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).

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 naive 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.

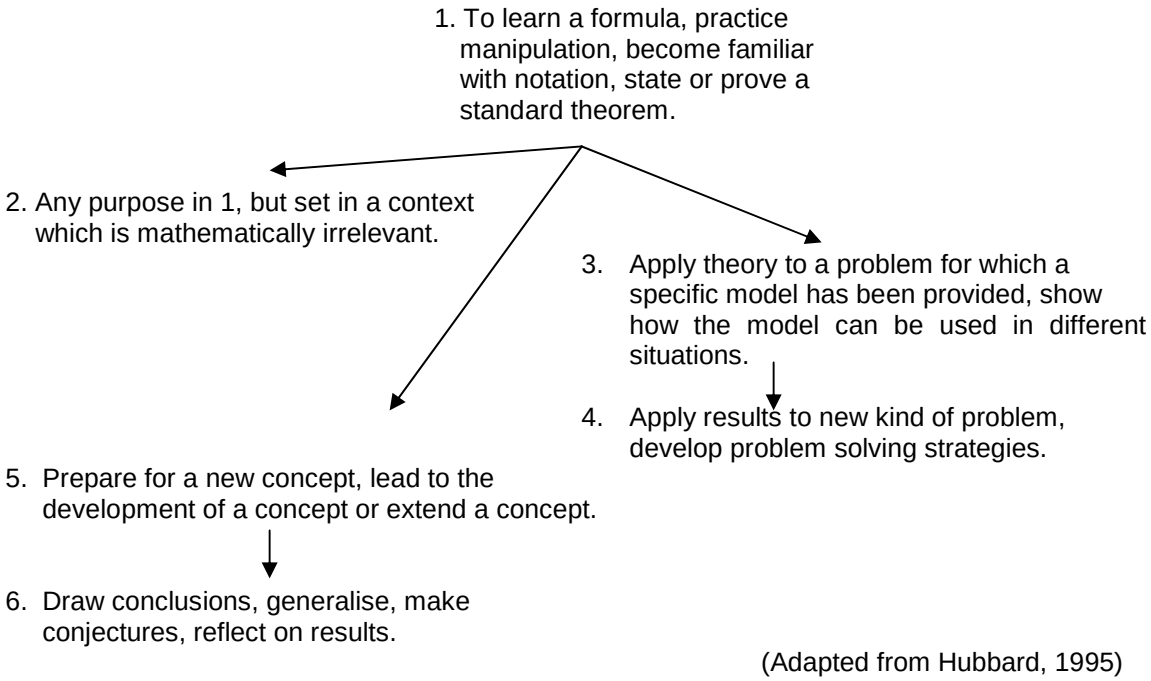
(Adapted from Biggs, 1991)

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.



(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.

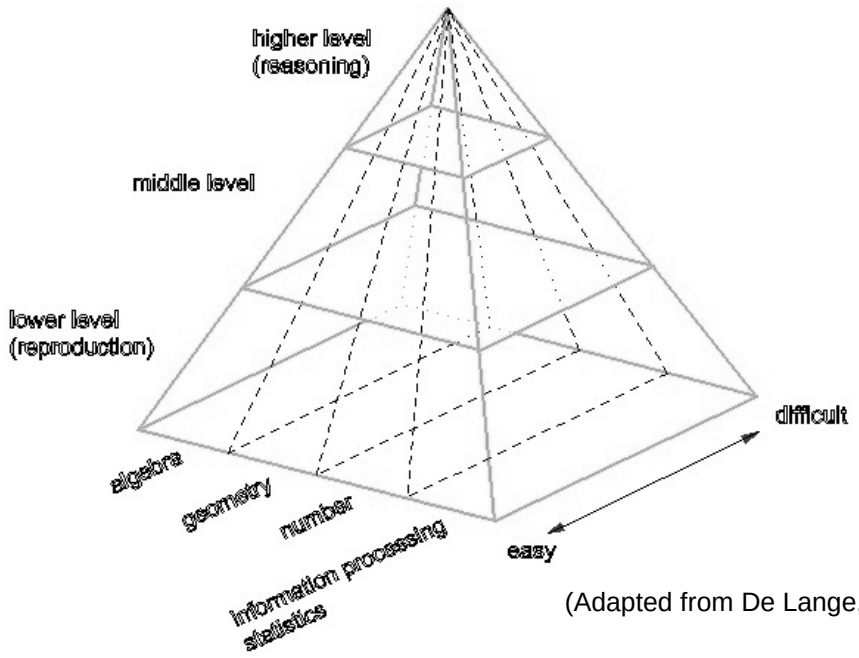
1. Recognition of key words and symbols which trigger memorised, standard procedures.
↓
2. Some understanding of standard procedures so that they can be modified slightly for new situations.
↓
3. Ability to explain and justify procedures and to form them into a coherent system.
↓
4. Ability to synthesise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

> 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

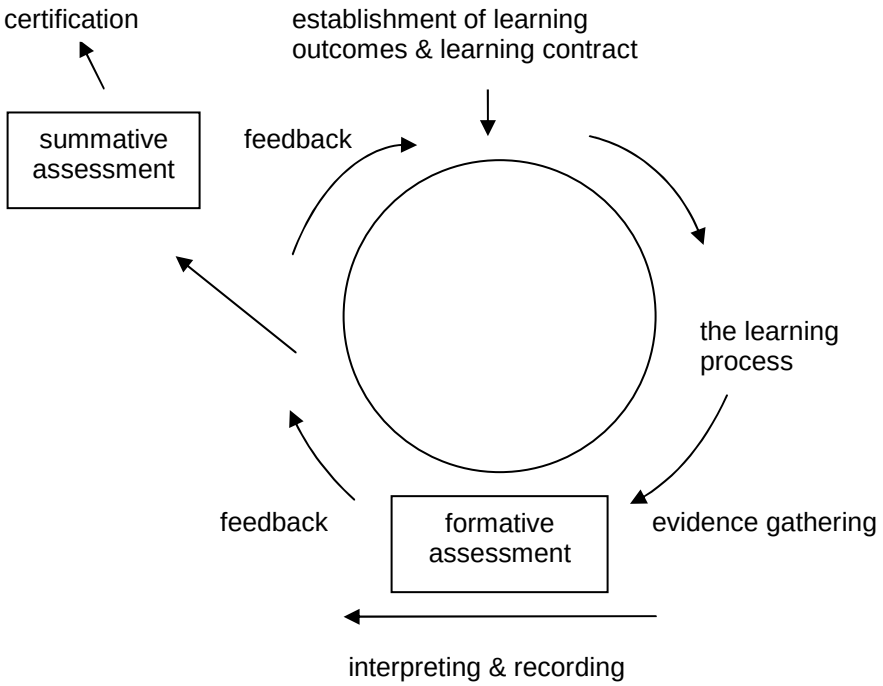
The term *'summative'* implies an overview of previous learning. Summative assessment is used to grade students at the end of a unit, or to accredit at the end of a programme (Biggs, 2000). Summative assessment is used to provide judgement on students' achievements in order to:

- 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



(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 10^{th} 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.

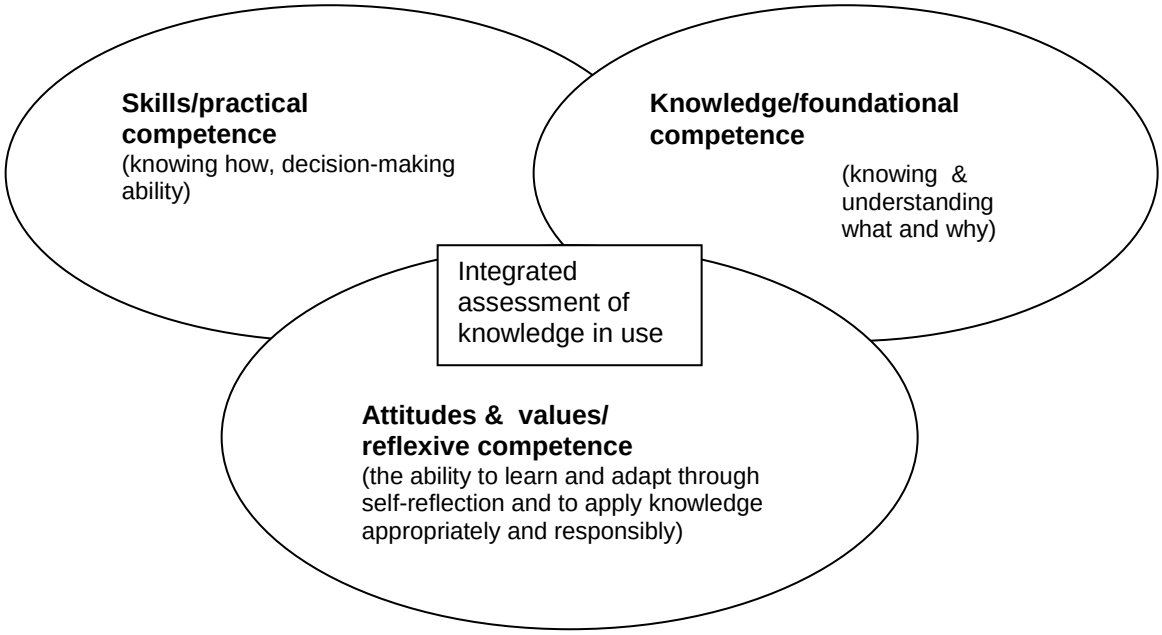
Learning is multidimensional and integrated. Integrated assessment is needed to ensure that students can bring together and integrate all the knowledge, skills and attitudes they have gleaned from a programme as a whole. Outcomes-based education requires integrated assessment of competence, which is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



(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

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

A. $(\| \underline{u} \| + \| \underline{v} \|)^2$
B. $(\| \underline{u} \| - \| \underline{v} \|)^2$
C. $\| \underline{u} \|^2 - \| \underline{v} \|^2$ — Distracters
D. $\| \underline{u} \|^2 + \| \underline{v} \|^2$ — Key

Options

Item

(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.