CHAPTER 4: QUALITATIVE INVESTIGATION

In this chapter I address the third research subquestion:

What are student preferences regarding different assessment formats?

4.1 QUALITATIVE DATA ANALYSIS

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

4.2 QUALITATIVE INVESTIGATION

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

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

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

- In favour of alternate assessment formats

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

> Interviewer: You saw that a percentage of your tests was multiple choice and a percentage was always long questions and your tutorial tests were only multiple choice. Did you like those different formats?
>
> Candidate: Ja, I did, ‘cos multiple choice gives you an option of , y’know, the right answer’s there somewhere so it kind of relieves you a bit and then you balance it off with a nice, um, long question so it’s not... you aren’t just depending on your luck but you’re also applying your knowledge and I think that’s.. that’s cool.

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

- MCQs test a higher conceptual level

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

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

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

Interviewer: So what does it represent?

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

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

Candidate: I think so.

Interviewer: Deeper?

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

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

- CRQs provide for partial credit

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

Interviewer: Which type of question do you prefer?

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

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

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

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

- Confidence plays an important role in assessment

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

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

Candidate: Fluctuated. Bit of a roller coaster.

Interviewer: Can you explain what you mean?

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

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

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

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

Candidate: We were, um, not as like... not such a broad spectrum but we were. We didn't really do MCQ as such in Maths but um... I think it… ja… the MCQs are definitely challenging because, I don't know, in most subjects they are, you know, like…

Interviewer: What makes them challenging?

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

Interviewer: Okay.

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

- MCQs require good reading and comprehension skills

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

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

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

Interviewer: Okay.

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

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

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

Interviewer: Okay.

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

Interviewer: More graph interpretation questions?

Candidate: Yes.

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

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

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

- PRQ format lends itself to guessing and cheating

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

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

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

Interviewer: Why?

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

Interviewer: Oh, so it's guessing?

Candidate: Ja! Ja, guessing, guessing.

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

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

Interviewer: Okay.

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

- Alternate formats add depth to assessment

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

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

Candidate: I think about seventy percent should be MCQ and the rest should be long questions because it's... sometimes it's harder to understand than MCQ questioning despite understanding the knowledge, you know, understanding the maths and the theory that you get 'cos it's very tricky sometimes. But I think it separates like your A's from your B's, you know, your like seventy-fives from your sixties. It's a good way to see what type of student you are.

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

- Diagnostic purpose

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

Interviewer: Do you like the different formats of assessment in your maths tests?

Candidate: Um, no, it's okay, but… Ja I think that… no, the papers have been up to standard so far. I don't think there really is a problem, especially like, um, the MCQs I felt really like gives you… it really tests your understanding of how to, you know, of all your calculations and stuff. I don't really think there's a problem with the way we've been tested so far.

Interviewer: Which type of questions do you prefer, MCQs or traditional long questions?

Candidate: Well, personally, I don't like the MCQs because sometimes you think you've got the right answer but, you know, you might have made a mistake somewhere in your calculations. You saw it or your right answer there then… but I think that the MCQs are probably designed that way. Like you would have probably picked up what kind of mistakes we would have made so… so I think, ja, there should be a variety of different questions.

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

- Distracters can cause confusion

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

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

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

This candidate was a poorly performing student, who admitted earlier in the interview that he had not been taking his studies seriously. He had not been attending classes regularly and had not studied for his tests. He did not have any preference for the type of assessment format, although he did feel more confident with the CRQ format. His lack of confidence in the MCQs was linked to the fact that often the distracters were very similar to each other and he found it difficult to make the correct choice. He did not have enough confidence to trust his calculation of the correct answer, and when faced with the situation of two answers very close in value or nature to each other, he doubted his calculation. This lack of confidence was also evident in his performance in the CRQ format.

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

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

From the students' responses, it seems as if the weaker students prefer CRQs. These students expressed a lack of confidence in PRQs, with one of the interviewees justifying her lack of confidence in this assessment format as a 'mental block'. The weaker students seemed to perform better in CRQ assessment format, thus resulting in a greater confidence in this format. The attitudes of weaker students to the PRQ format illustrate the important role that confidence plays in assessment. Weaker ability students also felt threatened by the fact that if their answer to an MCQ was incorrect, no marks were awarded, whereas with CRQs, partial marks were awarded even if the answer was incorrect. Weaker students often lack the necessary reading and comprehension skills required to answer MCQs successfully. One of the weaker students opposing MCQs felt that the PRQ format lends itself to 'guessing and cheating'. The weaker ability students also expressed their frustration with MCQs if two or more of the distracters were very similar to each other. They felt that distracters can cause confusion, and this in turn would affect their performance.

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

CHAPTER 5: THEORETICAL FRAMEWORK

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

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

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

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

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

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

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

5.1 MATHEMATICS ASSESSMENT COMPONENTS

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

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

Corresponding to Niss's assessment model (Niss, 1993) reviewed in Section 2.3, in this proposed set of seven mathematics assessment components, questions involving manipulation and calculation would be regarded as *technical.* Those that rely on memory and recall of knowledge and facts would fall under the *disciplinary* component. Assessment components (1) and (2) include questions based on mathematical facts and standard methods and techniques. The *conceptual* component (3) involves comprehension skills with algebraic, verbal, numerical and visual (graphical) questions linked to standard applications. The assessment components (4), (5) and (6) correspond to the *logical* ordering of proofs, *modelling* with translating words into mathematical symbols and *problem solving* involving word problems and finding mathematical methods to come to the solution. Assessment component (7), *consolidation*, includes the processes of synthesis (bringing together of different topics in a single question), analysis (breaking up of a question into different topics) and evaluation requiring exploration and the generation of hypothesis.

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

A modification of Bloom's taxonomy, adapted for assessment, called the MATH taxonomy (Smith *et al.*, 1996) was discussed in Section 2.4 in the literature review. The MATH taxonomy has eight categories, falling into three main groups. Group A tasks include those tasks which require the skills of factual knowledge, comprehension and routine use of procedures. In the proposed mathematics assessment component taxonomy, assessment components (1) and (2) -Technical and Disciplinary, would correspond to these Group A tasks. In the MATH taxonomy Group B tasks, students are required to apply their learning to new situations, or to present information in a new or different way. Such tasks require the skills of information transfer and applications in new situations, and would correspond to assessment components (3) - Conceptual and (4) - Logical. The third group in the MATH taxonomy, Group C encompasses the skills of justification, interpretation and evaluation. Such skills would relate to the mathematics assessment components (5) - Modelling, (6) - Problem solving and (7) - Consolidation. One of the main differences between Bloom's taxonomy and the MATH taxonomy is that the MATH taxonomy is context specific and is used to classify tasks ordered by the nature of the activity required to complete each task successfully, rather than in terms of difficulty.

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

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

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

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

Table 5.2: Mathematics assessment component taxonomy and cognitive skills.

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

5.1.1 Question examples in assessment components

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

Assessment Component 1: Technical

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

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

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

C. $3 - 4i$

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

MATH109 August 2005, Tutorial Test, Question 5.

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

Assessment Component 2: Disciplinary

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

A. f is not a function.

B. f is an even function.

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

D. f is an odd function.

MATH109 March 2005, Tutorial Test A, Question 1.

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

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

Assessment Component 3: Conceptual

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

on the interval $[-3,0]$

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

B. f is not continuous

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

D. Both A and B

E. None of the above

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

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

Assessment Component 3: Conceptual

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

A. 2

B. e^2

C. ∞

D. 1

E. Does not exist

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

In the *conceptual* question above, the student is required to apply his/her knowledge of the definition of Euler's number e, which is defined in lectures as:

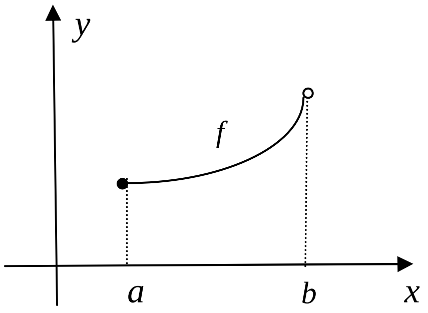
$$\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x = e$$

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

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

Assessment Component 3: Conceptual

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



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

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

C. No extrema.

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

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

In this *graphical conceptual* question, students are required to apply their knowledge of the Extreme Value theorem and the definition of relative extrema on an interval I. There is no algebraic calculation necessary of the values of the extrema on the closed interval [a,b]. The Extreme Value theorem is an existence theorem because it tells of the existence of minimum and maximum values, but does not show how to find these values. Students need to examine the graph of the given function f and consider how f behaves at the end points as well as how the continuity (or lack of it) has affected the existence of extrema on the given interval. The choice of the correct option is assisted by having a *visual* figure when the decision is made.

Assessment Component 4: Logical (PRQ)

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

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

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

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

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

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

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

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

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

Assessment Component 4: Logical (CRQ)

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

Theorem:

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

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

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

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

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

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

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

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

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

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

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

Correct order: (Only list the statement numbers.)

(b) What is the theorem called?

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

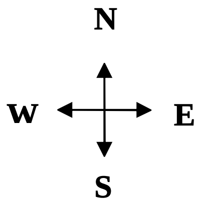
MATH109 June 2006, Section C: Calculus, Question 4.

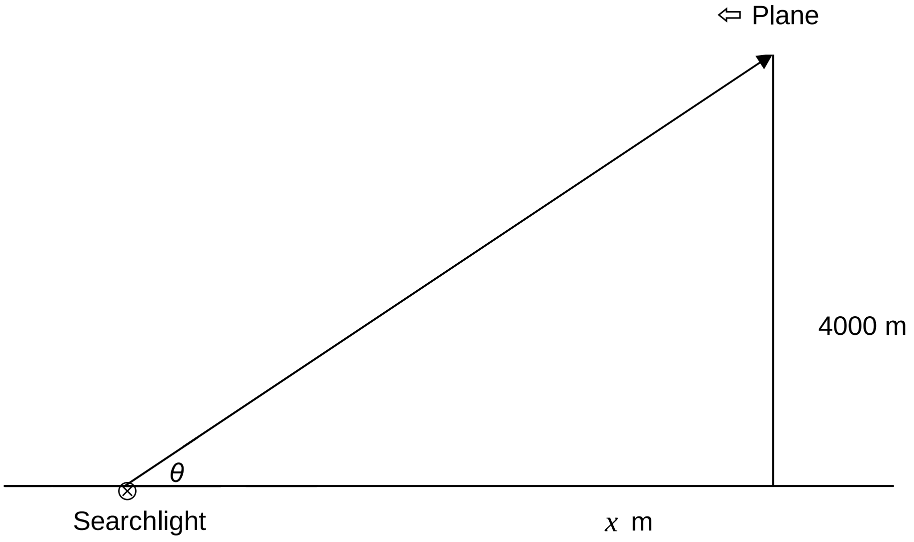
This *logical* CRQ requires the students to recall the proof of the Fundamental Theorem of Calculus. Although the proof is given, the statements appear in the incorrect order. The students are required to reorder the given statements to correct the proof. Such a reordering process involves the cognitive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

Assessment Component 5: Modelling (CRQ)

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

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





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

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

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

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

Assessment Component 6: Problem solving (PRQ)

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

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

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

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

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

E. None of the above.

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

In this *problem solving* MCQ, the student is required to find his or her own method to arrive at the solution. Firstly, the student has to know what the antiderivative of a function is in order to decide on a method. The solution can be arrived at by either integrating $f(x)$ using the technique of integration by parts, since $f(x)$ is a product of two differentiable functions, or by differentiating each function $F(x)$ provided in the distracters, using the Product Rule, until the original function $f(x)$ is obtained.

Assessment Component 6: Problem solving (CRQ)

This question deals with the statement

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

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

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

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

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

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

In the *problem solving* CRQ, the students are required to use the principle of Mathematical Induction to prove that the statement $P(n)$ is true for all natural numbers $n\geq 2$. The CRQ has been subdivided into smaller subquestions involving different cognitive skills to assist the student with the method of solving using mathematical induction. In subquestion (1.1), the students need to establish truth for $n=2$ by actually testing whether the statement $P(n)$ is true for $n=2$. Hence (1.1) assess within the technical mathematics assessment component. Subquestion (1.2) involves a numerical calculation, the result of which will be used in the proof by induction. Hence (1.2) also assesses within the technical assessment component. In subquestion (1.3), students are required to complete the proof by induction, by assuming the inductive hypothesis that $P(k)$ is true for $k > 2, k \in \mathrm{N}$, and proving that $P(k+1)$ is true. Since subquestion (1.3) requires the cognitive skills of identifying and applying the principle of Mathematical Induction to arrive at a solution, (1.3) assesses within the problem solving mathematics assessment component. Subquestion (1.4) concludes the proof by requiring the students to justify that both of the conditions of the principle hold, and therefore by the principle of induction $P(n)$ is true for every $n \geq 2, n \in \mathrm{N}$. Hence (1.4), requiring no more than a simple manipulation, assesses within the technical assessment component. This problem solving CRQ illustrates that often those questions involving higher order cognitive skills subsume the lower order cognitive skills.

Assessment Component 7: Consolidation (PRQ)

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

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

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

C. $\{y | -\frac{\pi}{2} < y < \frac{\pi}{2}\}$

D. $\{y | -\frac{\pi}{2} \leq y \leq \frac{\pi}{2}\}$

E. None of the above.

MATH 109 May 2006, Section A: MCQ, Question 1.

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

Assessment Component 7: Consolidation (CRQ)

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

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

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

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

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

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

5.2 DEFINING THE PARAMETERS

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

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

- A good question should discriminate well. In other words, high performing students should score well on this question and poor performing students are not expected to do well.
- Students' confidence when dealing with the question should correspond to the level of difficulty of the question. There is a problem with a question when it is experienced as misleadingly simple by students and subsequently leads to an incorrect response. In this case, students are over confident and do not judge the level of difficulty of the question correctly. Similarly, there is a problem if a simple question is experienced as misleadingly difficult and students have no confidence in doing it.
- The level of difficulty of the question should be judged correctly by the lecturer. When setting a question, the lecturer judges the level of difficulty intuitively. There is a problem with the question when the lecturer over or underestimates the level of difficulty as experienced by students.
- The level of difficulty of a question does not make it a good or poor question. Difficult questions can be good or poor, just as easy questions can be.

With these premises as background, three parameters were identified:

(i) Discrimination index

(ii) Confidence index

(iii) Expert opinion

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

(iv) Level of difficulty

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

5.2.1 Discrimination index

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

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

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

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

N = number of students in both groups.

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

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

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

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

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

   A second issue involving test homogeneity is that of the precision of a student's obtained test score as an estimate of that student's "true" score on the skill tested. Precision (reliability) increases as the average item-test correlation increases.
2. Distracters for PRQs that are not chosen by any students should be replaced or eliminated. They are not contributing to the test's ability to discriminate the good students from the poor students. One should be suspicious about the correctness of any item in which a single distracter is chosen more often than all other options, including the answer, and especially so if the distracter's correlation with the total score is positive.
3. Items that virtually everyone gets right are unsuccessful for discriminating among students and should be replaced by more difficult items (Ebel, 1965).

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

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

$P_{vix} =$ probability that person v of ability β_v is observed in category x of a rating scale applied to item i with difficulty level δ_i;

$F_x =$ Rasch-Andrich threshold.

In Winsteps, item discrimination is not a parameter. It is merely a descriptive statistic. The Winsteps reported values of item discrimination are a first approximation to the precise value of a_i. The possible range of a_i is $-\infty$ to $+\infty$, where $+\infty$ corresponds to a Guttman data pattem (perfect discrimination) and $-\infty$ to a reversed Guttman pattem. The Guttman scale (also called 'scalogram') is a data matrix where the items are ranked from easy to difficult and the persons likewise are ranked from lowest achiever on the test to highest achiever on the test. Rasch estimation usually forces the average item discrimination to be near 1.0. An *estimated discrimination* of 1.0 accords with Rasch model expectations. Values greater than 1.0 indicate over-discrimination, and values less than 1.0 indicate under-discrimination. Over-discrimination is thought to be beneficial under classical (raw-score) test theory conventions (Linacre, 2005).

In classical test theory, the ideal item acts like a switch i.e. high performers pass, low performers fail. This is perfect discrimination, and is ideal for sample stratification. Such an item provides no information about the relative performance of low performers, or the relative performance of high performers. Rasch analysis, on the other hand, requires items that provide indication of relative performance along the latent variable as discussed in section 3.4. It is this information which is used to construct measures. From a Rasch perspective, over-discriminating items tend to act like switches, not measuring devices. Under-discriminating items tend neither to stratify nor to provide information about the relative performance of students on those items.

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

5.2.2 Confidence index

The *confidence index* (CI) has its origins in the social sciences, where it is used particularly in surveys and where a respondent is requested to indicate the degree of confidence he has in his own ability to select and utilise well-established knowledge, concepts or laws to arrive at an answer. In the science education literature, as well as the measurement literature (as discussed in section 2.14), a range of studies has considered some aspects of student confidence and how such confidence may impact students' test performance. Students' self-reported confidence levels have also been studied in the field of educational measurement to assess over- and underconfidence bias in students' test-taking practices (Pallier, Wilkinson, Danthiir, Kleitman, Knezevic, Stankov & Robertsw, 2002). In physics education research, Hasan *et al.* (1999) used a confidence index in conjunction with the correctness or not of a response, to distinguish between students' embedded misconceptions (wrong answer and high confidence) and lack of knowledge (wrong answer and low confidence) and to restrict guessing (Table 5.3). The CI is usually based on some scale. For example, in Hasan's (1999) study, a six-point scale (0 – 5) was used in which 0 implies no knowledge (total guess) of methods or laws required for answering a particular question, while 5 indicates complete confidence in the knowledge of the principles and laws required to arrive at the selected answer. When a student is asked to provide an indication of confidence along with each answer, we are in effect requesting him to provide his own assessment of the certainty he has in his selection of the laws and methods utilised to get to the answer (Webb, 1994).

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

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

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

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

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

Hasan *et al.* (1999) recommend that if the answers and related CI values indicate the presence of misconceptions, then feedback to students can be modified with the explicit intent of removing the misconceptions. Furthermore, the information obtained by utilising the CI can also be used to address other areas of instruction. In particular, it can be used:

- as a means of assessing the suitability of the emphasis placed on different sections of a course
- as a diagnostic tool, enabling the teacher to modify feedback
- as a tool for assessing progress or teaching effectiveness when both pre- and post-tests are administered
- as a tool for comparing the effectiveness of different teaching approaches, including technology-integrated approaches, in promoting understanding and problem-solving proficiency.

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

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

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

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

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

In this research study, the CI values per item were calculated according to a 4-point Likert scale in which 1 implied a ‘complete guess’, 2 implied a ‘partial guess’, 3 for ‘almost certain’, while 4 indicated ‘certain’. In terms of the Rasch model, a Likert scale is a format for observing responses wherein the categories increase in the level of the variable they define, and this increase is uniform for all agents of measurement. The polytomous Rasch-Andrich rating scale model, discussed in section 3.4.1.3, was used in the Winsteps calculation of the CI.

5.2.3 Expert opinion

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

1: student should find the question easy

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

3: student should find the question difficult or challenging.

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

While giving their opinions, experts could reflect on the learning outcomes of the course, and on the assessment components corresponding to each test item. Such reflection would assist experts to write questions that guide students towards the kinds of intellectual activities they wish to foster, and raise their awareness of the effects of the kinds of questions they ask on their students' learning. In this context, Hubbard (2001) refers to Ausubel's meaningful learning, Skemp's description of relational understanding, Tall's definition of different types of generalisation and abstraction and Dubinsky and Lewin's reflective abstraction as all investigating in different ways, the kinds of intellectual activities which we desire our students to engage in. The experts involved in giving their opinions were not asked to familiarise themselves with any of the above research papers. However, it was hoped that because they were successful mathematics thinkers themselves, the task of giving their opinions would enable them to recognise the intellectual activities required to solve different types of questions, in both the PRQ and CRQ formats.

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

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

5.2.4 Level of difficulty

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

In traditional test theory, difficulty level is defined as:

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

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

5.3 MODEL FOR MEASURING A GOOD QUESTION

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

5.3.1 Measuring criteria

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

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

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

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

(1) Point measure as a discrimination index

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

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

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

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

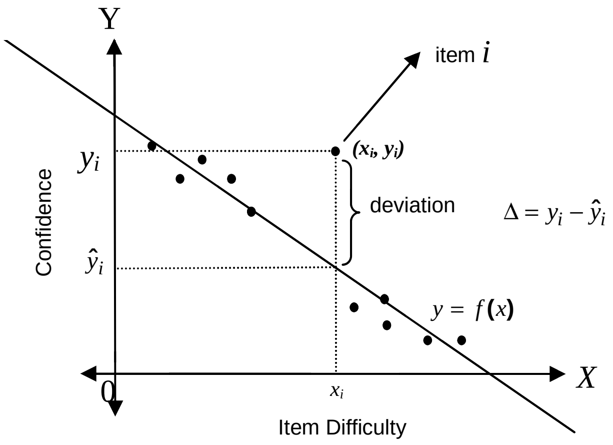
(2) Confidence deviation

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

1 : complete guess
2 : partial guess
3 : almost certain
4 : certain

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

Figure 5.1: Illustration of confidence deviation from the best fit line between item difficulty and confidence.



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

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

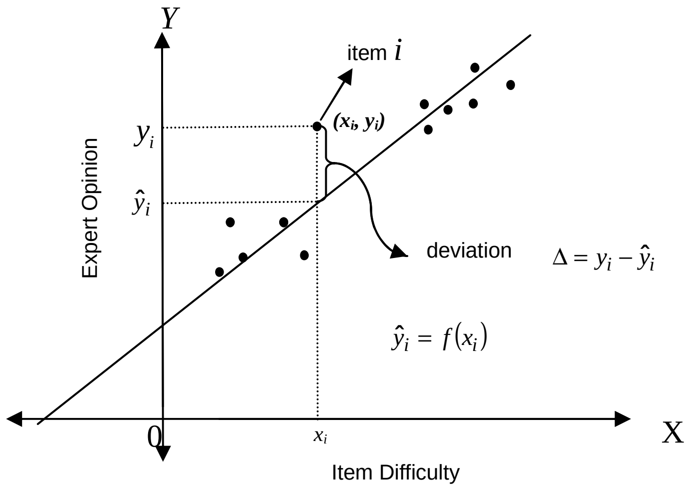
(3) Expert opinion deviation

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

1: student should find the question easy
2: student should find the question of average difficulty, but fair
3: student should find the question difficult or challenging.

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

Figure 5.2: Illustration of expert opinion deviation from the best fit line between item difficulty and expert opinion.



Figures 5.1 and 5.2 show that the larger the deviation of the predicted value from the observed value, the further the observed value is from the regression line and the worse the situation is in terms of an indication of quality.

5.3.2 Defining the Quality Index (QI)

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

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

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

Figure 5.3: Visual representation of the three axes of the QI.

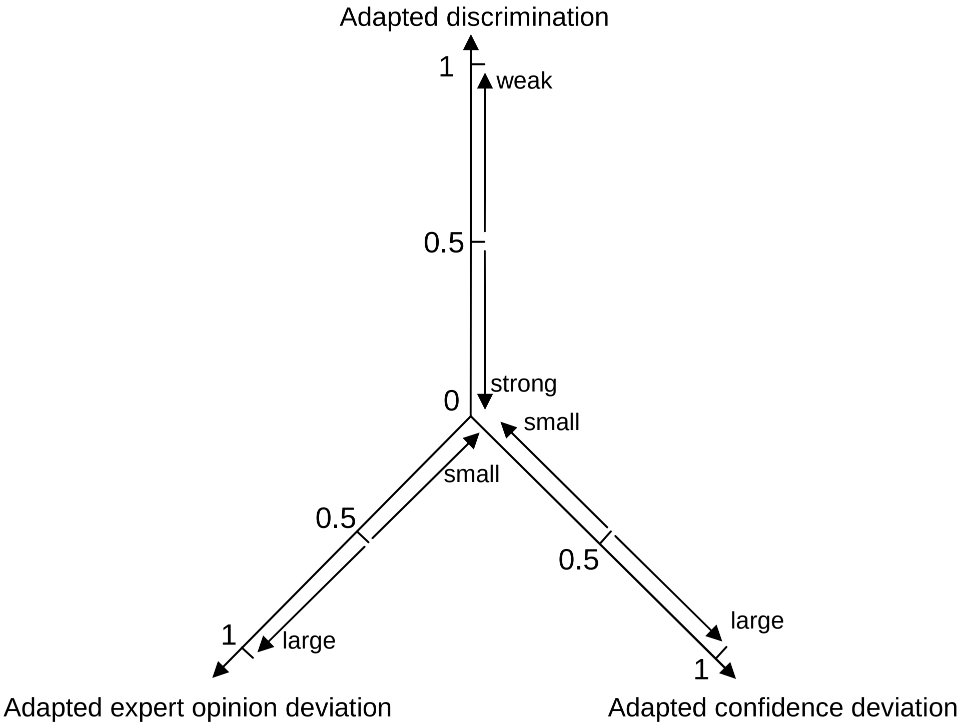
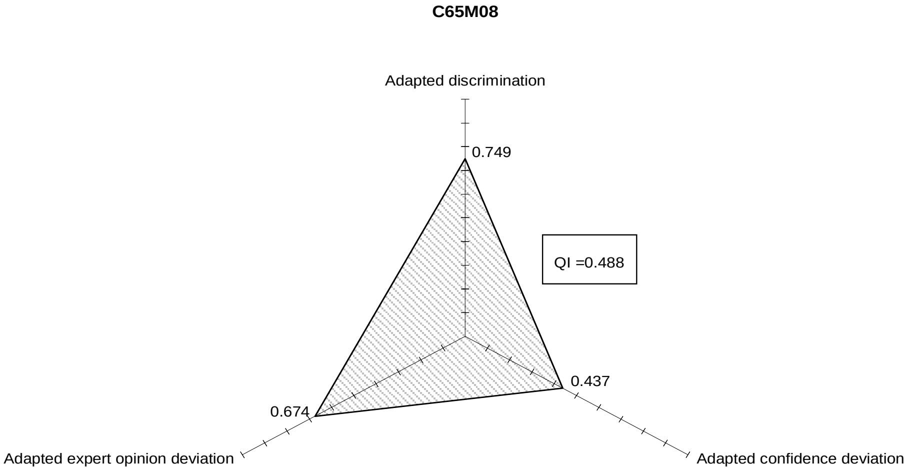


Figure 5.4 depicts an example of a radar plot.

Figure 5.4: Quality Index for PRQ



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

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

Discr = Adapted discrimination;

Conf dev = Adapted confidence deviation;

EO dev = Adapted expert opinion deviation

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

Good quality	:	QI < 0.282
Poor quality	:	QI ≥ 0.282

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

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

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

Figure 5.5: A good quality item.

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

CRQ, Algebra, June 2005, Q1b.

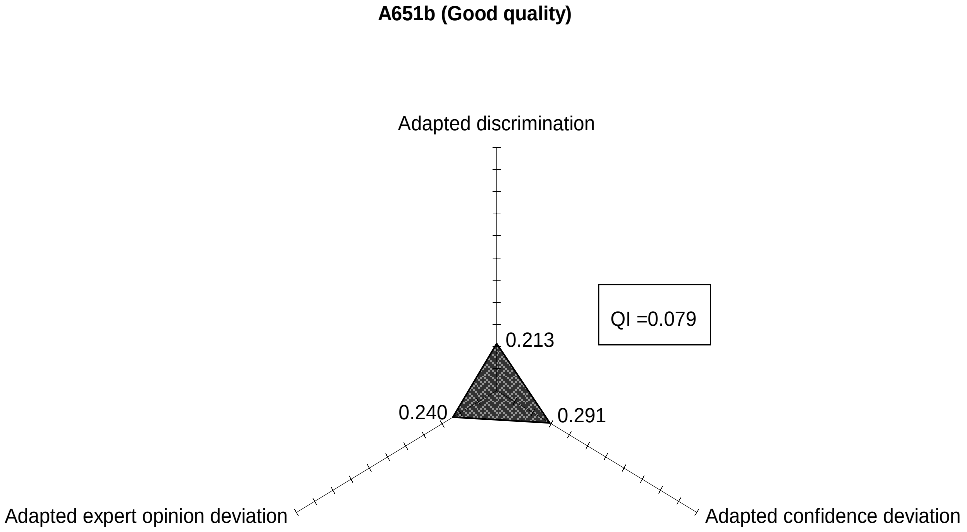


Figure 5.6: A poor quality item.

Consider the following theorem:

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

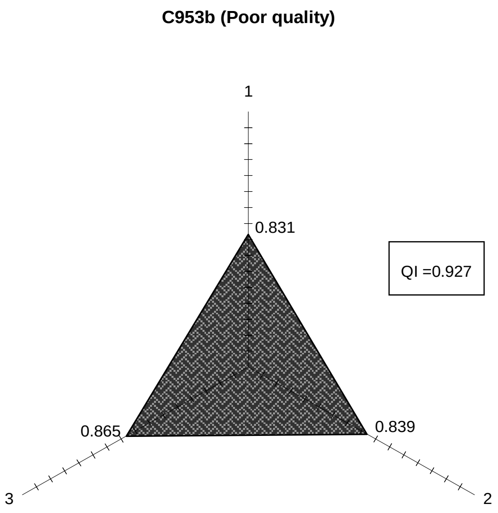
Consider the proof to this theorem:

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

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

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

CRQ, Calculus, September 2005, Q3b.

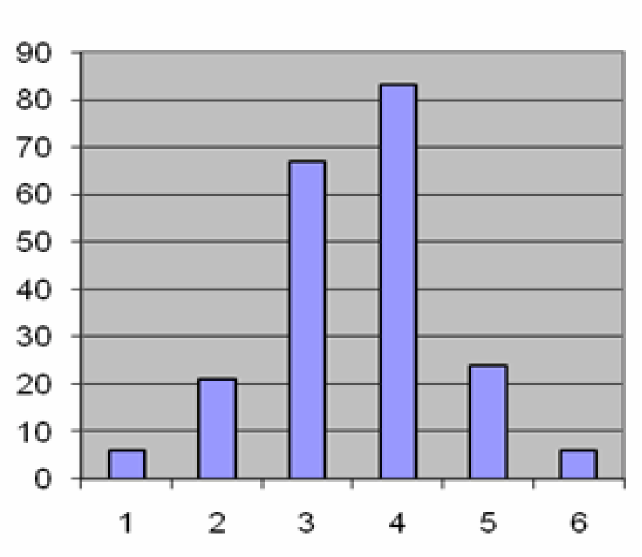


5.3.3 Visualising the difficulty level

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

In this study, the range of difficulty levels over the 207 test items was calculated to be a value of 0.12 using the maximum difficulty value of 4.56 and the minimum difficulty value of -5.56. The standard deviation for this range was calculated to be a value of 1.59. Using these parameters, the distribution of the difficulty levels was investigated by creating a histogram with six intervals of difficulty of 1.5 logits each, as indicated in Figure 5.7.

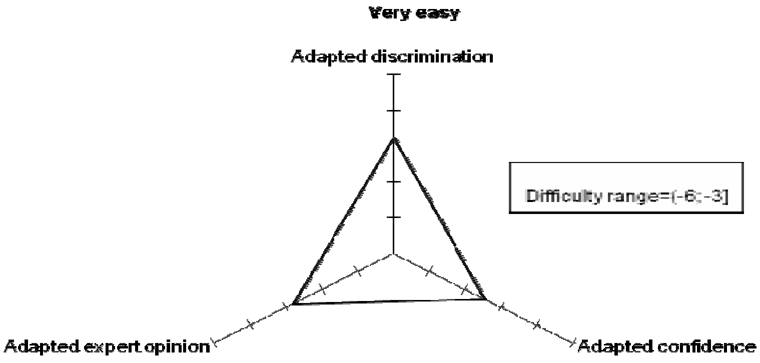
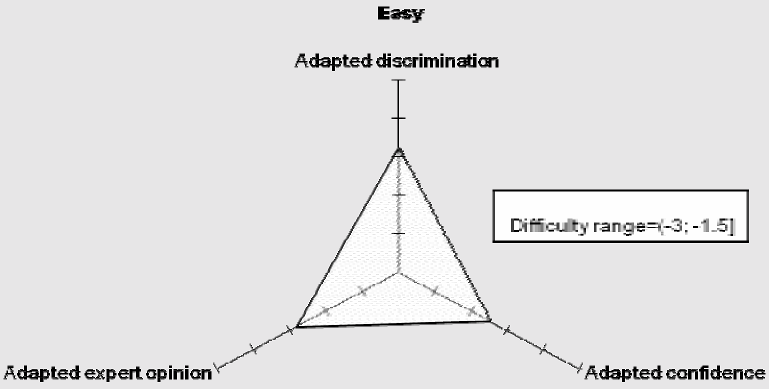
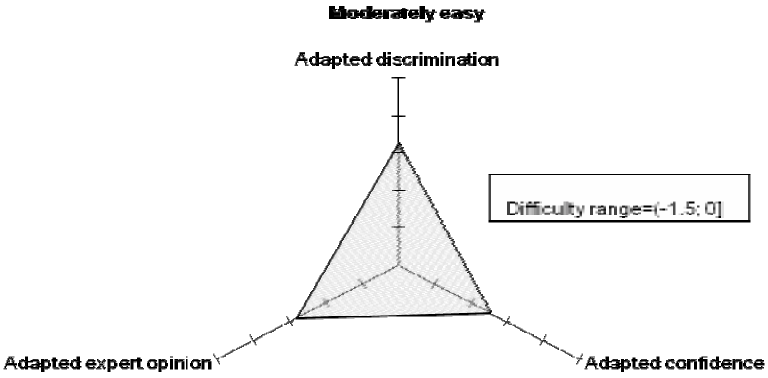
Figure 5.7: Distribution of six difficulty levels.

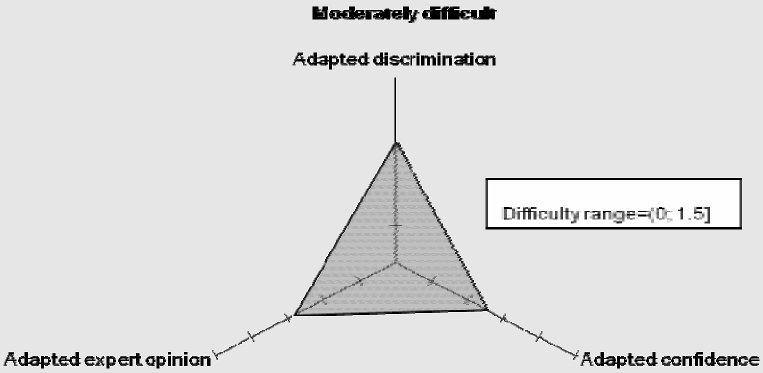
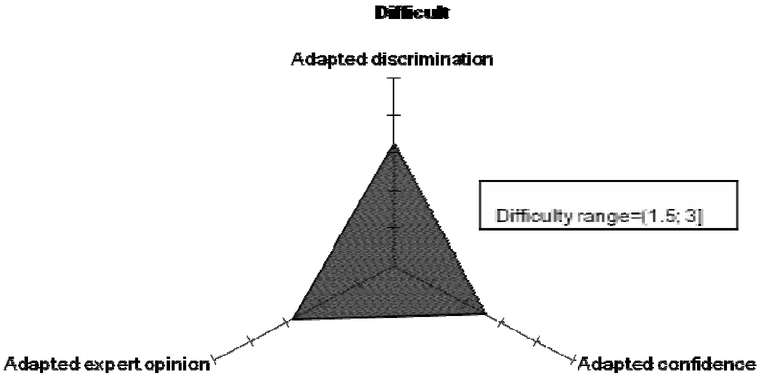
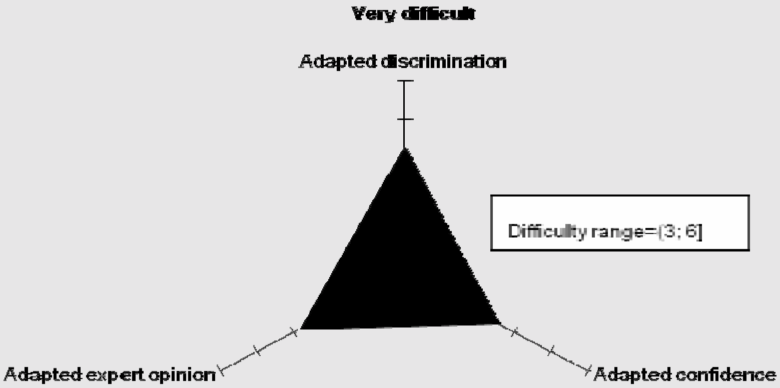


For each of the six intervals, a corresponding shading of the radar chart was chosen to represent the six difficulty levels: very easy; easy; moderately easy; moderately difficult; difficult; very difficult.

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

Table 5.4: Classification of difficulty intervals.

Interval	Degree of difficulty	Shading
(-6; -3]	Very easy	
(-3; -1.5]	Easy	
(-1.5; 0]	Moderately easy	

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

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