.5	0.75	-2.1	0.66
CC1154CII	244	105	-0.61	0.12	0.99	0	1.04	0.3	0.59
CC1155	228	113	-0.1	0.12	0.91	-0.7	1.06	0.5	0.6
CC1156A	227	108	-0.29	0.12	0.71	-2.5	0.76	-1.8	0.66
CC1156B	232	100	-0.61	0.13	1.12	1	1.09	0.7	0.59
CC1157A	181	104	0.39	0.14	1.06	0.5	0.92	-0.4	0.62
CC1157B	196	92	-0.31	0.13	0.93	-0.5	0.89	-0.7	0.64
CI036M01	382	220	0.26	0.1	1.03	0.3	1.14	1.2	0.51
CI036M02	165	130	2.07	0.18	1.06	0.4	0.98	0	0.33
CI036M03	373	218	0.31	0.1	0.85	-1.6	0.84	-1.4	0.58
CI036M04	240	180	1.57	0.14	0.9	-0.7	0.78	-1.4	0.47
CI036M05	363	221	0.46	0.1	0.71	-3.1	0.72	-2.6	0.61
CI036M06	461	228	-0.34	0.09	1.21	2.2	1.27	2.5	0.56
CI036M07	510	233	-0.65	0.08	0.92	-0.9	0.96	-0.4	0.66
CI036M08	393	224	0.2	0.1	1.03	0.3	0.95	-0.4	0.54
CA36A	192	128	0.89	0.14	1.28	1.8	1.08	0.5	0.41
CA36B	275	140	-0.27	0.11	0.89	-0.9	0.86	-1.1	0.61
CA36C	280	124	-0.84	0.12	0.67	-3.2	0.68	-2.8	0.73

276

TEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MnSQ	ZSTD	MnSQ	ZSTD	
CA36D	272	115	-1	0.12	0.87	-1.1	0.82	-1.4	0.72
CA36E	239	105	-0.87	0.13	0.64	-3.2	0.62	-3	0.75
CC361A	220	150	0.93	0.14	0.9	-0.7	0.9	-0.6	0.39
CC361B	97	79	2.28	0.25	0.98	0	0.78	-0.8	0.3
CC361C	227	144	0.66	0.13	1.11	0.8	1.07	0.5	0.46
CC362A	260	143	0.01	0.12	1.18	1.5	1.25	1.8	0.5
CC362B	260	150	0.22	0.12	0.89	-0.9	1	0.1	0.49
CC363A	226	142	0.56	0.13	1.02	0.2	1.05	0.4	0.41
CC363B	281	120	-1.04	0.12	0.93	-0.6	0.92	-0.6	0.67
CC364A	202	131	0.7	0.14	0.88	-0.8	0.85	-1	0.45
CC364BI	308	141	-0.65	0.11	0.8	-1.9	0.78	-1.9	0.67
CC364BII	252	124	-0.41	0.12	0.91	-0.7	0.94	-0.4	0.62
CA46MA4	402	171	-0.98	0.1	0.88	-1.2	0.91	-0.9	0.72
CC46MA5	299	170	0.05	0.11	0.77	-2.2	0.79	-1.7	0.66
CC46MA6	303	171	0.03	0.11	0.88	-1.1	0.88	-1	0.64
CC46MA7	275	173	0.43	0.12	0.85	-1.3	0.83	-1.2	0.62
CC46MA8	228	148	0.7	0.13	0.81	-1.5	0.8	-1.4	0.58
CA46MB4	182	73	-0.89	0.15	0.77	-1.6	0.72	-1.9	0.77
CC46MB5	152	71	-0.31	0.16	0.98	-0.1	1.2	1.1	0.64
CC46MB6	87	65	1.69	0.24	1.16	0.7	0.77	-0.8	0.46
CC46MB7	146	73	-0.05	0.16	0.87	-0.8	0.78	-1.3	0.68
CC46MB8	121	72	0.6	0.18	0.9	-0.5	0.81	-0.8	0.61
CI56M01	340	171	-0.16	0.1	0.99	0	1.15	1.2	0.67
CI56M02	290	168	0.39	0.12	0.97	-0.2	0.91	-0.6	0.65
CI56M03	288	165	0.33	0.11	1.19	1.6	1.04	0.4	0.63
CI56M04	296	167	0.27	0.11	0.95	-0.4	0.99	0	0.65
CI56M05	261	163	0.71	0.12	1	0.1	0.92	-0.5	0.64
CI56M06	357	163	-0.54	0.1	1.25	2.2	1.38	3	0.65
CI56M07	309	166	0.07	0.11	0.85	-1.4	0.83	-1.3	0.7
CI56M08	279	168	0.55	0.12	0.89	-0.9	0.87	-0.9	0.66
CA561A	198	98	-0.27	0.13	0.93	-0.5	0.88	-0.7	0.66
CA562A	209	106	-0.15	0.13	0.88	-0.9	0.94	-0.4	0.63
CA562B	192	96	-0.25	0.14	0.74	-2.1	0.71	-2	0.67
CA562C	202	94	-0.47	0.13	0.87	-1	0.84	-1.1	0.67
CA562D	181	89	-0.37	0.14	1.35	2.3	1.28	1.7	0.59
CC561AI	187	107	0.32	0.14	0.71	-2.2	0.72	-1.9	0.61
CC561AII	164	103	0.66	0.15	1.03	0.3	0.98	0	0.52
CC561AIII	190	93	-0.28	0.14	1.1	0.8	1.04	0.3	0.59
CC561B	172	102	0.43	0.15	0.83	-1.1	0.75	-1.5	0.6
CC562	203	93	-0.53	0.13	0.92	-0.6	0.93	-0.4	0.67
CC563AI	120	89	1.61	0.21	1.22	1.1	1.22	1	0.46
CC563AII	195	91	-0.53	0.14	0.94	-0.4	1.14	0.9	0.61
CC563C	173	86	-0.33	0.14	0.92	-0.5	1.15	0.9	0.59
CI66M06	234	125	-0.07	0.12	0.87	-1	1.33	2.1	0.59
CI66M08	215	121	0.16	0.13	1.15	1.1	0.97	-0.1	0.59
CI66M09	256	129	-0.36	0.12	0.79	-1.8	0.76	-1.8	0.69
CI66M10	284	116	-1.15	0.12	1.4	3	1.39	2.6	0.67
CA6611	114	69	0.44	0.18	1.15	0.8	0.98	0	0.58
CA6612	117	61	-0.22	0.18	1.04	0.3	1.09	0.5	0.55
CA6613	124	61	-0.52	0.17	1.09	0.6	1.01	0.1	0.62
CA6614	97	56	0.13	0.2	0.89	-0.5	0.77	-1	0.64
CA6621	97	60	0.52	0.2	0.83	-0.8	0.87	-0.5	0.67
CA6622	89	51	0	0.21	0.92	-0.3	0.96	-0.1	0.59

TEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.
					MnSQ	ZSTD	MnSQ	ZSTD	
CC661A	101	65	0.62	0.2	0.77	-1.2	0.79	-0.9	0.61
CC661B	95	62	0.75	0.21	1	0.1	0.87	-0.4	0.59
CC662C	114	59	-0.2	0.18	0.69	-1.8	0.64	-2	0.67
CC662D	110	57	-0.2	0.18	0.59	-2.6	0.59	-2.3	0.68
CC662F	105	56	-0.15	0.19	0.77	-1.2	0.7	-1.5	0.63
CC663A	85	51	0.51	0.21	1.11	0.6	0.9	-0.3	0.59
CC663B	80	51	0.71	0.23	0.98	0	0.8	-0.7	0.61
CC663C	83	50	0.29	0.22	0.8	-0.9	0.84	-0.6	0.63
CC663D	94	53	0.08	0.2	0.57	-2.4	0.53	-2.3	0.66
CC664A	103	58	0.24	0.19	0.88	-0.6	0.87	-0.5	0.62
CC664B	73	53	1.39	0.26	0.9	-0.3	0.9	-0.2	0.54
CC664C	79	55	1.16	0.24	1.1	0.5	1.19	0.7	0.51
CC665	61	47	1.79	0.3	1.24	0.9	1.07	0.3	0.51

Item analysis data

Item	Diff	Adapted discrimi-nation	Adapted confidence deviation	Adapted expert opinion deviation		QI_3	Com-ponent	Good/poor)
A6622	1.52	0.048	0.495	0.251		0.069	4	1
A35M06	0.26	0.192	0.271	0.267		0.076	1	1
A651B	2.97	0.213	0.291	0.240		0.079	1	1
C1151A	-0.14	0.336	0.244	0.285		0.107	3	1
A55M06	4.56	-0.035	0.537	0.550		0.112	1	1
A651A	1.1	0.295	0.385	0.236		0.119	1	1
C1157B	0.28	0.295	0.398	0.239		0.123	3	1
C85M0884	1.08	0.378	0.258	0.299		0.125	3	1
C1152B	2.64	0.357	0.247	0.342		0.128	2	1
C1151B	0.8	0.378	0.266	0.329		0.135	2	1
I65M09	1.72	0.110	0.351	0.608		0.138	3	1
A1152B	2.93	0.357	0.255	0.373		0.138	3	1
A45MB146	0.34	0.275	0.416	0.301		0.140	2	1
A36E	2.28	0.254	0.447	0.303		0.141	2	1
C651C	0.27	0.378	0.360	0.268		0.144	1	1
A953C	2	0.295	0.249	0.492		0.148	2	1
C1152A	-1.37	0.439	0.352	0.272		0.160	6	1
A95M01	-0.61	0.357	0.412	0.303		0.164	3	1
A35M08	2.25	0.069	0.842	0.355		0.165	2	1
C662D	1.38	0.398	0.326	0.351		0.166	3	1
C363B	3.94	0.295	0.274	0.574		0.177	2	1
A652B	2.81	0.295	0.465	0.360		0.178	5	1
A36M06	-0.57	0.275	0.570	0.307		0.180	2	1
I65M10	1.73	0.213	0.352	0.609		0.181	6	1
C95M08	1.52	0.233	0.524	0.398		0.183	3	1
C951	0.86	0.419	0.392	0.323		0.185	7	1
I65M0466	0.14	0.522	0.358	0.280		0.188	6	1
C36M03	-0.38	0.460	0.381	0.311		0.189	3	1
A562B	-0.41	0.233	0.477	0.461		0.190	3	1
A1154BII	2.66	0.295	0.229	0.713		0.191	1	1
C115M02	1	0.316	0.583	0.286		0.191	3	1
C652D	2.81	0.233	0.230	0.843		0.192	2	1
C36M05	-0.31	0.357	0.502	0.314		0.195	7	1
A6613	1.36	0.419	0.357	0.390		0.196	1	1
A45MA146	0.2	0.357	0.542	0.290		0.197	2	1
C115M01	0.67	0.481	0.352	0.343		0.197	1	1
C66M09	-0.14	0.275	0.508	0.406		0.198	7	1
A45MB246	0.25	0.316	0.367	0.501		0.198	3	1
A36M07	1.19	0.419	0.481	0.289		0.200	2	1
C45MA7	0.13	0.275	0.523	0.402		0.201	3	1
C953D	0.13	0.336	0.313	0.557		0.202	7	1
A953A	-1.04	0.604	0.315	0.308		0.205	2	1
C45MA5	-0.7	0.481	0.377	0.346		0.207	2	1

A45MA4	1.11	0.275	0.698	0.296		0.207	7	1
C651D662E	0.1	0.481	0.257	0.487		0.209	2	1
C561AIII	0.5	0.398	0.337	0.476		0.210	2	1
C954	-1.2	0.522	0.264	0.449		0.212	3	1
A45MB4	1.56	0.522	0.473	0.247		0.213	1	1
C1154A	0.39	0.357	0.342	0.537		0.215	3	1
C1157A	-1.45	0.522	0.249	0.483		0.218	3	1
C55M03	0.9	0.563	0.374	0.318		0.221	2	1
A36M08	0.98	0.336	0.544	0.371		0.221	1	1
C561AI	-2.63	0.543	0.460	0.262		0.222	2	1
C35M05	-0.32	0.625	0.349	0.304		0.223	2	1
C56M06	0.79	0.460	0.473	0.324		0.225	7	1
A953B	-1.22	0.378	0.267	0.655		0.227	2	1
C651B662B	-0.33	0.543	0.354	0.371		0.227	3	1
C1153B	1.78	0.419	0.470	0.369		0.228	2	1
A95M03	0.84	0.357	0.443	0.460		0.228	5	1
A95M04	-0.3	0.522	0.309	0.449		0.231	2	1
C45MB8	-1.94	0.604	0.410	0.284		0.232	3	1
C1154B	0.91	0.460	0.250	0.593		0.232	3	1
A85M0484	0.41	0.378	0.305	0.623		0.234	6	1
C651E662G	0.8	0.378	0.238	0.736		0.235	3	1
A55M07	-0.76	0.790	0.294	0.290		0.236	2	1
C362A	0.99	0.419	0.408	0.455		0.237	4	1
A45MA346	-0.85	0.439	0.601	0.277		0.239	3	1
A35M07	-0.89	0.481	0.312	0.508		0.239	1	1
A951	0.67	0.439	0.480	0.372		0.239	6	1
C664A	0.81	0.481	0.542	0.290		0.241	5	1
A952D	1.8	0.522	0.553	0.251		0.242	1	1
C652C	0.1	0.419	0.445	0.432		0.242	3	1
C1154CI	0.31	0.336	0.602	0.382		0.243	3	1
C95M06	-0.16	0.398	0.656	0.287		0.244	7	1
A6612	0.02	0.666	0.379	0.301		0.246	7	1
A85M0184	1.22	0.398	0.519	0.397		0.247	1	1
C46MA7	-0.23	0.378	0.495	0.441		0.247	3	1
I65M0566	0.7	0.439	0.244	0.680		0.248	6	1
A1156A	1.2	0.378	0.509	0.430		0.248	4	1
A653	-0.15	0.543	0.350	0.432		0.249	2	1
C661A	0.94	0.275	0.840	0.314		0.251	5	1
A952C	0.61	0.378	0.743	0.268		0.252	2	1
C1153A	0.76	0.316	0.240	0.919		0.254	3	1
C115M05	0.03	0.666	0.283	0.424		0.256	2	1
C953CII	0.08	0.357	0.267	0.796		0.256	7	1
C35M01	-0.36	0.460	0.587	0.309		0.257	5	1
A45MA246	-0.5	0.378	0.443	0.524		0.258	2	1
C651A662A	-0.9	0.357	0.448	0.543		0.259	5	1
C663D	0.55	0.419	0.381	0.554		0.261	7	1
C115M08	-0.19	0.584	0.294	0.492		0.261	3	1
A1153A	-0.03	0.604	0.345	0.418		0.262	1	1
C115M03	0.98	0.522	0.248	0.623		0.264	3	1
A1152A	-0.58	0.439	0.334	0.601		0.265	2	1
A55M08	0.15	0.378	0.479	0.504		0.265	4	1

C1156A	-0.22	0.295	0.472	0.617		0.265	5	1
A36B	-0.79	0.481	0.559	0.336		0.267	2	1
A1155AII	0.16	0.336	0.304	0.804		0.267	1	1
C85M0784	-1.17	0.687	0.230	0.514		0.272	7	1
A562A	-1.62	0.295	0.620	0.487		0.272	4	1
A652A	-0.33	0.501	0.318	0.574		0.273	1	1
I65M0766	-1.04	0.625	0.488	0.308		0.281	7	1
C952	0.7	0.439	0.251	0.779		0.281	5	1
C115M07	-1.12	0.666	0.343	0.416		0.281	1	1
A46MA4	1.41	0.501	0.680	0.263	Median QI	0.282	3	0
C115M06	1.01	0.584	0.420	0.409		0.284	7	0
C663A	0.86	0.419	0.746	0.295		0.284	3	0
A561A	-1.51	0.254	0.687	0.519		0.287	5	0
A1153B	0.34	0.584	0.459	0.379		0.287	1	0
I65M0266	0.98	0.357	0.598	0.475		0.289	6	0
A952A	0.63	0.398	0.545	0.490		0.294	3	0
C652B	0.2	0.378	0.484	0.577		0.295	2	0
C653B	-1.07	0.666	0.443	0.349		0.295	3	0
C46MA8	-3.18	0.996	0.284	0.322		0.301	2	0
C46MB5	-0.41	0.646	0.520	0.314		0.304	3	0
A95M02	-3.22	0.769	0.406	0.333		0.305	3	0
A36C	0.02	0.192	0.826	0.536		0.305	2	0
C652A	-0.84	0.625	0.487	0.361		0.306	2	0
A1155AI	-0.3	0.357	0.400	0.755		0.309	1	0
C654	0.29	0.481	0.248	0.819		0.310	7	0
A1156B	0.42	0.501	0.337	0.663		0.314	2	0
A6621	0.16	0.316	0.629	0.561		0.315	1	0
C1156B	1.46	0.336	0.405	0.799		0.315	2	0
A56M01	3.07	0.460	0.655	0.389		0.318	7	0
C56M05	-0.94	0.604	0.571	0.335		0.320	3	0
C46MB8	-2.96	1.058	0.317	0.298		0.323	2	0
C662C	-0.47	0.439	0.452	0.613		0.323	1	0
A36A	-1.7	0.687	0.565	0.287		0.324	2	0
A85M0384	-0.08	0.398	0.548	0.569		0.328	3	0
C56M04	-1.2	0.666	0.242	0.657		0.328	2	0
C85M0984	-0.08	0.563	0.391	0.571		0.332	7	0
C36M01	1.85	0.584	0.742	0.256		0.334	2	0
C46MB7	0.73	0.604	0.319	0.630		0.335	6	0
A56M03	-0.71	0.728	0.337	0.501		0.337	4	0
A85M05	-2.31	0.687	0.652	0.249		0.338	4	0
C953CI	-1.83	0.563	0.391	0.589		0.338	3	0
A1152C	3.83	0.584	0.300	0.691		0.340	2	0
C66M10	0.73	0.233	0.924	0.500		0.344	3	0
C364BI	4.19	0.501	0.501	0.547		0.346	2	0
A45MB346	-1.18	0.666	0.449	0.450		0.347	2	0
C55M01	-0.5	0.728	0.288	0.587		0.349	6	0
A562C	-0.72	0.233	0.691	0.703		0.351	3	0
A1155BII	2.23	0.522	0.347	0.736		0.356	1	0
C55M04	1.5	0.336	0.723	0.546		0.356	3	0
C95M07	1.49	0.481	0.587	0.510		0.358	4	0
C563AI	-4.74	0.831	0.545	0.273		0.359	2	0

C663C	-0.13	0.604	0.411	0.577		0.361	6	0
A46MB4	0.45	0.481	0.786	0.367		0.365	3	0
A36D	1.27	0.378	0.720	0.522		0.366	2	0
A6611	-2.35	0.975	0.324	0.410		0.367	1	0
A1151I	-2.92	0.687	0.468	0.470		0.374	2	0
C1154CII	1.36	0.646	0.422	0.561		0.378	1	0
C66M06	-1	0.687	0.452	0.496		0.379	5	0
C563AII	-0.46	0.481	0.688	0.466		0.379	4	0
C55M02	-0.13	0.522	0.686	0.430		0.380	2	0
C45MA8	-2.98	1.058	0.414	0.300		0.381	2	0
C46MA5	2.47	0.481	0.700	0.470		0.386	4	0
C361C	-1.02	0.460	0.520	0.673		0.389	2	0
C1155	-1.42	0.687	0.548	0.424		0.390	3	0
A56M02	0.77	0.563	0.643	0.453		0.393	1	0
C45MB5	1.91	0.749	0.521	0.409		0.394	2	0
I65M0366	-1.1	0.625	0.578	0.459		0.395	3	0
A1151II	-2.08	0.646	0.507	0.561		0.422	1	0
C364BII	0.48	0.666	0.434	0.657		0.438	1	0
A85M0284	0.24	0.728	0.650	0.395		0.441	1	0
I65M1166	0.18	0.439	0.437	0.945		0.442	7	0
A562D	-1.42	0.625	0.743	0.424		0.452	4	0
C562	-0.31	0.398	0.661	0.742		0.455	2	0
C115M04	1.07	0.625	0.770	0.415		0.459	2	0
C66M08	-2.02	0.769	0.467	0.568		0.460	2	0
C55M05	0.12	0.419	0.694	0.696		0.461	7	0
C653A	-1.93	0.831	0.561	0.431		0.462	1	0
A1154A	-1.05	0.501	0.446	0.896		0.465	1	0
A6614	1.49	0.419	0.584	0.827		0.465	3	0
A1155BIII	0.56	0.625	0.377	0.848		0.470	3	0
C363A	-0.39	0.522	0.560	0.745		0.475	2	0
A1155BI	-0.25	0.563	0.420	0.882		0.478	3	0
C663B	0	0.790	0.738	0.348		0.482	3	0
C36M02	-5.05	0.872	0.822	0.239		0.486	2	0
C65M08	-1.04	0.749	0.437	0.674		0.488	1	0
C364A	0.02	0.378	0.734	0.789		0.500	5	0
A1154BI	-0.43	0.295	0.661	1.048		0.518	2	0
C361A	-1.68	0.707	0.598	0.605		0.525	1	0
A1154BIII	0.35	0.316	0.717	0.964		0.529	6	0
C35M04	-2.27	0.790	0.796	0.394		0.543	2	0
C362B	-2.09	0.913	0.436	0.643		0.548	1	0
C955	-2.09	0.769	0.554	0.643		0.553	3	0
C561AII	-4.51	0.810	0.549	0.613		0.553	5	0
I65M06	-1.36	0.810	0.727	0.457		0.559	7	0
C953A	-5.56	1.161	0.497	0.434		0.562	3	0
I65M0166	0.27	0.707	0.681	0.596		0.567	7	0
C56M07	2.13	0.790	0.654	0.551		0.568	4	0
C662F	3.75	0.646	0.783	0.609		0.595	7	0
C35M03	-0.97	0.563	1.013	0.521		0.603	3	0
A952B	-1.15	0.852	0.897	0.370		0.611	3	0
C85M0684	1.96	0.687	0.475	0.935		0.611	4	0
I65M1266	1.62	0.460	0.679	0.972		0.615	2	0

282

C95M05	1.3	0.398	0.729	1.007		0.617	3	0
C563C	-1.31	0.357	0.695	1.144		0.628	2	0
C45MA6	-2.84	0.852	0.998	0.333		0.634	3	0
C56M08	0.33	0.398	0.681	1.112		0.637	3	0
C661B	0.09	0.563	0.782	0.797		0.655	3	0
C85M1084	0.53	0.460	0.656	1.090		0.658	7	0
C664C	-1.61	1.058	0.776	0.612		0.842	2	0
C953B	2.4	0.831	0.839	0.865		0.927	3	0
C46MB6	-2.24	0.996	1.047	0.544		0.933	7	0
C664B	0.96	0.398	1.399	0.891		0.935	5	0
C665	-0.27	0.625	1.469	0.758		1.085	3	0
Average diff	0.0617				Median QI	0.282		
Median diff	0.13							