CHAPTER 7: DISCUSSION AND CONCLUSIONS

In Chapter 7, I set about discussing my research results. The discussion in this chapter will include the interpretation of the results and the implications for future research. I intend to discuss how the research results could have implications for assessment practices in undergraduate mathematics.

Using the Quality Index model, as developed in section 5.3, I will illustrate which items can be classified as good or poor quality mathematics questions. A comparison of good and poor quality mathematics questions in each of the PRQ and CRQ assessment formats will be made. Furthermore, I draw conclusions from my research about which of the mathematics assessment components, as defined in section 5.1, can be successfully assessed with respect to each of the two assessment formats, PRQ and CRQ.

In this way, I endeavour to probe and clarify the first two research subquestions as stated in section 3.2 i.e. How do we measure the quality of a good mathematics question? and; Which of the mathematics assessment components can be successfully assessed using the PRQ assessment format and which of the mathematics assessment components can be successfully assessed using the CRQ assessment format?

7.1 GOOD AND POOR QUALITY MATHEMATICS QUESTIONS

Section 7.1 summarises the development and features of the QI model for the sake of completeness of this chapter.

In section 5.3, the Quality Index (QI) was defined in terms of the three measuring criteria: discrimination, confidence deviation and expert opinion deviation. Each of these three criteria represented the three arms of a radar plot. In the proposed QI model, all three criteria were considered to be equally important in their contribution to the overall quality of a question.

The QI model can be used both to quantify and visualise how good or how poor the quality of a mathematics question is. The following three features of the radar plots could assist us to visualise the quality and the difficulty of the item:

(1) the shape of the radar plot;
(2) the area of the radar plot;
(3) the shading of the radar plot.

1. Shape of the radar plot

When comparing the radar plots for the good quality items with those of the poor quality items, it is evident that the shapes of these radar plots are also very different. For the good mathematics questions, the shape seems to resemble a small equilateral triangle. This ideal shape is achieved when all three arms of the radar plot are shorter than the average length of 0.5 on each axis i.e. are all very close to 0, as well as all three arms being almost equal in magnitude. Such a situation would be ideal for a mathematics question of good quality, since all three measuring criteria would be close to zero which indicates a small deviation from the expected confidence level as well as a small deviation from the expected student performance, and would also indicate an item that discriminates well. In contrast, those radar plots corresponding to items of a poor quality did not display this small equilateral triangular shape. One notices that these radar plots are skewed in the direction of one or more of the three axes. This skewness in the shape of the radar plot reflects that the three measuring criteria do not balance each other out. The axis towards which the shape is skewed reflects which of the criteria contribute to the overall poor quality of the question. However, there are poor quality items which have radar plots resembling the shape of a large equilateral triangle. The difference is that although the plot has three arms equal in magnitude, all three arms are longer than the average length of 0.5 and are in fact all very close to 1 (i.e. very far from 0).

2. Area of the radar plot

Another visual feature of the radar plot is its area. In this study, the area of the radar plot represents the Quality Index (QI) of the item. By defining the QI as the area, a balance is obtained between the three measuring criteria. If the QI value is less than 0.282 (the median QI), then the question is classified as a good quality mathematics question. If the QI value is greater than or equal to 0.282, the question is considered to be of a poor quality. When investigating the area of the good quality items, it is evident that such items have a small area i.e. a QI value close to zero. In such radar plots, the three arms are all shorter than the average length of 0.5 on each axis, and are all close to 0. For the poor quality items, the corresponding radar plot has a large area with QI values far from 0 (i.e. close to 1). In such radar plots, the three arms are generally longer than the average length of 0.5 on each axis, and are all far away from 0. The closer the QI value is to 0, the better the quality of the question.

We can conclude that both the area and the shape of the radar plot assist us to form an opinion on the quality of a question.

In Figure 7.1, both the shape and the area of the radar plot indicate a good quality assessment item. The shape resembles an equilateral triangle and the area is small.

Figure 7.2 visually illustrates an assessment item of poor quality. The shape is skewed in the direction of both the discrimination and confidence axes and the radar plot has a large area. The poor performance of all three measuring criteria contributes to this item being a poor quality item. The item does not discriminate well and both students and experts misjudged the difficulty of the question. The large, skewed shape of the radar plot indicates an item of poor quality.

Figure 7.1: A good quality item.

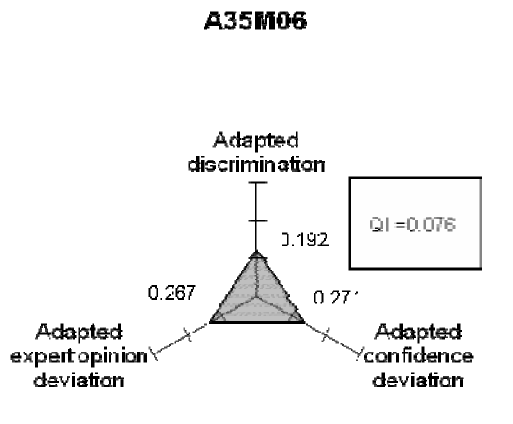
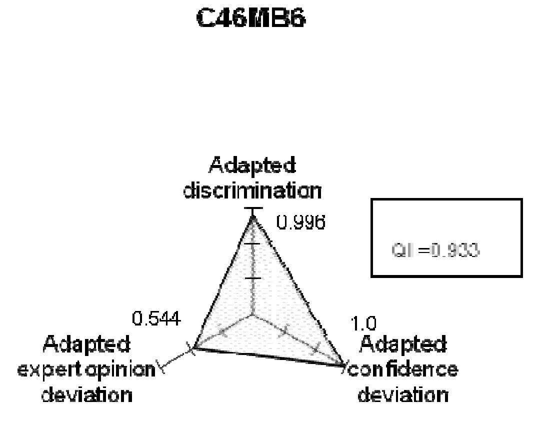


Figure 7.2: A poor quality item.



3. Shading of the radar plot

In this study, the shading of the radar plot helped us to visualise the difficulty level of the question. Six shades of grey, ranging from white through to black (as shown in Table 5.4), represented the six corresponding difficulty levels chosen in this study ranging from very easy through to very difficult. Difficulty level is an important parameter, but does not contribute to classifying a question as good or not. Both easy questions and difficult questions can be classified as good or poor. Not all difficult questions are of a good quality, and not all easy questions are of a poor quality. For example, in Figure 7.3, the dark grey shading of the radar plot represents a difficult item. The large area and skew shape of the plot represents a poor quality item. So Figure 7.3 visually represents a difficult, poor quality item. In Figure 7.4, the very light shading of the radar plot represents an easy item. The small area and shape of the radar plot represents a good quality item. So Figure 7.4 visually represents an easy, good quality item.

Figure 7.3: A difficult, poor quality item.

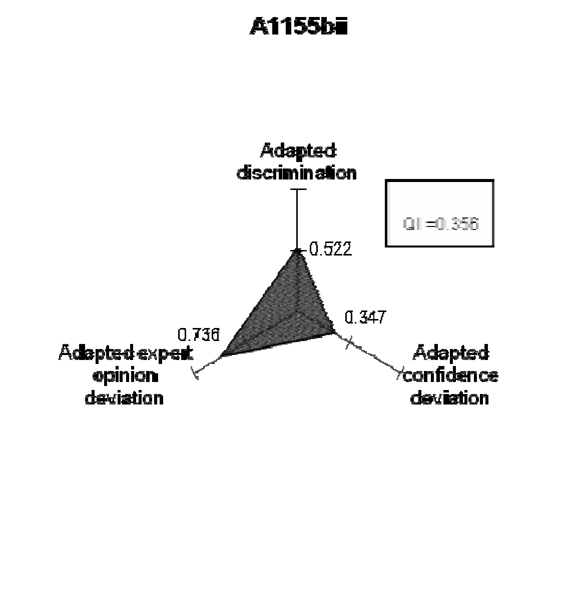
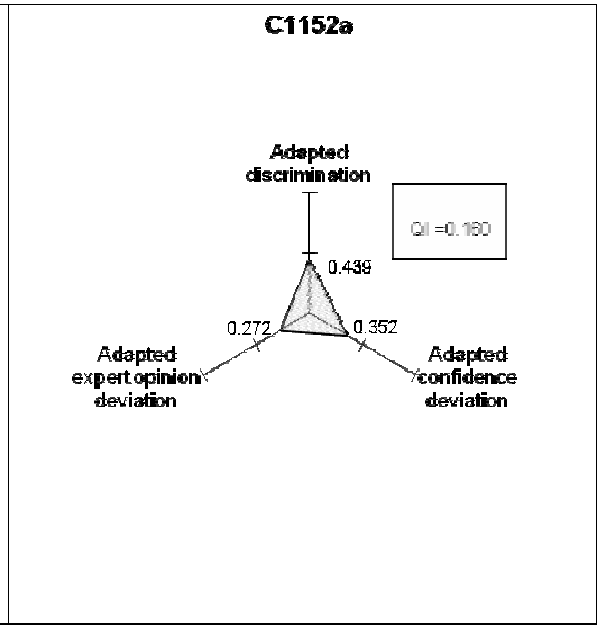


Figure 7.4: An easy, good quality item.



7.2 A COMPARISON OF PRQs AND CRQs IN THE MATHEMATICS ASSESSMENT COMPONENTS

In section 6.3, Table 6.3 summarised the quality of both PRQs and CRQs within each assessment component. It was noted that certain assessment components lend themselves better to PRQs than to CRQs. For example, in the **technical assessment component**, there were almost twice as many good quality PRQs than good quality CRQs. For the assessor, this means that the PRQ assessment format can be successfully used to assess mathematics content which requires students to adopt a routine, surface learning approach. In this component, PRQs can successfully assess content which students will have been given in lectures or will have practised extensively in tutorials. In addition there were more than twice as many poor quality CRQs than poor quality PRQs. The conclusion is that the PRQ format successfully assesses cognitive skills such as manipulation and calculation, associated with the technical assessment component.

Another component in which PRQs can be used successfully is the **disciplinary assessment component**. In this component, there was no difference between the good quality PRQs and the poor quality PRQs, with very little difference between the good quality CRQs and the poor quality CRQs. The PRQ format can be used to assess cognitive skills involving recall (memory) and knowledge (facts) equally successfully as the CRQ format. Thus in the disciplinary assessment component, results show that it is easy to set PRQs of a good quality, thus saving time in both the setting and marking of questions involving knowledge and recall.

As we proceed to the higher order **conceptual assessment component**, it is once again encouraging that the results indicate that PRQs can hold more than their own against CRQs. PRQs could be used successfully as a format of assessment for tasks involving comprehension skills whereby students are required to apply their learning to new situations or to present information in a new or different way. The results challenge the viewpoint of Berg and Smith (1994) that PRQs cannot successfully assess graphing abilities. The shift away from a surface approach to learning to a deeper approach, as mentioned by Smith *et al.* (1996), can be just as successfully assessed with PRQs as with the more traditional open-ended CRQs. The conclusion is that the PRQ assessment format can be successfully used in the conceptual assessment component.

The **modelling assessment component** tasks, requiring higher order cognitive skills of translating words into mathematical symbols, have traditionally been assessed using the CRQ format. The results from this study show that although there are few PRQs corresponding to this component, probably due to the fact that it is more difficult to set PRQs than CRQs of a modelling nature, the PRQs were highly successful. The perhaps somewhat surprising conclusion is that PRQs can be used very successfully in the modelling component. This result disproves the claim made by Gibbs (1992) that one of the main disadvantages of PRQs is that they do not measure the depth of student thinking. It also puts to rest the concern expressed by Black (1998) and Resnick & Resnick (1992) that the PRQ assessment format encourages students to adopt a surface learning approach. Although PRQs are more difficult and time consuming to set in the modelling assessment component (Andresen *et al.*, 1993), these results encourage assessors to think more about our attempts at constructing PRQs which require words to be translated into mathematical symbols. The results show that there is no reason why PRQs cannot be authentic and characteristic of the real world, the very objections made by Bork (1984) and Fuhrman (1996) against the whole principle of the PRQ assessment format.

Another very encouraging result was the high percentage of good quality PRQs as opposed to poor quality PRQs in the **problem solving assessment component**. This component encompasses tasks requiring the identification and application of a mathematical method to arrive at a solution. It appears that PRQs are slightly more successful than CRQs in this assessment component which encourages a deep approach to learning. Greater care is required when setting problem-solving questions, whether PRQs or CRQs, but the results show that PRQ assessment can add value to the assessment of the problem solving component. Once again this result shows that PRQs do not have to be restricted to the lower order cognitive skills so typical of a surface approach to learning (Wood & Smith, 2002).

The results indicate that PRQs were not as successful in the logical and consolidation assessment components. In the **logical assessment component**, there were noticeably more poor quality PRQs than poor quality CRQs. The nature of the tasks involving ordering and proofs lends itself better to the CRQ assessment format. There were very few good PRQs in the logical assessment component. The high percentage of the poor quality PRQs in the logical assessment component leads to the conclusion that this component lends itself better to CRQs than to PRQs.

In the **consolidation assessment component**, involving cognitive skills of analysis, synthesis and evaluation, there were noticeably more good quality CRQs than good quality PRQs. This trend towards more successful CRQs than PRQs indicates that CRQs add more value to the assessment of this component. This is not an unexpected result, as at this highest level of conceptual difficulty, assessment tasks require students to display skills such as justification, interpretation and evaluation. Such skills would be more difficult to assess using the PRQ format. However, as shown by many authors (Gronlund, 1988; Johnson, 1989; Tamir, 1990), the 'best answer' variety in contrast to the 'correct answer' variety of PRQs does cater for a wide range of cognitive abilities. In these alternative types of PRQs the student is faced with the task of carefully analysing the various options and of making a judgement to select the answer which best fits the context and the data given. The conclusion is that the consolidation assessment component encourages the educator or assessor to think more about their attempts at constructing suitable assessment tasks. According to Wood and Smith (2002), assessment tasks corresponding to a high level of conceptual difficulty should provide a useful check on whether we have tested all the skills, knowledge and abilities that we wish our students to demonstrate. As the results have shown, PRQs can be used as successfully as CRQs as an assessment method for those mathematics assessment components which require a deeper learning approach for their successful completion.

7.3 CONCLUSIONS

The mathematics assessment component taxonomy, proposed by the author in section 5.1, is hierarchical in nature, with cognitive skills that need a surface approach to learning at one end, while those requiring a deeper approach appear at the other end of the taxonomy. The results of this research study have shown that it is not necessary to restrict the PRQ assessment format to the lower cognitive tasks requiring a surface approach. The PRQ assessment format can, and does add value to the assessment of those components involving higher cognitive skills requiring a deeper approach to learning. According to Smith *et al.* (1996), many students enter tertiary institutions with a surface approach to learning mathematics and this affects their results at university. The results of this research study have addressed the research question of whether we can successfully use PRQs as an assessment format in undergraduate mathematics and the mathematics assessment component taxonomy was proposed to encourage a deep approach to learning. In certain assessment components, PRQs are more difficult to set than CRQs, but this should not deter the assessor from including the PRQ assessment format within these assessment components. As the discussion of the results has shown, good quality PRQs can be set within most of the assessment components in the taxonomy which do promote a deeper approach to learning.

In the Niss (1993) model, discussed in section 2.3, the first three content objects require knowledge of facts, mastery of standard methods and techniques and performance of standard applications of mathematics, all in typical, familiar situations. Results of this study have shown that PRQs are highly successful as an assessment format for Niss's first three content objects. As we proceed towards the content objects in the higher levels of Niss's assessment model, students are assessed according to their abilities to activate or even create methods of proofs; to solve open-ended, complex problems; to perform mathematical modelling of open-ended real situations and to explore situations and generate hypotheses. Results of this study again show that even though PRQs are more difficult to set at these higher cognitive levels, they can add value to the assessment at these levels.

Results of this study show that the more cognitively demanding conceptual and problem solving assessment components are better for CRQs. Traditional assessment formats such as the CRQ assessment format have in many cases been responsible for hindering or slowing down curriculum reform (Webb & Romberg, 1992). The PRQ assessment format can successfully assess in a valid and reliable way, the knowledge, insights, abilities and skills related to the understanding and mastering of mathematics in its essential aspects. As shown by the qualitative results, PRQs can provide assistance to the learner in monitoring and improving his/her acquisition of mathematical insight and power, while also improving their confidence levels. Furthermore, PRQs can assist the educator to improve his/her teaching, guidance, supervision and counselling, while also saving time. The PRQ assessment format can reduce marking loads for mathematical educators, without compromising the value of instruction in any way. Inclusion of the PRQ assessment format into the higher cognitive levels would bring new dimensions of validity into the assessment of mathematics.

Table 7.1 presents a comparison of the success of PRQs and CRQs in the mathematics assessment components.

Table 7.1: A comparison of the success of PRQs and CRQs in the mathematics assessment components.

Mathematics assessment Component	**Comparison of success**
1. Technical	PRQs can be used successfully
2. Disciplinary	No difference
3. Conceptual	PRQs can be used successfully
4. Logical	CRQs more successful
5. Modelling	PRQs can be used successfully
6. Problem solving	PRQs can be used successfully
7. Consolidation	CRQs more successful

As Table 7.1 illustrates, the enlightening conclusion is that there are only two components where CRQs outperform PRQs, namely the logical and consolidation assessment components. In two other components, PRQs are observed to slightly outperform CRQs, namely the conceptual and problem solving assessment components. The PRQs outperform the CRQs substantially in the technical and modelling assessment components. In one component there is no observable difference, the disciplinary assessment component.

7.4 ADDRESSING THE RESEARCH QUESTIONS

In this study, a model has been developed to measure the quality of a mathematics question. This model, referred to as the Quality Index (QI) model, was used to address the research question and subquestions as follows:

Research question:

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

Subquestion 1:

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

Subquestion 2:

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

Subquestion 3:

What are student preferences regarding different assessment formats?

- ### Addressing the first subquestion:

There is no single way of measuring the quality of a good question. I, as author of the thesis, have proposed one model as a measure of the quality of a question. I have illustrated the use of this model and found it to be an effective and quantifiable measure.

The QI model can assist mathematics educators and assessors to judge the quality of the mathematics questions in their assessment programmes, thereby deciding which of their questions are good or poor. Retaining unsatisfactory questions is contrary to the goal of good mathematics assessment (Kerr, 1991). Mathematics educators should optimise both the quantity and the quality of their assessment, and thereby optimise the learning of their students (Romberg, 1992).

The QI model for judging how good a mathematics question is has a number of apparent benefits. The model is visually satisfying; whether a question is of good or poor quality can be witnessed at a single glance. Visualising the difficulty level in terms of shades of grey adds convenience to the model. Another visual advantage of this model is that shortcomings in different aspects of an item, such as that experts completely under estimate the expected level of student performance in the particular item, can also be instantly visualised. In addition, the model provides a quantifiable measure of the quality of a question, an aspect that makes the model useful for comparison purposes. The fact that the model can be applied to judge the level of difficulty of both PRQs and CRQs makes it useful for both traditional "long question" environments, as well as the increasingly popular online, computer centred environments.

- Addressing the second subquestion:

In terms of the mathematics assessment components, it was noted that certain assessment components lend themselves better to PRQs than to CRQs. In particular, the PRQ format proved to be more successful in the technical, conceptual, modelling and problem solving assessment components, with very little difference in the disciplinary component, thus representing a range of assessment levels from the lower cognitive levels to the higher cognitive levels. Although CRQs proved to be more successful than PRQs in the logical and consolidation assessment components, PRQs can add value to the assessment of these higher cognitive component levels. Greater care is needed when setting PRQs in the logical and consolidation assessment components. The inclusion of the PRQ format in all seven assessment components can reduce marking loads for mathematics educators, without compromising the validity of the assessment. The PRQ assessment format can successfully assess in a valid and reliable way. The results have shown, both quantitatively and qualitatively, that PRQs can improve students' acquisition of mathematical insight and knowledge, while also improving their confidence levels. The PRQ assessment format can be used as successfully as the CRQ format to encourage students to adopt a deeper approach to the learning of mathematics.

- Addressing the third subquestion:

With respect to the student preferences regarding different mathematics assessment formats, the results from the qualitative investigation seemed to indicate that there were two distinct camps; those in favour of PRQs and those in favour of CRQs. Those in favour of PRQs expressed their opinion that this assessment format did promote a higher conceptual level of understanding and greater accuracy; required good reading and comprehension skills and was very successful for diagnostic purposes. Those in favour of CRQs were of the opinion that this assessment format promoted a deeper learning approach to mathematics; required good reading and comprehension skills; partial marks could be awarded for method and students felt more confident with this more traditional approach. Furthermore, from the students' responses, it also seemed as if the weaker ability students preferred the CRQ assessment format above the PRQ assessment format. The reasons for this preference were varied: CRQs provide for partial credit; there was a greater confidence with CRQs than with PRQs; PRQs require good reading and comprehension skills; PRQs encourage guessing and the distracters cause confusion.

- Addressing the main research question:

As this study aimed to show, PRQs can be constructed to evaluate higher order levels of thinking and learning, such as integrating material from several sources, critically evaluating data and contrasting and comparing information. The conclusion is that PRQs can be successfully used as an assessment format in undergraduate mathematics, more so in some assessment components than in others.

7.5 LIMITATIONS OF STUDY

The tests used in this study were conducted with tertiary students in their first year of study at the University of the Witwatersrand, Johannesburg, enrolled for the mainstream Mathematics I Major course. The study could be extended to other tertiary institutions and to mathematics courses beyond the first year level.

The judgement of how good or poor a mathematics question is, is modulo the QI model developed in this study. In the proposed QI model, I assumed that the three arms of the radar plot contribute equally to the overall quality of the mathematics question. This assumption needs to be investigated.

The qualitative component of this study was not the most important part of the research. The small sample of students interviewed was carefully selected to include differences in mathematical ability, from different racial backgrounds and different gender classes. Consequently, I regarded their responses as being indicative of the opinions of the Mathematics I Major cohort of students. The third research subquestion, dealing with student preferences regarding the different assessment formats, was included as a small subsection of the study and was not the main focus of this study. The qualitative component could be expanded in future by increasing the sample size of interviewees and by using questionnaires in which all the students in the first year mathematics major course could be asked to express their feelings and opinions regarding different mathematics assessment formats.

7.6 IMPLICATIONS FOR FURTHER RESEARCH

Collection of confidence-level data in conceptual mathematics tests provides valuable information about the quality of a mathematics question. The analysis suggests that confidence of responses should be collected, but also that it is critical to consider not only students' overall confidence but to consider separately confidence in both correct and incorrect answers. The prevalence of overconfidence in the calibration of performance presents a paradox of educational practice.

> On the one hand, we want students to have a healthy sense of academic self-concept and persist in their educational endeavours. On the other hand, we hope that a more realistic understanding of their limitations will be the impetus for educational development. The challenge for educators is to implement constructive interventions that lead to improved calibration and performance

without destroying students' self-esteem and confidence (Bol & Hacker, 2008, p2).

In this study, three parameters were identified to measure the quality of a mathematics question: discrimination index, confidence index and expert opinion. Further work needs to be carried out to investigate whether more contributing measuring criteria can be identified to measure the overall quality of a good mathematics question, and how this would affect the calculation of the Quality Index (QI) as discussed in section 5.3.2. As the assumption was made that the three parameters contributed equally to the quality of a mathematics question, the QI was defined as the area of the radar plot. The QI model could be adjusted or refined using other formulae.

It is common practice in the South African educational setting to use raw scores in tests and examinations as a measure of a student's ability in a subject. According to Planinic *et al.* (2006), misleading and even incorrect results can stem from an erroneous assumption that raw scores are in fact linear measures. Rasch analysis, the statistical method used in this research, is a technique that enables researchers to look objectively at data. The Rasch model (1960), can provide linear measures of item difficulties and students' confidence levels. Often, analysis of raw test score data or attitudinal data is carried out, but it is not always the case that such raw scores can be immediately assumed to be linear measures, and linear measures facilitate objective comparison of students and items (Planinic *et al.* 2006). According to Wright and Stone (1979), the Rasch model is a more precise and moral technique that can be used to comment on a person's ability and that the introduction thereof is long overdue. The Rasch method of data analysis could be valuable for other researchers in the fields of mathematics and science education research.

It might be important for mathematics educators and researchers to further explore the QI model with questions not limited to Calculus and Linear Algebra topics of many traditional first year tertiary mathematics courses. In doing so, mathematics educators and assessors can be provided with an important model to improve the overall quality of their assessment programmes and enhance student learning in mathematics.

This research study could be expanded to other universities. Tertiary mathematics educators need to use models of the type developed in this study to quantify the quality of the mathematics questions in their undergraduate mathematics assessment programmes. The QI model can also be used by tertiary mathematics educators to design different formats of assessment tasks which will be significant learning experiences in themselves and will provide the kind of feedback that leads to success for the individual student, thus reinforcing positive attitudes and confidence levels in the students' performance in undergraduate mathematics.

> The way students are assessed influences what and how they learn more than any other teaching practice (Nightingale *et al.*, 1996, p7).

Good quality assessment of students' knowledge, skills and abilities is crucial to the process of learning. In this research study, I have shown that the more traditional CRQ format is not always the only and best way to assess our students in undergraduate mathematics. PRQs can be constructed to evaluate higher order levels of thinking and learning. The research study conclusively shows that the PRQ format can be successfully used as an assessment format in undergraduate mathematics.

As mathematics educators and assessors, we need to radically review our assessment strategies to cope with changing conditions we have to face in South African higher education.

> The possibility that innovative assessment encourages students to take a deep approach to their learning and foster intrinsic interest in their studies is widely welcomed (Brown & Knight, 1994, p24).