Can Multiple Choice Questions be Successfully Used as an Assessment Format in Undergraduate Mathematics?

Belinda Huntley; Johann Engelbrecht; Ansie Harding

University of Pretoria

huntley@sjc.co.za; johann.engelbrecht@up.ac.za; ansie.harding@up.ac.za

In this study we propose a taxonomy for assessment in mathematics, which we call the *assessment component taxonomy*, to identify those components of mathematics that can be successfully assessed using alternative assessment formats. Based on the literature on assessment models and taxonomies in mathematics, this taxonomy consists of seven mathematics assessment components, hierarchically ordered by cognitive level, as well as the nature of the mathematical tasks associated with each component. Using a model that we developed earlier for measuring the quality of mathematics test items, we investigate which of the assessment components can be successfully assessed in the provided response question (PRQ) format, in particular multiple choice questions (MCQs), and which can be better assessed in the constructed response question (CRQ) format. The results of this study show that MCQs can be constructed to evaluate higher order levels of thinking and learning. The conclusion is that MCQs can be successfully used as an assessment format in undergraduate mathematics, more so in some assessment components than in others. The inclusion of the PRQ assessment format in all seven assessment components can reduce the large marking loads, associated with continuous assessment practices in undergraduate mathematics, without compromising the validity of the assessment.

In South Africa, as in the rest of the world, the changes in society and technology have imposed pressures on academics to review current assessment approaches. Changes in education assessment are currently being called for, both within the fields of measurement and evaluation as well as in specific academic disciplines such as mathematics. Geyser (2004) summarises the paradigm shift that is currently under way in tertiary education as follows:

> The main shift in focus can be summarized as a shift away from assessment as an add-on experience at the end of learning, to assessment that encourages and supports deep learning. It is now important to distinguish between learning for assessment and learning from assessment as two complementary purposes of assessment. (p. 90)

Mathematics at tertiary level remains conservative in its use of alternative formats of assessment. As goals for mathematics education change to broader and more ambitious objectives (NCTM, 1989), such as developing mathematical thinkers who can apply their knowledge to solving real problems, a mismatch is revealed between traditional assessment and the desired student outcomes. It is no longer appropriate to assess student mathematical knowledge using general assessment taxonomies, because these taxonomies are not pertinent to mathematics and do not identify those levels of mathematics that can be assessed using alternative formats of assessment.

With this background, we propose a taxonomy of mathematics, which we call the *assessment component taxonomy*, to identify those components of mathematics that can be successfully assessed using alternative assessment formats. Using a model that we developed earlier (Huntley, Engelbrecht &

Harding, 2008; 2009) for measuring the quality of mathematical test items, we investigate which of the assessment components can be successfully assessed in the provided response question (PRQ) format, in particular multiple choice questions (MCQs), and which can be better assessed in the constructed response question (CRQ) format where students have to construct and supply their own responses (Engelbrecht & Harding, 2003).

Mathematics assessment models

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

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

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

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

(a) *Mathematical facts*, which include definitions, theorems, formulae, certain specific proofs and historical and biographical data.
(b) *Standard methods* and *techniques* for obtaining mathematical results. These include qualitative or quantitative conclusions, solutions to problems and display of results.
(c) *Standard applications* which include familiar, characteristic types of mathematical situations which can be treated by using well-defined mathematical tools.

To a lesser extent, objects of assessment also include:

(d) *Heuristic* and *methods of proof* as ways of generating mathematical results in non-routine contexts.
(e) *Problem solving* of non-familiar, open-ended, complex problems.
(f) *Modelling* of open-ended, real mathematical situations belonging to other subjects, using whatever mathematical tools at one's disposal.

In mathematics, we rarely encounter

(g) *Exploration* and *hypothesis generation* as objects of assessment.

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

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

In the Niss assessment model, objects (a) – (g) and the corresponding students' abilities are widely considered to be essential representations of what mathematics and mathematical activity are really about. The first three objects in the list emphasise routine, low-level features of mathematical work, whereas the

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

Assessment taxonomies

According to the World Book Dictionary (1990), a *taxonomy* is any classification or arrangement. Taxonomies are used to ensure that examinations contain a mix of questions to test skills and concepts. A leader in the use of a taxonomy for test construction and standardization was Ralph W. Tyler, the "father of educational evaluation" (Romberg, 1992, p. 19) who in 1931 reported on his efforts to construct achievement tests for various university courses.

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

Bloom's taxonomy has often been seen as fitting mathematics especially poorly (Romberg, Zarinnia & Collis, 1990). It is quite good for structuring assessment tasks, but Freeman and Lewis (1998) suggest that Bloom's taxonomy is not helpful in identifying which levels of learning are involved.

As Ormell (1974) noted in a strong critique of the taxonomy, Bloom's categories of behaviour "are extremely amorphous in relation to mathematics. They cut across the natural grain of the subject, and to try to implement them – at least at the level of the upper school – is a continuous exercise in arbitrary choice" (p. 7).

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

Smith et al. (1996) propose a modification of Bloom's taxonomy called the MATH taxonomy (Mathematical Assessment Task Hierarchy) for the structuring of assessment tasks. The categories in the taxonomy are summarised in Table 1.

Table 1: *MATH Taxonomy* (Adapted from Smith et al., 1996)

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

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

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

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

Mathematics assessment components

Based on the literature on assessment models and taxonomies in mathematics (Bloom, 1956; Niss, 1993; Smith et al., 1996), we argued that for purposes of this study it was necessary to adapt the reviewed taxonomies in order to address the issue of assessing the cognitive level of difficulty of mathematical tasks, as well as the cognitive skills associated with each level. With this background, we propose a taxonomy of mathematics, which we call the *assessment component taxonomy*, to identify those components of mathematics that can be successfully assessed using alternative assessment formats such as MCQs. This taxonomy consists of a set of seven items, hereafter referred to as the *mathematics assessment components*. This set of seven mathematics components was 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

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.

Using Bloom's taxonomy (Bloom, 1956), and the MATH taxonomy (Smith et al., 1996), the proposed mathematics assessment components can be classified according to the cognitive level of difficulty of the tasks as shown in Table 2.

Table 2: *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 3 summarises the proposed mathematics assessment components and the corresponding cognitive skills required within each component. Based on the literature on assessment, the necessary cognitive skills required by students to complete the mathematical tasks within each mathematics assessment component were identified.

Table 3: *Mathematics assessment component taxonomy and cognitive skills*

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

Question examples in assessment components

In the following discussion, question examples within two of the mathematics assessment components have been identified according to Table 3. These items, one MCQ and one CRQ, were selected from the tests and examinations in the first year Mathematics Major course (MATH109) at the University of the Witwatersrand, Johannesburg. 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.

Example 1: *Conceptual mathematics assessment component*

Assessment Component 3: Conceptual (MCQ)

State why the Mean Value Theorem does not apply to the function defined by $f(x) = \frac{2}{(x+1)^2}$ 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 *conceptual* question (Example 1), 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 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.

Example 2: *Problem solving assessment component*

Assessment Component 6: Problem solving (CRQ)

This question deals with the statement

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

(1.1) Show that the statement is true for $n = 2$.
(1.2) Use Pascal's triangle to expand and then simplify $(k+3)^3$.
(1.3) Hence, assuming that $P(k)$ is true for $k > 2$ with $k \in \mathrm{N}$, prove that $P(k+1)$ is true.
(1.4) Based on the above results, justify what you can conclude about the statement $P(n)$.

MATH109 June 2006, Section B: Algebra. Question 1

In the *problem solving* CRQ (Example 2), 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 sub-questions involving different cognitive skills to assist the student with the method of solving using mathematical induction. In sub-question (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) assesses within the technical mathematics assessment component. Sub-question (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 sub-question (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 sub-question (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. Sub-question (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.

The quality index (QI) model

The quality index (QI) model, developed by Huntley et al. (2008, 2009), can be used both to quantify and visualize the quality and nature of a mathematics question. Three parameters, namely discrimination index, confidence index and expert opinion were used to develop the QI model to quantify the quality of a question. In addition, a fourth parameter, namely the level of difficulty, was used to qualitatively contribute to the characteristics of a question.

In order to graphically represent the qualities and characteristics of a question, 3-axes radar charts are constructed, where each of the three measuring criteria are represented as one of the three arms of the radar plot.

In this model, no distinction is made between the relative importance of the three criteria in their contribution to the overall quality of a question. The quality index (QI) is defined to be the *area* of the triangle in the radar chart. For the QI model, the smaller the QI value of the radar plot i.e. the closer the QI value is to zero, the better the quality of the question. To visualize the difficulty level of the test item, a corresponding shading of the radar plot was chosen to represent the six difficulty levels: very easy; easy; moderately easy; moderately difficult; difficult; very difficult. The shading for the easy items is a lighter shade of grey and for the more difficult items, a darker shade of grey is used, with black representing very difficult items.

Briefly, 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:

- the shape of the radar plot;
- the area of the radar plot;
- the shading of the radar plot.

Further discussion of the application of the QI model and radar charts will be presented in the component analysis of this study.

Research questions

The quality index model (QI), briefly outlined in this study, was used in order to address the following research question:

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

In order to answer the research question, the following sub-question was formulated:

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?

In order to address the research sub-question, the QI model was used to identify those components that can be successfully assessed using the provided response format, in particular MCQs, and which can be better assessed in the constructed response question format. To assist with this process, we used the proposed mathematics assessment component taxonomy.

Research design

Response data from 14 different mathematics tests, both CRQ and PRQ format, written between August 2004 and June 2006 were collected. The study was set in the context of a first-year level mathematics major course at the University of the Witwatersrand, Johannesburg. In total, 207 test items were analysed in this study.

The Rasch model, as a statistical tool, was used in the quantitative data analysis of this study. It is a probabilistic model by which linear measures are created to be used in subsequent parametric tests (Rasch, 1980).

In comparison to traditional analysis techniques, the Rasch model can be used (1) to analyse and improve a test instrument, and (2) to generate linear (interval strength) learner scores, thus meeting the assumptions of parametric statistical tests such a t-tests and ANOVA (Birnbaum, 1968).

The Rasch model focuses on the interaction of a person with an item rather than upon the total test score. One of the basic assumptions of the Rasch model is that a relatively stable latent trait underlies test results (Boone & Rogan, 2005).

For this reason, the model is also sometimes called the latent trait model. A feature of traditional test theory is that many of the statistics depends on the assumption that the true scores of people are normally distributed (Andrich, 1988). An important advantage of the Rasch latent trait model is that no assumptions need to be made about this distribution, and indeed, the distribution of abilities may be studied empirically. It was for this reason that the Rasch model was chosen above other traditional statistical procedures for the quantitative research methodology of this study.

Component analysis

Using the QI model, a component analysis of selected questions, both MCQs and CRQs, in the seven different mathematics assessment components was conducted.

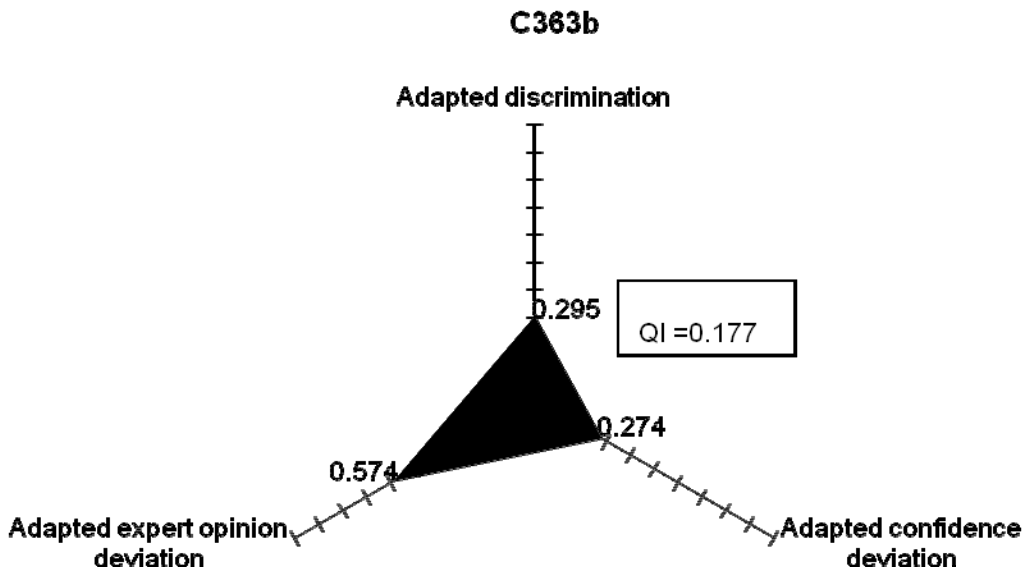
Two such examples of the component analysis will now be illustrated. Example 1 is a CRQ item within the disciplinary component and Example 2 is a MCQ item within the conceptual component. For each item, the question is followed by a radar plot and a table summarising the quality parameters of the test item, i.e. item difficulty; discrimination; confidence index; expert opinion and the final quality index. Each of the axes of the radar plots are labelled with the corresponding values for discrimination, confidence index and expert opinion. The Quality Index (QI) is displayed alongside the radar plot. The shading of the radar plot corresponds to one of the six item difficulty levels as outlined in the brief discussion of the QI model in this study. The black shading of the radar plot in Example 1 indicates a very difficult item while the lighter grey shading of the radar plot in Example 2 indicates an item of moderate difficulty. The comments briefly summarise the difficulty level, the three measuring criteria and the overall quality of the item. All these parameters were used to define the QI model developed earlier by Huntley (2008).

Example 1: Disciplinary component

C363b

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

CRQ, Calculus, March 2006, Q3b



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

Example 2: Consolidation component

A45MA4

If f is an odd function and g is an even function then

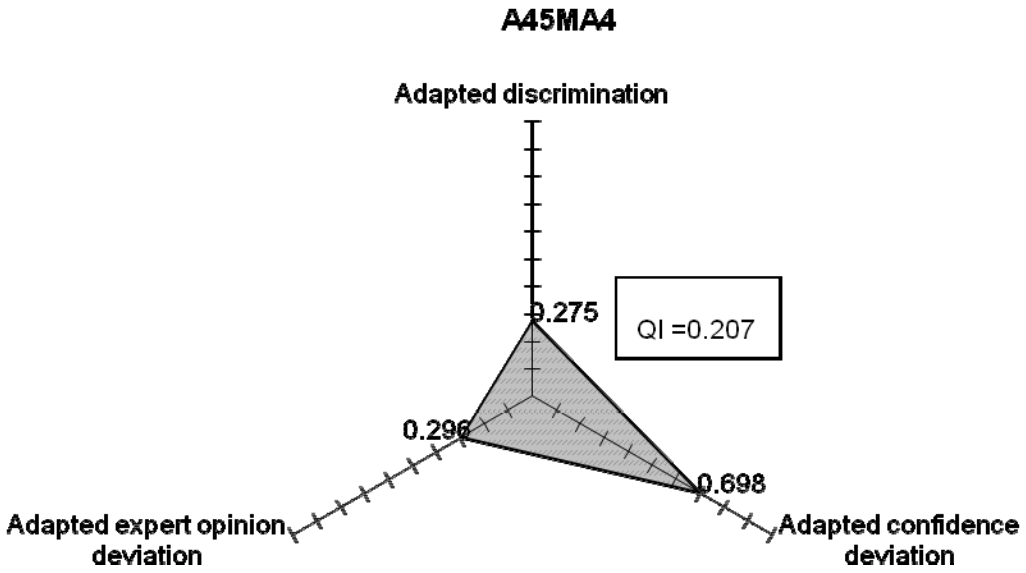
A. $f \circ g$ is an even function

B. $f \circ g$ is an odd function

C. f is a one-to-one function

D. g is a one-to-one function

PRQ, Algebra, March 2005, Tut Test A, Q4



A45MA4		Comment
Assessment Component	Consolidation	
PRQ/CRQ	PRQ	
Item Difficulty	1.11	Moderately difficult
Discrimination	0.275	Discriminates well
Confidence Index	0.698	Large deviation from expected confidence level
Expert Opinion	0.296	Small deviation from expected performance
Quality Index	0.207	Good quality PRQ

Research results and discussion

In the presentation of the results, a comparison of test items, both PRQs and CRQs, within each assessment component will be made.

Table 4 summarises the quality of the item, both PRQs and CRQs, within each assessment component. Within each component the number of good and poor quality items are given, both for the PRQ and CRQ formats. The numbers are also given as percentages of the total number of items.

Technical

In the technical assessment component, there is a higher percentage (73%) of good PRQs than good CRQs (41%). 73% good PRQs compared to 41% good CRQs shows us that PRQs are more successful than CRQs as an assessment format in the technical component. There is also a much higher percentage (73%) of good PRQs than poor PRQs (27%). CRQs, however, are not that successful in this component, with the results showing 59% poor CRQs compared to 41% good CRQs. The conclusion is that the technical assessment component lends itself better to PRQs than to CRQs.

Disciplinary

In this study, the disciplinary component is the assessment component with the most items (58), of which 34 were CRQs and 24 were PRQs. In this component it is interesting to note that the percentages of good PRQs (50%) and good CRQs (47%) are almost equal. In addition, there is no difference between the good PRQs (50%) and the poor PRQs (50%), with very little difference between the good CRQs (47%) and poor CRQs (53%). PRQs and CRQs can be considered as equally successful assessment formats in the disciplinary component.

Table 4: *Component analysis – trends*

COMPONENT	No. of PRQs	No. of CRQs	Total no. of items	Good quality items	Poor quality items	Good PRQs	Good CRQs	Poor PRQs	Poor CRQs
Technical	11	22	33	17 (52%)	16 (48%)	8 (73%)	9 (41%)	3 (27%)	13 (59%)
Disciplinary	24	34	58	28 (48%)	30 (52%)	12 (50%)	16 (47%)	12 (50%)	18 (53%)
Conceptual	26	30	56	28 (50%)	28 (50%)	14 (54%)	14 (47%)	12 (46%)	16 (53%)
Logical	7	6	13	5 (39%)	8 (61%)	1 (14%)	4 (67%)	6 (86%)	2 (33%)
Modelling	3	10	13	8 (62%)	5 (38%)	2 (67%)	6 (60%)	1 (33%)	4 (40%)
Problem solving	7	4	11	6 (55%)	5 (45%)	4 (57%)	2 (50%)	3 (43%)	2 (50%)
Consolidation	16	7	23	12 (52%)	11 (48%)	7 (44%)	5 (71%)	9 (56%)	2 (29%)

Conceptual

The conceptual component also contained many items (56), with an almost equal number of PRQs and CRQs (26 PRQs versus 30 CRQs). 50% of the items are of good quality and 50% are of poor quality. In this component, there is no clear trend that PRQs are better than CRQs or vice versa. There is a slight leaning towards good PRQ assessment (47% good CRQs compared to 54% good PRQs). Therefore, in the conceptual assessment component, PRQs could be used as successfully as CRQs as a format of assessment.

Logical

In this study, it is interesting to note that the majority of questions within the logical component were of a poor quality mainly due to the large percentage of poor PRQs. There are noticeably more good quality CRQs (67%) than good quality PRQs (14%), and noticeably more poor quality PRQs (86%) than poor quality CRQs (33%). A very high percentage of the PRQs (86%) in the logical component were of a poor quality. The conclusion is that the logical assessment component lends itself better to CRQs than to PRQs.

Modelling

In the modelling component, very few PRQs were used as assessment items in comparison to CRQs, 3 PRQs versus 10 CRQs, probably because it is difficult to set PRQs in this component. Despite the small number of PRQs, it was encouraging to note that the good PRQs (67%) far outweighed the poor PRQs (33%). So in terms of quality, the PRQs were highly successful in the modelling component. There are also more good CRQs (60%) than poor CRQs (40%). It appears that although more difficult to set in the modelling component, PRQs could be used as successfully in the modelling assessment component as CRQs.

Problem solving

Although the problem solving component had the least number of items (11), it is interesting to note that there are more PRQs (7) than CRQs (4). There is a slightly higher percentage (57%) of good PRQs than good CRQs (50%). Although the sample is too small to make definite conclusions, there is no reason to disregard the use of PRQs in this assessment component. In fact, PRQs seem to be slightly more

successful than CRQs, and the conclusion is that PRQ assessment format can add value to the assessment of the problem solving component.

Consolidation

It was somewhat surprising to note that corresponding to the highest level of conceptual difficulty, the consolidation component displayed an unusually higher proportion of PRQs (16) to CRQs (7). This supports the earlier claim that PRQs are not only appropriate for testing lower level cognitive skills (Haladyna, 1999; Thorndike, 1997; Williams, 2006). In the consolidation component there is a significant higher percentage (71%) of good CRQs than good PRQs (44%). In addition, there is a higher percentage of poor PRQs (56%) than good PRQs (44%). The high percentage of good CRQs (71%) in comparison to poor CRQs (29%) indicates that the consolidation assessment component lends itself better to CRQs than to PRQs.

Conclusion

The mathematics assessment component taxonomy, proposed by the authors in this study, 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 study have shown that it is not necessary to restrict MCQs 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 MCQs 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, MCQs 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 MCQs can be set within most of the assessment components in the taxonomy which do promote a deeper approach to learning.

Results of this study (Huntley, 2008) 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 results of this study, MCQs 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, MCQs 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 5 presents a comparison of the success of MCQs and CRQs in the mathematics assessment components.

As Table 5 illustrates, the enlightening conclusion is that there are only two components where CRQs outperform MCQs, namely the logical and consolidation assessment components. In two other components, MCQs are observed to slightly outperform CRQs, namely the conceptual and problem solving assessment components. The MCQs outperform the CRQs substantially in the technical and modelling assessment components. In one component there is no observable difference, the disciplinary assessment component.

In addressing the research question formulated as “Can we successfully use MCQs as an assessment format in undergraduate mathematics?”, this paper has addressed the sub-question formulated as “Which of the mathematics assessment components can be successfully assessed using the PRQ assessment

Table 5: *Comparison of the success of MCQs and CRQs in the mathematics assessment components*

Mathematics assessment Component	Comparison of success
Technical	MCQs can be used successfully
Disciplinary	No difference
Conceptual	MCQs can be used successfully
Logical	CRQs more successful
Modelling	MCQs can be used successfully
Problem solving	MCQs can be used successfully
Consolidation	CRQs more successful

format and which of the mathematics assessment components can be successfully assessed using the CRQ assessment format?". To address the sub-question, the mathematics assessment component taxonomy was proposed.

In terms of the mathematics assessment components, it was noted that certain assessment components lend themselves better to MCQs 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 MCQs in the logical and consolidation assessment components, MCQs can add value to the assessment of these higher cognitive component levels. Greater care is needed when setting MCQs 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 that MCQs can improve students' acquisition of mathematical insight and knowledge. The PRQ assessment format (including MCQs) can be used as successfully as the CRQ assessment format to assess undergraduate mathematics.

References

Andrich, D. (1988). *Rasch models for measurements*. Beverley Hills, CA: Sage Publications.

Begle, E. G., & Wilson, J. W. (1970). Evaluation of mathematics programs. In E. G. Begle (Ed.), *Mathematics Education* (69th Yearbook of the National Society for the Study of Education, Part I, 376-404). Chicago: University of Chicago Press.

Ball, G., Stephenson, B., Smith, G. H., Wood, L. N., Coupland, M., & Crawford, K. (1998). Creating a diversity of experiences for tertiary students. *International Journal of Mathematical Education in Science and Technology*, *29*(6*)*, 827-841.

Biggs, J. (1991). Student learning in the context of school. In J. Biggs (Ed.), *Teaching for learning: The view from cognitive psychology* (pp. 7-20). Hawthorn, Victoria: Australian Council for Educational Research.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: The cognitive domain*. New York: David McKay.

Boone, W., & Rogan, J. (2005). Rigour in quantitative analysis: The promise of Rasch analysis techniques. *African Journal of Research in Mathematics, Science and Technology Education, 9*(1), 25-38.

Eisenberg, T. A. (1975). Behaviorism: The bane of school mathematics. *International Journal of Mathematical Education in Science and Technology, 6*(2)*,* 163-171.

Engelbrecht, J., & Harding, A. (2003). Online assessment in mathematics: Multiple assessment formats. *New Zealand Journal of Mathematics, 32* (Supp.), 57-66.

Freeman, R., & Lewis, R. (1998). *Planning and implementing assessment.* London: Kogan Page.

Geyser, H. (2004). Learning from assessment. In S. Gravett & H. Geyser. (Eds.), *Teaching and learning in higher education* (pp. 90-110). Pretoria: Van Schaik.

Haladyna, T. M. (1999). *Developing and validating multiple choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Huntley, B. (2008). *Comparing different assessment formats in undergraduate mathematics*. (Unpublished doctoral dissertation). University of Pretoria, Pretoria. Retrieved from http://upetd.up.ac.za/thesis/available/etd-01202009-163129/

Huntley, B., Engelbrecht, J., & Harding, A. (2008). How good are your mathematics questions? In O. Nam Kwon & A. Harding (Eds.), *Proceedings of the Eleventh International Congress on Mathematical Education: Topic Study Group 5*. Retrieved from http://tsg.icme11.org/document/get/554.

Huntley, B., Engelbrecht, J., & Harding, A. (2009). An assessment component taxonomy for alternative mathematics assessment formats. In D. Wessels (Ed.), *Proceedings of the 7th Southern Right Delta Conference on the Teaching and Learning of Undergraduate Mathematics and Statistics* (pp. 117–128). Gordons Bay, South Africa: International Delta Steering Committee.

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.

Niss, M. (1993). *Investigations into assessment in mathematics education. An ICMI study*. Dordrecht: Kluwer Academic Publishers.

Ormell, C. P. (1974). Bloom's taxonomy and the objectives of education. *Educational Research, 17,* 3-18.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.).Chicago: University of Chicago Press.

Romberg, T. A., Zarinnia, E. A., & Collis, K. F. (1990). A new world view of assessment in mathematics. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 21-38). Washington, DC: American Association for the Advancement of Science.

Romberg, T. A. (1992). *Mathematics assessment and evaluation. Imperatives for mathematics educators*. Albany: State University of New York Press.

Smith, G. H., Wood, L. N., Crawford, K., Coupland, M., Ball, G., & Stephenson, B. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, *27*(1), 65-77.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Webb, N., & Romberg, T.A. (1992) Implications of the NCTM standards for mathematics assessment. In T.A. Romberg (Ed.), *Mathematics Assessment and Evaluation: Imperatives for Mathematics Educators* (pp. 37-60). Albany: State University of New York Press.

Williams, J. B. (2006). Assertion-reason multiple-choice testing as a tool for deep learning: A qualitative analysis. *Assessment and Evaluation in Higher Education, 31*(3), 287-301. doi:10.1080/02602930500352857

World Book Dictionary (1990). Chicago: World Book Inc.