Perspectives in Education 2014: 32(1)
ISSN 0258-2236

http://www.perspectives-in-education.com

A Rasch analysis to determine the difficulty of the National Senior Certificate Mathematics examination

Joyce Sewry

Paul Mokilane

The National Senior Certificate (NSC) examinations were written for the second time in 2009 amid much criticism. In this study, scripts of candidates who wrote the NSC Mathematics examinations (papers 1 and 2) in 2009 were used as data to analyse the marks scored and then polytomous Rasch analysis was conducted for all the sub-questions to determine the level of difficulty of the questions. The purpose of applying Rasch measurement models is to explore the extent to which a test or an examination and its associated data set permit the interpretation of an underlying linear scale of ability against which to interpret overall performance and item difficulty. In the NSC data, some questions discriminated well at the lower-ability levels of candidates, but no questions were found to discriminate among higher-ability candidates.

Keywords: Mathematics education, Grade 12 examinations, National Senior Certificate, Rasch analysis

Introduction

Large-scale studies, including examinations, tests and questionnaires have been used for data collection for research and, in the case of examinations, teachers use the results of the analysis to guide their teaching (Edwards & Alcock, 2010). A number of studies have also been undertaken to determine levels of mathematical ability at different stages of schooling (Wendt, Bos & Goy, 2011; Wilson & Macgillivray, 2007).

The National Senior Certificate (NSC) was written for the second time in 2009, followed by much criticism when the results were released (Association for Mathematics Education of South Africa [AMESA], 2009; Keeton, 2010). AMESA, which reported on the 2009 and 2010 Mathematics examinations, stated that the

Joyce Sewry
Department of Chemistry, Rhodes University
E-mail: j.sewry@ru.ac.za
Telephone: 046 603 8259

Paul Mokilane
Statistical Information and Research, Umalusi
E-mail: paul@umalusi.org.za
Telephone: 012 349 1510

2009 paper 1 was at too high a level, while the standard of the 2010 paper was fairer, despite the fact that there were not many questions at the lower level. Furthermore, Mathematics paper 2 of 2009 was a fair paper, and that of 2010 was at an appropriate level (AMESA, 2009; 2010).

The NSC examinations are high-stakes examinations in the South African schooling system, because they are school-leaving examinations. Also, they are used to select candidates for higher education programmes, hence, the maintenance of high standards of these examinations. It is, therefore, important that examination papers be analysed, paying particular attention to the quality of the questions (Grussendorf, Booyse & Burroughs, 2010).

Mathematics, Physical Sciences and Accounting are seen as 'gateway' subjects that facilitate entry into tertiary education for school leavers. Passing these subjects is critical because university study has the potential to address the lack of skills in South Africa (Grussendorf *et al*., 2010). Based on this issue, more emphasis has been placed on the analysis of the examinations of different learning areas such as Mathematics (Umalusi, 2009). Results in the NSC examinations for Mathematics, in particular, have been poor for a number of years. An illustration of this is given in table 1 which shows, for instance, that in 2009, 29% of the candidates obtained a mark of 40% or more nationally, while 31% achieved this in 2010. These results point to the need for perusing candidates' responses in the examinations. In this regard, the Rasch model was used to analyse the 2009 Mathematics scripts.

Table 1*: Percentage of learners who achieved greater than 30% and 40% in 2009 and 2010*

	2009			*2010*		
	Number of candidates writing subject	*% achieved >30%*	*% achieved >40%*	*Number of candidates writing subject*	*% achieved >30%*	*% achieved >40%*
Mathematics						
National [1, 2]	290 407	46	29	263 034	47	31
Eastern Cape [1]	43 251	38	21	38 801	37	21
Grahamstown	530	38	25	378	43	28

[1]Department of Basic Education, 2011

[2]Keeton, 2010

The Rasch model

Item response theory (IRT) is based on two postulates: the performance of an examinee on an item (test question) is related to the examinee's ability or latent trait; and the relationship between the examinee's performance and the difficulty of an item can be related by an item characteristic curve (ICC).

Examinees' abilities are scaled such that an 'average' person has a latent trait of zero and an 'average' examinee will have a 50% probability of answering correctly a question of 'average' difficulty. Also, IRT has the property of invariance, in other words, the characteristics of difficulty of an item are not dependent on the ability distribution of the examinees, and the ability of an examinee is not dependent on the item characteristics (Baker, 2001; Hambleton, Swaminathan & Rogers, 1991). So, if a question is asked in a different test with a different set of examinees, it should have a similar level of difficulty. The item parameters (item difficulty and discrimination index) are independent of the test takers' characteristics, and the test takers' parameter (ability level) is independent of the item characteristics.

Rasch analysis is a specific application of IRT. In Rasch analysis a distinction is made between dichotomous and polytomous analysis. The dichotomous model is used in simple questions where an answer is either right or wrong, such as multiple-choice questions. On the other hand, the polytomous Rasch model is used when a variety of marks can be awarded for a question (Wu & Adams, 2007), as was the case with questions in the NSC Mathematics papers. Rasch analysis is based on the model that, mathematically, the probability of a candidate of a certain ability to answer correctly a question of a specific difficulty can be represented as (Yu, 2010):

$$Probability = 1/(1+exp(-(ability - difficulty)))$$

The model assumes that the more proficient candidates (candidates with high latent traits) in the subject are more likely to get the difficult questions correct, while the less proficient candidates are expected to answer only the easy questions correctly. When plotting the range of difficulties over the range of candidates' latent traits, an ICC is formed. From the shape of the ICC, decisions about the item can be made, for example, whether a question is easy or difficult. Also, it may be explored whether a question discriminates among the high-performing and low-performing candidates, and it may reveal whether a question is at all confusing. The latter is evident when, for instance, high-ability candidates answered a question wrongly, but low-ability candidates answered the question correctly (Van der Berg & Taylor, 2010).

The polytomous Rasch model, also known as the partial credit model (PCM), is used for question items with a scale of answers (Wu & Adams, 2007). For instance, if a question has a total mark of 2, then candidates could be awarded 0, 1 or 2 marks. In such a situation the ICC will consist of three curves. Each curve will represent a probability of attaining no mark (0), 1 or 2 marks respectively. For questions scoring higher marks, the ICC will produce a corresponding number of curves. Since an ICC with many curves could become 'messy', an 'expected score' curve is used instead. The expected score, defined by E, is based on the probabilities of achieving each of the marks (Wu & Adams, 2007). Mathematically, the expected score is represented by:

$$E = 0 \times Pr(X = 0) + 1 \times Pr(X = 1) + 2 \times Pr(X = 2)$$

Thus, a plot of expected scores versus latent traits is drawn. The expected score curve gives an indication of the difficulty of each part of the question. The results of Rasch analysis give the level of difficulty of a test item as well as expected score curves. In essence, a score curve is a plot of the expected scores for the question against the abilities (latent traits) of the candidates. Even though the difficulty levels are invariant in Rasch analysis, it does not mean that the numerical values of the difficulty of each item would be the same in different tests.

> *The obtained numerical values will be subject to variation due to sample size, how well-structured the data is, and the goodness-of-fit of the curve to the data. Even though the underlying item parameter values are the same for two samples, the obtained item parameter estimates will vary from sample to sample. Nevertheless, the obtained values should be 'in the same ballpark' (Baker, 2001: 62).*

Baker goes on to say that ICCs for two different groups should be similar, since the number of candidates with a certain ability is not the issue, but rather the fact that there are different abilities present.

Analysis of fit of different items is also conducted during Rasch analysis. Two statistics are used to look at the item fit, namely the weighted fit mean square and the unweighted fit. The weighted fit mean square (MNSQ) (infit) is used to indicate that the standardised residuals are weighted by the variance of the item response (Wu & Adams, 2007); the weighting gives more emphasis to the anomalous residuals of the examinees whose ability levels are near to the item difficulties, with much less weight to the residuals when examinees have abilities far from the item difficulty (Yu, 2010).

The unweighted fit MNSQ (outfit) is the mean of the squared standardised residuals which are considered to have a common weight of 1 (Wu & Adams, 2007). The outfit will show up presence of cases with unexpected responses where a candidate's response is not in line with the difficulty of a 'very easy' or 'very hard' item (Prieto, Alonso & Lamarca, 2003).

The ideal values of both the infit and outfit statistics are equal to 1 (Wilson & Macgillivray, 2007; Wu & Adams, 2007; Yu, 2010) when a measurement-like model is valid for the data set, but 'acceptable' values for both statistics are between 0.7 and 1.3 (Hwang & Davies, 2009; Prieto *et al*., 2003; Smith, Rush, Fallowfield, Velikova & Sharpe, 2008; Wilson & Macgillivray, 2007). These limits may be used regardless of the size of the data set (Smith *et al*., 2008).

Data collection

To collect data, NSC Mathematics examination papers 1 and 2 written in 2009 were analysed. The analysis focused on the candidates' scripts from the Grahamstown education district.

Initially, marks of each sub-question in each paper were entered into an Excel spreadsheet. These marks were for all the candidates in a school. Packs of scripts (per school) were selected randomly, but this strategy was soon changed, since many of the schools had no candidates who scored more than 30% for any paper. In fact, of the 284 candidates sampled, 119 scored less than 15/150 for Mathematics paper 1. Therefore, a decision was made to also include schools whose learners had better marks compared with the initial sample. Thus, scripts from former model C schools were included. This inclusion is consistent with Baker's contention that the number of candidates with a specified ability is not important, but rather the fact that there are different abilities present (Baker, 2001).

The Rasch model was fitted to Mathematics papers 1 and 2 data using the Conquest (Generalised Item Response Modelling) software. From Rasch analysis and the corresponding ICC, it can also be determined which questions discriminated well at different levels of ability (Hambleton *et al*., 1991). For each question of the examination papers, an estimate of difficulty was obtained – the higher the estimate of difficulty, the more difficult the question. Expected score curves for each sub-question were also drawn.

Rasch analysis: Mathematics paper 1

Table 2 represents the Rasch analysis results including the maximum marks and the content on which each question was based for the 2009 Mathematics paper 1 scripts (n = 290) in the Grahamstown District. Since the PCM was used to estimate the difficulty levels of items, the estimates could not be calculated in the case where no learner scored some of the marks between the minimum and maximum possible marks in an item. For example, if an item was scored out of 2 marks and no learner in the sample scored 1 mark, which is between 0 (minimum) and 2 (maximum), the difficulty level of the item could not be estimated using the Conquest software. In this study there were six instances where this was the case – these are depicted by xxx (see table 2).

Table 2: Rash analysis results, maximum marks and content on which questions were based

Question	Max mark	Rasch difficulty	Content	Question	Max mark	Rasch difficulty	Content
1.1.1	3	–2.475	A and E*	**8.2**	1	1.959	F and G†
1.1.2	4	–1.688	A and E	**8.3**	1	1.656	F and G
1.1.3	4	–0.925	A and E	**8.4.1**	1	1.111	F and G
1.2	5	–1.301	A and E	**8.4.2**	3	0.680	F and G
1.3	3	xxx	A and E	**8.5**	3	1.190	F and G
1.4	3	0.412	A and E	**9.1**	4	–1.137	A and F§

Question	Max mark	Rasch difficulty	Content	Question	Max mark	Rasch difficulty	Content
2.1.1	3	–1.485	P and S**	9.2.1	3	–0.492	A and F
2.1.2	2	–0.796	P and S	9.2.2(a)	3	0.127	A and F
2.2	5	0.784	P and S	9.2.2 (b)	1	xxx	A and F
3.1	1	–1.377	P and S	9.2.3	4	–0.072	A and F
3.2	4	–0.530	P and S	9.2.4	1	–0.189	A and F
4.1	3	–0.494	P and S	10.1	5	–1.270	Calculus
4.2	2	0.053	P and S	10.2	2	–0.647	Calculus
4.3	4	–0.736	P and S	11.1	5	–0.723	Calculus
4.4	2	1.627	P and S	11.2	5	–0.604	Calculus
5.1	2	1.064	P and S	11.3	3	–0.123	Calculus
5.2	3	0.438	P and S	11.4	2	–0.667	Calculus
5.3	3	0.754	P and S	11.5	2	0.777	Calculus
6.1	6	–0.278	F and G†	12.1	2	xxx	Calculus
6.2	2	0.474	F and G	12.2	3	0.403	Calculus
6.3	2	0.419	F and G	12.3	2	1.742	Calculus
6.4	3	2.027	F and G	12.4	3	xxx	Calculus
7.1	1	0.556	F and G	13.1	7	0.080	LP‡
7.2	3	–0.031	F and G	13.2	2	–0.905	LP
7.3	2	–0.126	F and G	13.3	3	0.026	LP
7.4	2	xxx	F and G	13.4	2	xxx	LP
8.1	1	0.988	F and G				

A and E = Algebra and equation; ** P and S = Patterns and sequences; †F and G = Functions and graphs; §A and F = Annuities and finance; ‡LP = Linear programming

The paper started out easy, became more difficult, then questions 9-11 were easy again. In question 6.4 the learners were asked to algebraically show that:

$$g(x)+g(1/x)=g(-x).g(x-1) \quad (x\neq 0 \text{ or } x\neq 1)$$

This was the most difficult question (difficulty = 2.027), while question 1.1.1 where, learners were asked to solve for x:

$$x(x-1)=30$$

was the easiest question (–2.475). The analysis revealed, however, that 58 (20.4%) candidates did not get a mark for this question.

Figure 1 represents the expected score curve for question 1.1.1. It can be seen that the average-ability student (latent trait value = 0.0) could be expected to get full marks, that is, 3 out of 3.

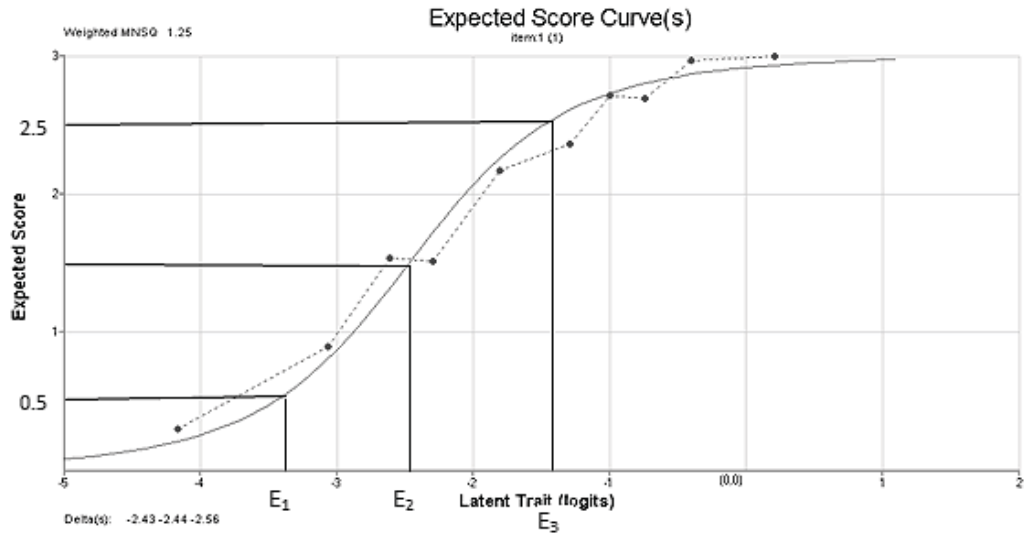


Figure 1: Expected score curve for question 1.1.1

Figure 1 shows that candidates with ability measures below E_1 (expected score, E <0.5) were expected to average near 0 for this question; those with abilities between E_1 and E_2 (expected score, 0.5< E <1.5) were expected to average at 1 mark; those between E_2 and E_3 (expected score, E >1.5) 2 marks, and those with abilities above E_3 should average close to the full 3 marks on this question.

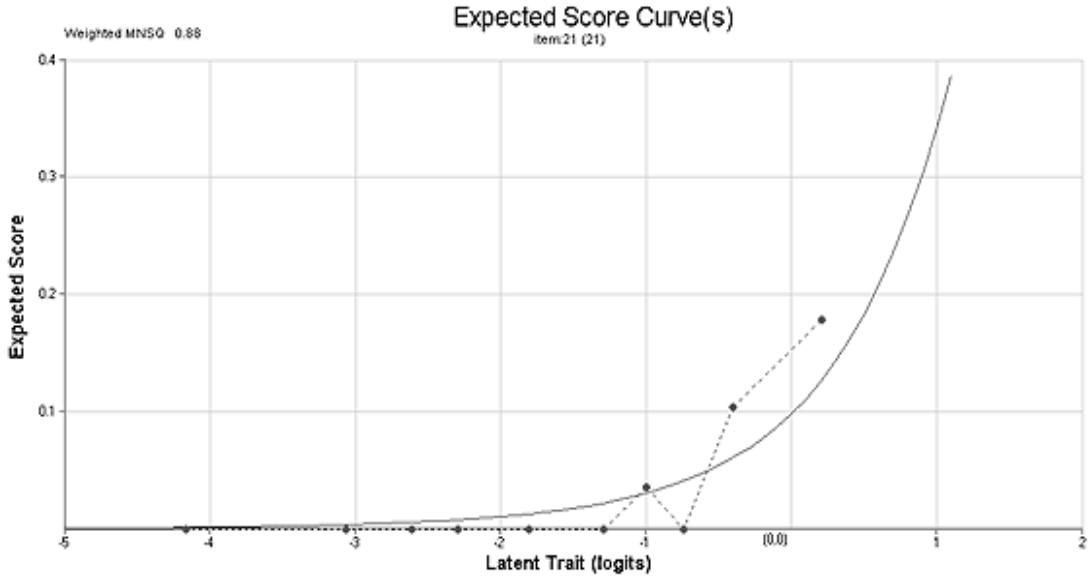


Figure 2: Expected score curve for question 6.4

Figure 2 shows the expected score curve for question 6.4, the most difficult question in the examination. The figure indicates that the average score of students at the

reference value (latent trait value = 0.0) is 0.1 out of 3. The chief marker of the Eastern Cape commented that 'educators tend to neglect the topics that fell outside of the former S.G curriculum', and also that it was evident that students were poor at algebraic manipulation (East Cape Education Department, 2009). AMESA's (2009) comments were: 'It is a good question although Q6.4 demanded a great deal of work and effort for very few marks'.

By studying item response curves, questions which can discriminate among different latent traits can be identified. Thus, from figure 1, question 1.1.1 was a good question to differentiate candidates at the low-ability level, since at an 'ability' between –3 and –2, the curve was at its steepest.

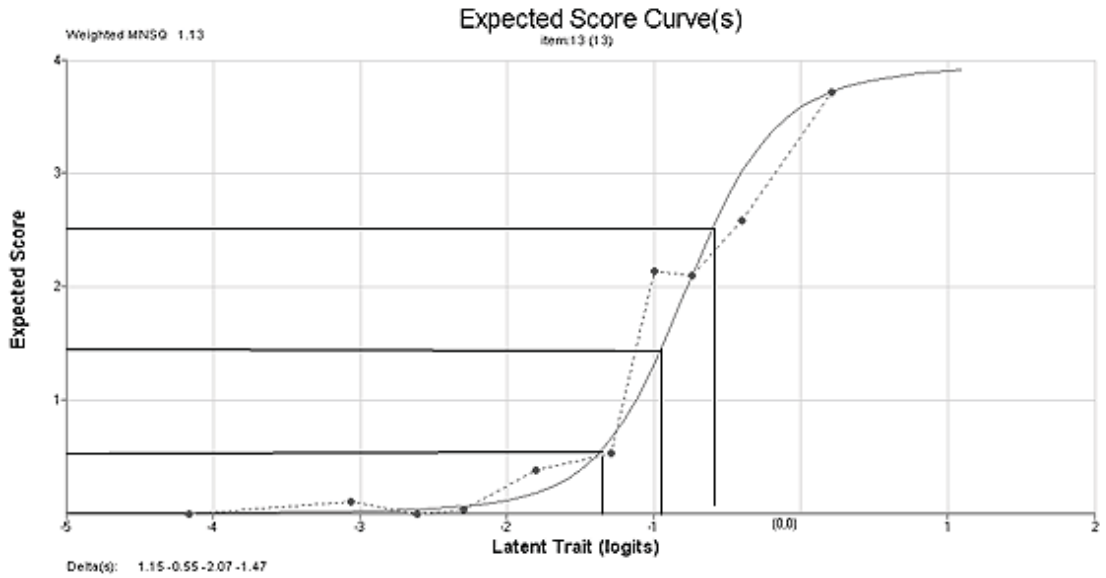


Figure 3: Expected score curve for question 4.3

Figure 3 illustrates that question 4.3 was a good question to differentiate among candidates at latent trait values just below average (0). The average score of candidates at latent trait values between –0.9 and –0.5 should be roughly 2 out of 4.

Unfortunately, according to the expected score curves, there were no questions that differentiated well at the upper end of the latent trait values. In this data set there were very few candidates with high scores in this examination. In the Mathematics paper 1, only four of the 290 candidates scored more than 70% in the examination paper and these marks were: 87%, 81%, 71% and 77% respectively. However, the presence of candidates with high scores should be enough for Rasch analysis to fit expected score curves at all levels (Baker, 2001).

As far as the fit of questions to the Rasch analysis is concerned, the weighted fit MNSQ (infit) of most questions were within or close to the 0.7 to 1.3 range. However, 13 of the unweighted fit MNSQ (outfit) statistics were not within this range. Question 6.4, the most difficult question, had an outfit statistic of 0.37, which is very low. The reason for this could be that only one candidate scored 2, seven candidates scored 1, and the rest scored 0 out of three.

Rasch analysis: Mathematics paper 2

Table 3: Rasch analysis results, maximum marks and content on which questions were based

Question	Max mark	Rasch difficulty	Content	Question	Max mark	Rasch difficulty	Content
1.1	2	0.172	DH*	**6.1.2**	2	–0.599	transformations
1.2	5	–0.939	DH	**6.1.3**	3	–1.045	transformations
1.3	3	xxx	DH	**6.1.4**	2	xxx	transformations
1.4	2	–0.925	DH	**6.2**	6	–0.135	transformations
1.5	3	–0.576	DH	**7.1**	2	–0.187	transformations
2.1	1	1.179	DH	**7.2**	2	0.020	transformations
2.2	2	–1.184	DH	**7.3**	7	xxx	transformations
2.3	1	1.053	DH	**8.1**	3	–0.040	trigonometry
2.4	1	–0.244	DH	**8.2**	2	–0.110	trigonometry
2.5	2	–0.066	DH	**8.3**	3	–0.153	trigonometry
2.6	1	–0.947	DH	**9.1**	7	–0.608	trigonometry
3.1	1	–1.437	DH	**9.2**	7	–0.773	trigonometry
3.2	1	1.366	DH	**9.3**	7	0.427	trigonometry
3.3	3	xxx	DH	**10.1**	3	0.400	trigonometry
4.1	2	–1.492	CG#	**10.2**	4	2.854	trigonometry
4.2	2	0.829	CG	**11.1.1**	6	0.406	trigonometry
4.3	4	–0.308	CG	**11.1.2**	3	1.089	trigonometry
4.4	2	0.546	CG	**11.2**	4	0.446	trigonometry
4.5	3	–0.609	CG	**12.1**	2	0.912	trigonometry
5.1	2	–2.229	CG	**12.2**	2	1.510	trigonometry
5.2	1	–1.171	CG	**12.3**	3	0.738	trigonometry
5.3	5	–0.045	CG	**12.4**	2	1.398	trigonometry
5.4	4	–2.020	CG	**12.5**	3	0.935	trigonometry
5.5	3	–0.760	CG				
5.6	1	–0.966	CG				
5.7	6	xxx	CG				
6.1.1	2	1.064	transfor-mations				

****DH = Data handling; #CG = Coordinate geometry***

Table 3 points to a spread of easy and difficult questions from the beginning of the examination paper up to question 9, but questions 10-12 were all difficult. This level of difficulty was supported by the AMESA (2009) report: 'Some teachers feel that Q10 which involved proving the tan identity, and an application of it, was "unfair". There were some questions (Q5.6-5.7, Q10.2, Q11.1.2 and Q12) that involved critical thinking, but the high achievers should have had no problem answering them' (AMESA, 2009: 12).

Most questions in the Mathematics paper 2 were relatively easy. The most difficult question was question 10.2, which was formulated as follows:

If $\tan(A+B) = \dfrac{\sin(A+B)}{\cos(A+B)}$, prove in any ΔABC that

$\tan A.\tan B.\tan C = \tan A + \tan B + \tan C.$

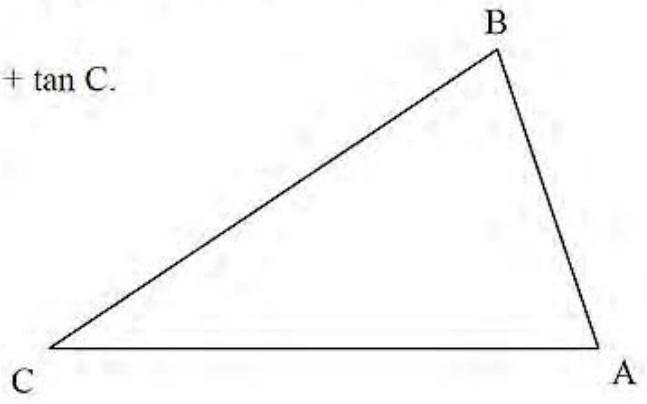


Figure 4: ***Diagram for question 10.2***

The difficulty level of this item is estimated as 2.854. This was a trigonometry identity equation, and figure 5 shows that candidates at the reference zero would be expected to score 2 out of 5.

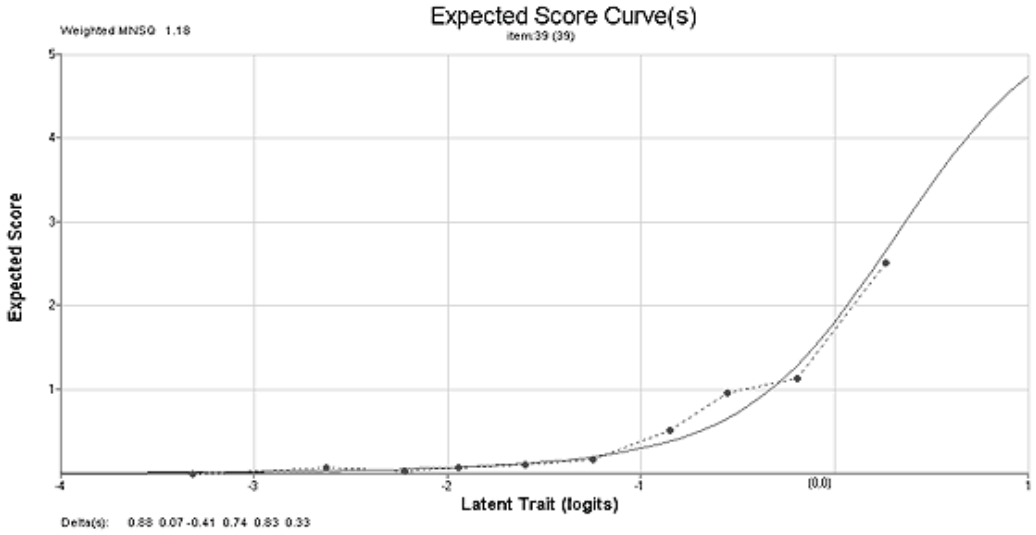
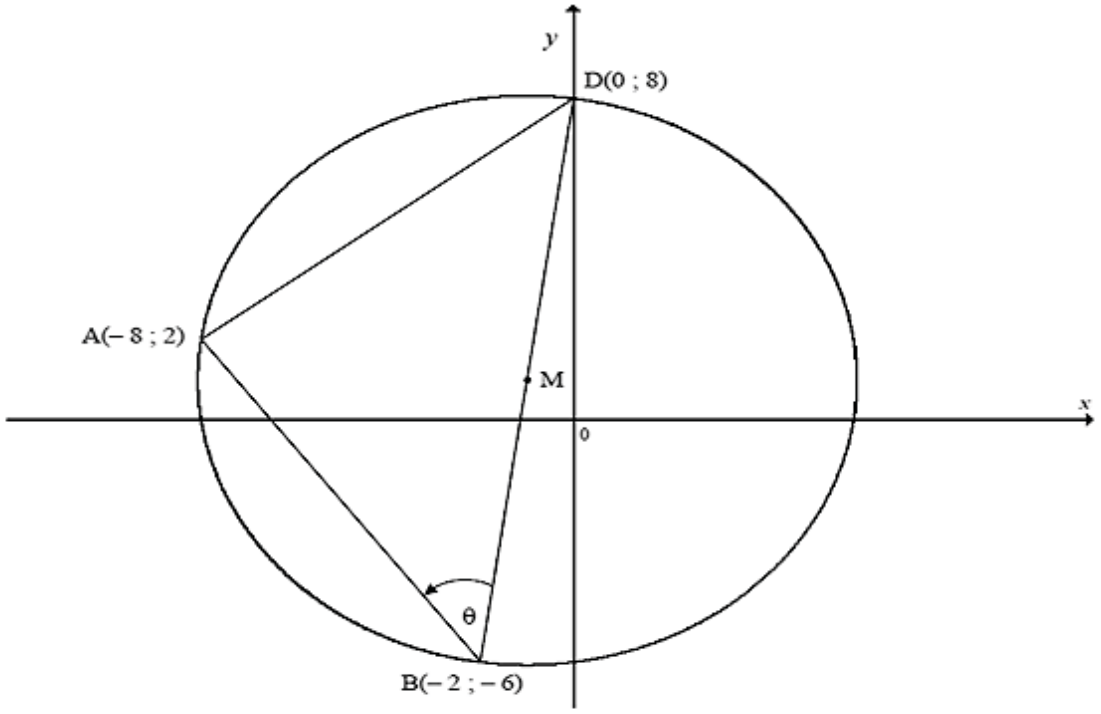


Figure 5: ***Expected score curve for question 10.2***

The easiest question was question 5.1 which asked the learners to calculate the coordinates of a point *M* in the picture. The difficulty level of this item was estimated as –2.229.



5.1 Calculate the coordinates of M.

Figure 6: ***Diagram for question 5.1***

Despite the easy level of question 5.1, 84 of the 295 (28.5%) candidates scored 0 for this question. Figure 7 shows that question 5.1 was a good discriminator for the very low level candidates, since the curve is steepest at latent trait values between –3 and –2, but candidates with latent trait values above –1 would average close to full marks, 2 out of

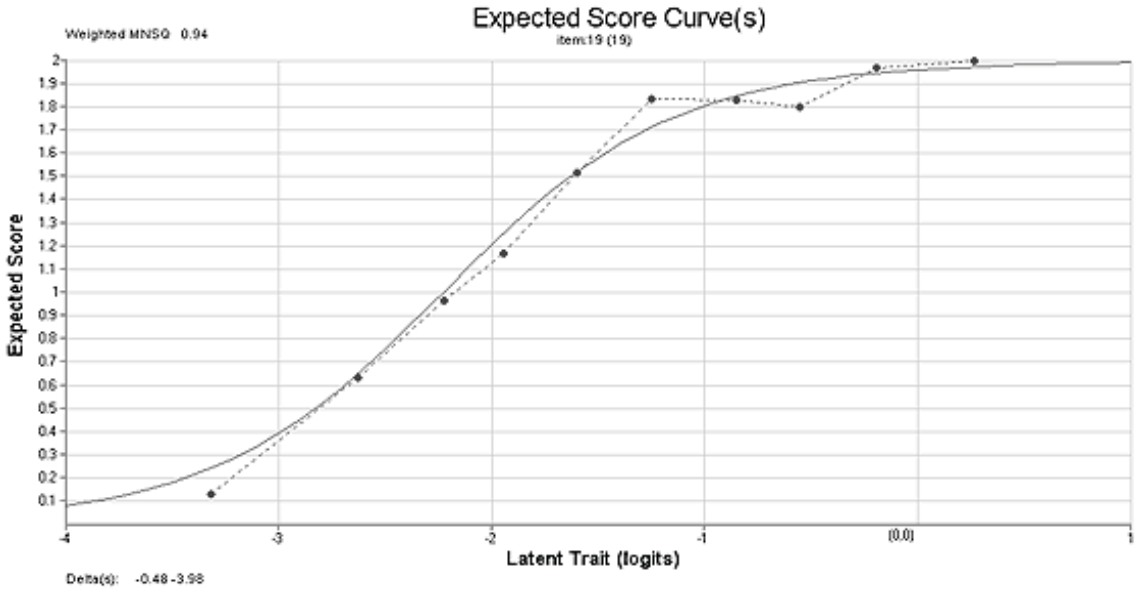


Figure 7: Expected score curve for question 5.1

In question 5.4 candidates were asked to calculate the lengths of AD and AB (figure 6). The estimated difficulty level of question 5.4 was –2.02. The item differentiated well at the very low end of candidates (see figure 8).

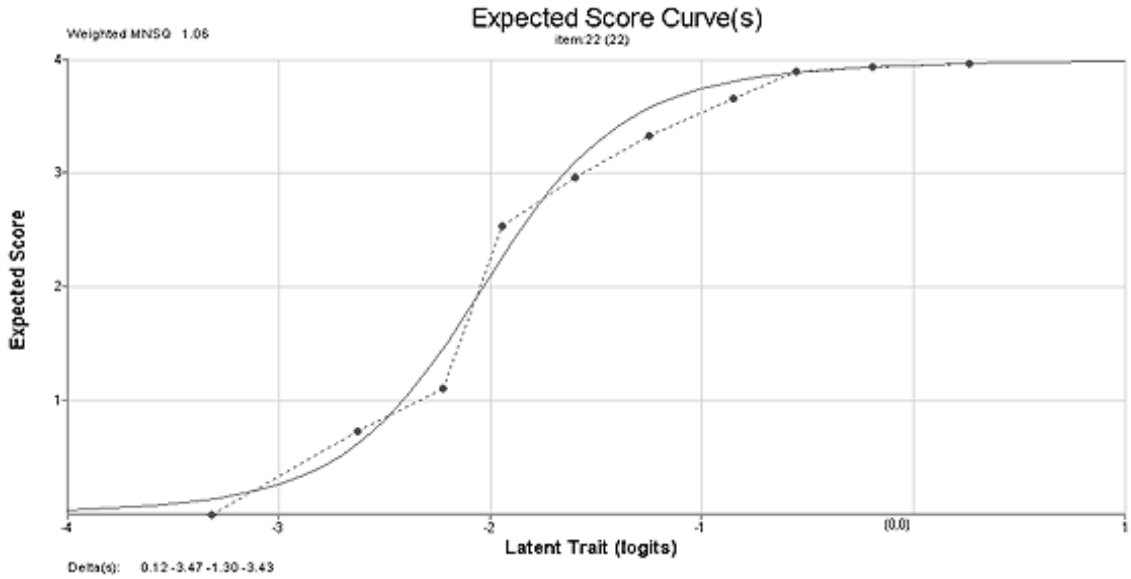


Figure 8: Expected score curve for question 5.4

In question 4.3 learners were asked to calculate *p*, the *x*-coordinate of point C (figure 9).

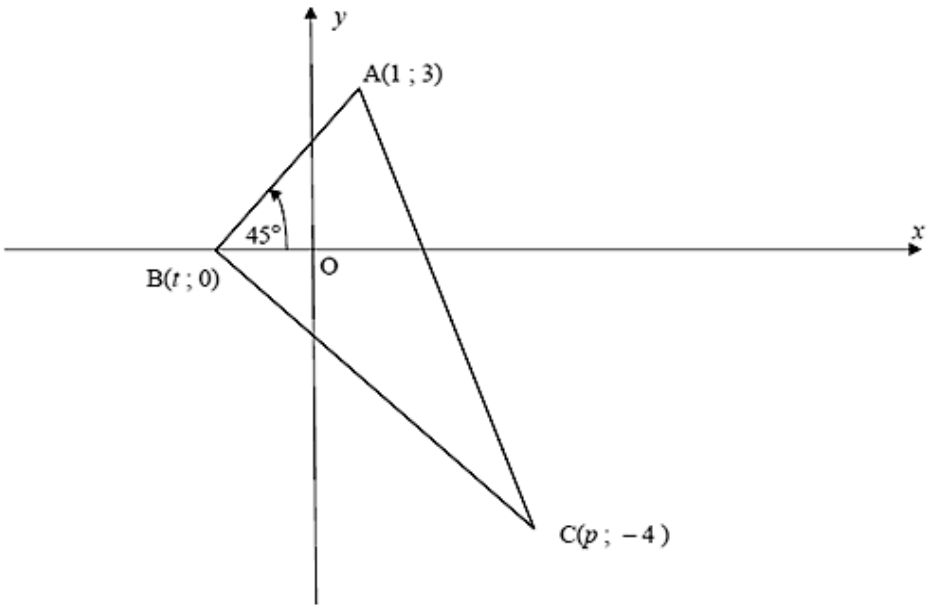


Figure 9: ***Diagram for question 4.3***

The expected score curve is depicted in figure 10. This would be a good question to differentiate among students who were just below the reference 0.

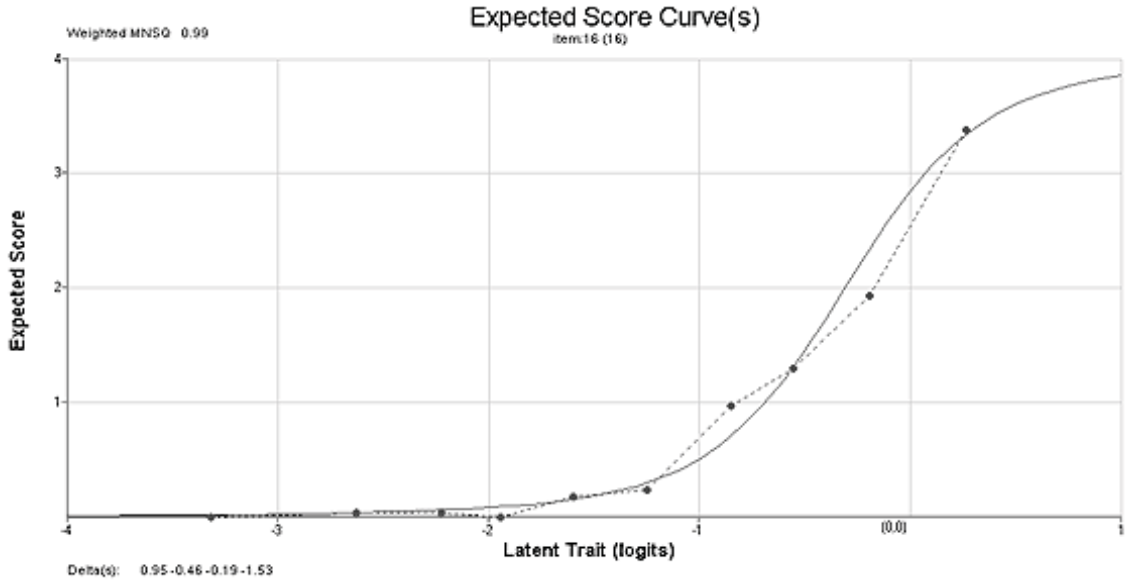


Figure 10: Expected score curve for question 4.3

Question 4.3 had a Rasch difficulty estimate of –0.308. Below reference zero candidates should average 1 to 2 marks, and candidates close to the reference 0 should average about 3 out of 4 for this question on coordinate geometry.

As with Mathematics paper 1, there was, unfortunately, no question for which the expected score curve showed good differentiation for candidates who were average or above average. (In this sample, only four candidates scored more than 70% for Mathematics paper 2, and these marks were 78%, 75%, 86% and 84% respectively.)

As in paper 1, there were few questions for which the infit statistic was not within the range of 0.7 to 1.3. Several questions had their outfit statistics outside the range, though. Both the infit and outfit statistics of question 9.3 ('Determine the general solution of; $\sin x + 2\cos^2 x = 1$') were out of the acceptable range. This outcome arises because most candidates did poorly in this question, but several candidates who scored 5 or 6 out of 7 for the question achieved only between 46 and 55 out of a maximum of 150 for the paper. Thus, these 'weak' candidates did unexpectedly well in question 9.3. Many candidates who did well in questions 9.1 and 9.2, also trigonometric identities, did poorly in question 9.3. The reason for the poor statistics is unclear, since the marking was reliable.

A similar situation arose with question 1.2 ('Write the five-number summary for the data'). One candidate who scored a total of only 20 out of 150, scored 5 out of 5 marks for this question, while another candidate whose total was 113 out of 150, scored only 2 out of 5 marks. There seems to be something wrong with questions 9.3 and 1.2, which needs to be scrutinised by the examiners.

```
                |20 25 26 27 42 85
  2             |
                |
                |91
                |24 58 89
                |
                |18 92
                |
                |
  1             |5 28 29 86
                |15 19 21
                |8 49 84 88 90
                |12 16 17 71
                |87
                |14 32 41 45
              X |23 34 39 62
  0           X |33 43 59 77 78 79 83
             XX |67 74 76 80
           XXXX |22 31 36 54 55 75
        XXXXXXX |10 11 38
        XXXXXXX |40 50 64
         XXXXXX |7 13 37 72
      XXXXXXXXX |3 44 61 69 82
         XXXXXX |47 70 81
 -1      XXXXXX |9 30 48 66 73
           XXXX |4 35 53 56 63
         XXXXXX |6 52
          XXXXX |2 60
       XXXXXXXX |57
```

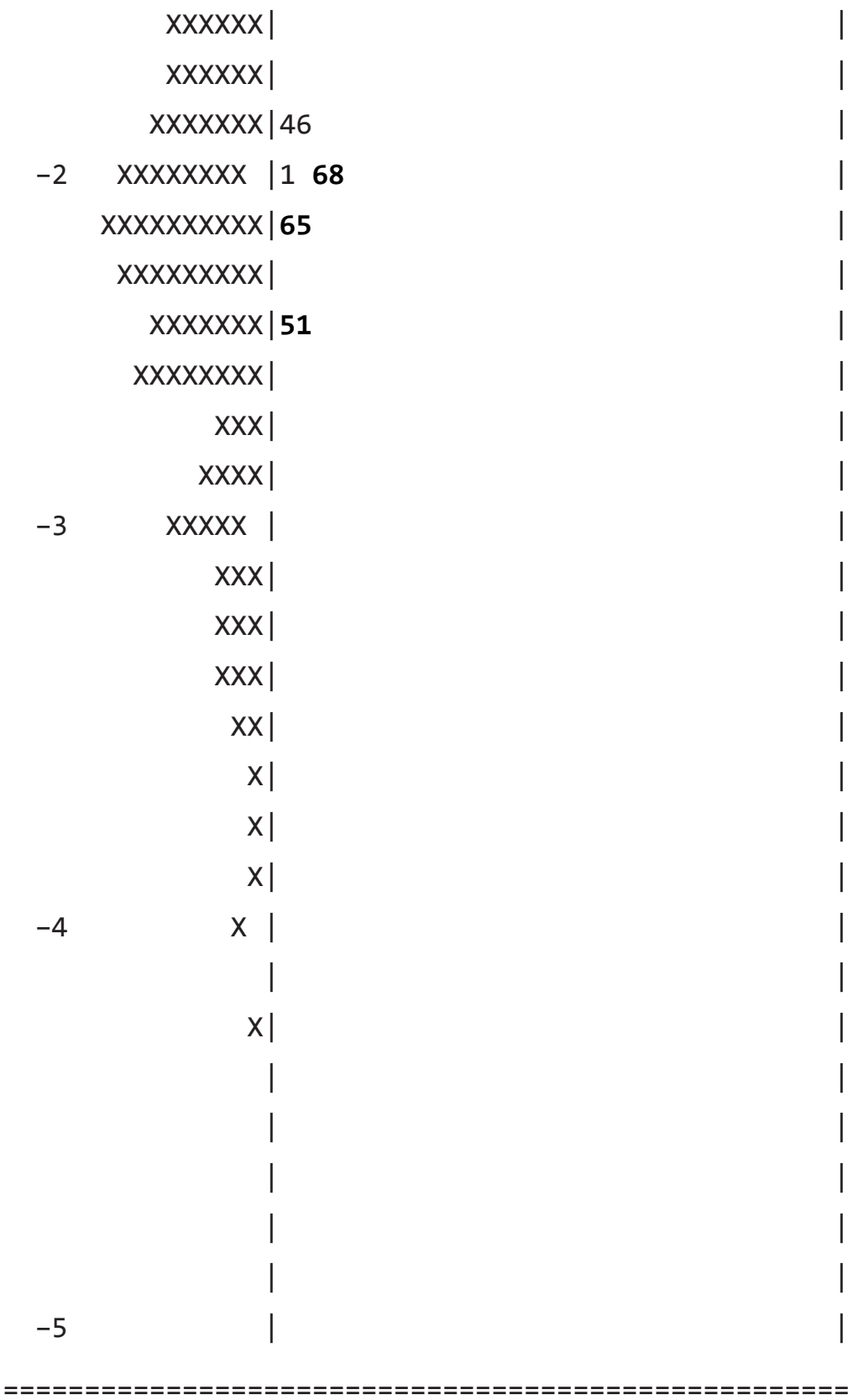


Figure 11: Item person map for Mathematics papers 1 and 2, 2009

In figure 11 the items on the right-hand side in bold represent the items of Mathematics paper 2, while the others represent the items of Mathematics paper 1. The higher the items are on the right-hand side, the more difficult they are. There are many items below 0, which shows that Mathematics papers 1 and 2 included many easy items. On the left-hand side of the histogram, the candidates are represented with Xs (one X = 1.5 cases; candidates wrote both papers 1 and 2), and the candidates are plotted against their latent trait values which range from –5 to +2. The higher the latent trait value, the more able the candidate; average candidates would be clustered around 0. It is evident that most of the candidates were below the zero-ability mark – most candidates were far below 'average'. Candidates on the lower left would have a very small chance of being able to do items on the upper right (Yu, 2010). For example, candidates with ability –1 should be able to do all items (on the right-hand side) to the right and below –1. Figure 11, thus, illustrates that most items were too difficult for most of the candidates.

Conclusion

In general, Mathematics 2009 was difficult for this cohort of learners, as there were too many questions that were not accessible to the learners as depicted in figure11, despite the fact that Rasch analysis indicated a fair number of easy items. The item is inaccessible to the learner if its difficulty level is higher than the ability of the learner. There were too many questions that were inaccessible to the most proficient learners as shown in figure 11. There were also a number of learners who did not have a chance on the simplest question on both papers. Expected score curves showed that, although a number of questions differentiated among low-ability candidates, there were no questions to discriminate among candidates with average to high abilities. The Rasch level of difficulty also confirmed the AMESA report that paper 1 was more difficult than paper 2. The infit and outfit statistics were acceptable for most of the questions. Examiners must be reminded of the questions which showed poor infit and outfit statistics.

Rasch analysis has been used extensively to test and analyse items of large-scale examinations (Wendt *et al*., 2011). This method can, therefore, also be used to test and analyse items and examination papers in the NSC examinations which are high-stakes examinations for South African school leavers. Rasch analysis could further assist in comparing standards across the assessment bodies (Independent Examination Body and Department of Basic Education examinations in the South African context) and across qualifications (NSC and the National Senior Certificate for Adults [NASCA]).

Since the candidates' abilities and the item difficulties are evaluated on the same measurement scale, one can tell whether the examination was too easy or too difficult for the cohort of candidates or whether the paper had a good targeting (a

paper that is not too difficult for the weak candidates and not too easy for the strong candidates). In the ideal paper the distribution of the difficulty levels of the items should be embedded in the distribution of the candidates' abilities.

References

Association for Mathematics Education of South Africa 2009. *Report of the AMESA curriculum committee on the 2009 senior certificate mathematics and mathematical literacy papers.* Retrieved on 21 April 2011 from: http://academic.sun.ac.za/mathed/amesa/2009%20SC%20exam.pdf.

Association for Mathematics Education of South AfricaAMESA 2010. *Report of the AMESA curriculum committee on the 2010 senior certificate mathematics and mathematical literacy papers.* Retrieved on 21 April 2011 from: http://academic.sun.ac.za/mathed/amesa/Maths%20Exam%20Response%20 2010.pdf.

Baker FB 2001. *The basics of item response theory.* 2nd ed. United States of America: ERIC Clearinghouse on Assessment and Evaluation.

Department of Basic Education 2011. *Report on the national senior certificate results 2010.* South Africa: Department of Basic Education.

Eastern Cape Education Department 2009. *Chief marker's report: Mathematics paper 1.* Eastern Cape Department of Education.

Edwards A & Alcock L 2010. Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics and its Applications,* 29: 165-175.

Grussendorf S, Booyse C & Burroughs E 2010. *Evaluating the South African national senior certificate in relation to selected international qualifications: A self-referencing exercise to determine the standing of the NSC.* Pretoria: HESA.

Hambleton RK, Swaminathan H & Rogers HJ 1991. *Fundamentals of item reponse theory*. California, SA: SAGE.

Hwang J & Davies PL 2009. Rasch analysis of the school function assessment provides additional evidence for the internal validity of the activity performance scales. *American Journal of Occupational Therapy*, 63: 363-373.

Keeton M 2010. Matric results not good enough, corporate social investment management CSI agency South Africa. Retrieved on 21 April 2011 from: http://www.tshikululu.org.za/thought-leadership/matric-results-not-good-enough-must-do-better/.

Prieto L, Alonso J & Lamarca R 2003. Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes*, 1: 27.

Smith AB, Rush R, Fallowfield LJ, Velikova G & Sharpe M 2008. Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, 8: 33.

Umalusi 2009. *2008 maintaining standards report: Part 3: Exam paper analysis.* Pretoria: Umalusi.

Van der Berg S & Taylor S 2010. *An analysis of 2009 IEB mathematics, DOE mathematics and DOE physical science papers using Rasch analysis tools*. Unpublished research report. Stellenbosch: University of Stellenbosch.

Wendt H, Bos W & Goy M 2011. On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation*, 17(6): 419-446.

Wilson TM & Macgillivray HL 2007. Counting on the basics: Mathematical skills among tertiary entrants. *International Journal of Mathematical Education in Science and Technology*, 38(1): 19-41.

Wu M & Adams R 2007. *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions.

Yu CH 2010. *A simple guide to the item response theory (IRT) and Rasch modelling.* Retrieved on 14 February 2011 from: http://www.creative-wisdom.com.