An investigation of Mathematical Literacy assessment supported by an application of Rasch measurement

Authors:
Caroline Long[1]
Sarah Bansilal[2]
Rajan Debba[2]

Affiliations:
[1]Faculty of Education, University of Pretoria, South Africa

[2]Department of Mathematics Education, University of KwaZulu-Natal, South Africa

Correspondence to:
Sarah Bansilal

Email:
Bansilals@ukzn.ac.za

Postal address:
Private Bag X03, Ashwood 3605, South Africa

Mathematical Literacy (ML) is a relatively new school subject that learners study in the final 3 years of high school and is examined as a matric subject. An investigation of a 2009 provincial examination written by matric pupils was conducted on both the curriculum elements of the test and learner performance. In this study we supplement the prior qualitative investigation with an application of Rasch measurement theory to review and revise the scoring procedures so as to better reflect scoring intentions. In an application of the Rasch model, checks are made on the test as a whole, the items and the learner responses, to ensure coherence of the instrument for the particular reference group, in this case Mathematical Literacy learners in one high school. In this article, we focus on the scoring of polytomous items, that is, items that are scored 0, 1, 2 … *m*. We found in some instances indiscriminate mark allocations, which contravened assessment and measurement principles. Through the investigation of each item, the associated scoring logic and the output of the Rasch analysis, rescoring was explored. We report here on the analysis of the test prior to rescoring, the analysis and rescoring of individual items and the post rescore analysis. The purpose of the article is to address the question: How may detailed attention to the scoring of the items in a Mathematical Literacy test, through theoretical investigation and the application of the Rasch model, contribute to a more informative and coherent outcome?

Background to the study

The subject Mathematical Literacy (ML), introduced in 2006 in South Africa, is a compulsory subject for those Grade 10–12 learners who do not study Mathematics. The purpose in ML is not that learners learn more and higher mathematics: the emphasis in ML is on the use of mathematics to explore the meaning and implications of quantitative information presented in many real-life situations.

The Department of Education (2003) defines ML as follows:

> Mathematical literacy provides learners with an awareness and understanding of the role that mathematics plays in the modern world. Mathematical literacy is a subject driven by life-related applications of mathematics. It enables learners to develop the ability and confidence to think numerically and spatially in order to interpret and critically analyse everyday solutions and to solve problems. (p. 3)

ML differs from Mathematics in purpose and in content. In Mathematics, emphasis is placed on engaging with increasingly more complex and abstract mathematical concepts, the relations between them and some applications to problems. However, in ML the emphasis is specifically on the application of basic mathematics to understand situations in real life. There is some lack of clarity evident in the description of Mathematical Literacy noted above, as being mathematically literate requires a sound mathematical base of algebraic concepts and skills. The debate about the subject content of Mathematics and Mathematical Literacy, though regarded as critical, is not the focus of this article.

The juxtaposition of mathematics content with real life contexts has meant that many people are unclear about how competence or proficiency in ML may be demonstrated. It is clear that ML as a subject in its infancy requires much research in respect of teaching, learning and assessment, in order to generate debate and establish some consensus on the many contrasting perspectives within the ML field.

In this study the focus is the Grade 12 ML preparatory examination, which was set by a provincial Department of Education and is intended to prepare students for the final examination. We pay attention to one aspect, that is, assessment in ML, by identifying some issues arising from the analysis of the empirical data obtained from learners' responses to this provincial preparatory assessment. The construct under scrutiny is the notion of proficiency in the subject ML. We apply Rasch measurement theory (RMT) to investigate the validity and accuracy of the test in providing

a measurement-like representation of ML proficiency in terms of person proficiency and item difficulty.

A valid and reliable test would provide teachers with some indication of the levels of mastery of curricular elements and of developing proficiency in ML. It should also provide the Department of Education with an overview of the entire learner cohort taking ML. An application of the Rasch model will help us to identify anomalies and inconsistencies amongst these assessment items and the accompanying scoring rubrics and working memoranda.

In this article we consider the implications of considering the purpose of the test and the construction of rubrics so that they work coherently in the interests of valid measurement-like properties and consequently provide reliable information for teachers. Some concepts underlying the Rasch measurement theory are introduced to clarify the analytic process. The aim of the article is to investigate the domain of ML and offer some observations concerning the assessment of ML.

Methodology

This study, focused on the scoring of items, is part of a larger study on ML (Debba, 2012). The instrument investigated here comprised 51 items, two of which were dichotomous items, marked either correct or incorrect. Twenty items had a maximum of two marks, 14 items a maximum of three marks, 9 items had four marks and 6 items had five marks. The maximum possible score was 150. The participants in the study were 73 Grade 12 ML learners.[1]

The Grade 12 KZN provincial preparatory ML examination paper is intended to assist Grade 12 learners in their preparation for the final examination. It is set by a team of examiners selected by the education department and written under examination conditions. For the purposes of this study the ML 2009 preparatory test was re-marked by the third author to ensure that the final version of the marking was entirely consistent with the marking memorandum supplied by the KZN Department of Education. A Rasch analysis supported the investigation of the test as a whole, the items and the ML learner responses. This analysis was conducted to identify factors that may have affected the coherence of the instrument for this sample of learners.

The first requirement for this analysis was to capture the score obtained in each of the 51 items for each of the 73 students.[2] The Rasch model offers various statistics to help diagnose where the data differs from what is expected by the model. Multiple means are applied to an analysis of this nature, to enable the subject expert to make an informed judgement.

The output provides statistics on the test as a whole as well as the individual item statistics, in particular the fit residual statistic and the chi square probability statistic, which provide information on the fit of the items. In addition to these statistics we investigated the item characteristic curves (ICCs) to identify which items were misfitting in the ways to be discussed. The research question directing the study is: How may detailed attention to the scoring of the items in a Mathematical Literacy test through theoretical investigation and the application of the Rasch model contribute to a more consistent outcome?

Assessment and measurement

We note that Mathematical Literacy and its assessment have been introduced relatively recently into the South African high school curriculum. We agree with Matters (2009) that assessment in the 21st century has a powerful influence, but this influence is only warranted if the assessment is of a sufficient quality to support the inferences, in this case the inferences about the mathematical literacy proficiency of learners that are drawn from the test results. The assessment process involves the theoretical exploration of the construct of mathematical literacy, the operationalisation of the construct in items designed to gauge proficiency, the compilation of a test instrument and the administration and marking.

From the classical theory of measurement, and measurement in the physical sciences (Wright, 1997), we note that the property of invariance of comparisons across the scale of measurement is a requirement. The application of the Rasch model enables the calibration of item measures and the estimation of person locations on a common continuum that together fit the criteria of invariance for a particular frame of reference (Rasch, 1960/1980; Humphry, 2005; Humphry & Andrich, 2008). The comparative difficulty of any two items should be constant regardless of the abilities of the persons responding to the items. Where the data do not conform to the measurement principles, the model will highlight the anomalies for further investigation. In the current study, the application of a Rasch analysis highlighted anomalies and inconsistencies that constituted threats to the construction of measures in the sense understood in classical measurement theory. In particular, the allocation of marks was inconsistent with the grading of proficiency along a continuum. In this article, we investigate the outcomes of both the initial scoring memorandum and the revisions, utilising both Rasch analysis and the educational considerations.

Rasch measurement theory

The fundamentals of Rasch measurement theory (RMT) are covered in many publications (Andrich, 1988; Rasch, 1960/1980; Wright & Stone, 1979, 1999). Here we note that with RMT there is an assumption that for the construct of interest there exists a latent trait in the learner that may be gauged through the operationalisation of the construct in various items. Both learner ability, denoted by β_n, and item difficulty, denoted by δ_i, may be represented on the same scale. This explanation is presented as follows by Dunne, Long, Craig and Venter (2012):

1.The small sample size may in some senses present as a limitation, but should not detract from the study's usefulness in alerting ML educators to the issues identified here. Any teacher of a Grade 12 ML class is likely to be concerned with fewer than 73 learners' performance on any such test. Larger counts of learners may occur in schools with several ML classes at Grade 12 level. The general rule of thumb for the construction and development of test instruments is that the learner count is about ten times the maximum score count. In the case of this study, the information obtained from the small group is cross-referenced with substantive analysis and therefore generalisable in the sense that the same principles will apply.

2.Missing responses were allocated zeros.

Each outcome of an interaction between a person and an item is uncertain but has a probability governed only by these two characteristics, that is by person ability (β_n) and item difficulty (δ_i). The Rasch model avers that the arrays of numbers β_n and δ_i are on the same linear scale, so that all differences between arbitrary pairs of these numbers such as ($\beta_n - \delta_i$) and hence also ($\beta_n - \beta_m$) and ($\delta_i - \delta_j$) are meaningful. Through these differences we may not only assign probabilities to item outcomes but also measure the contrasts between ability levels of items, and offer stochastic interpretations of these contrasts. (p. 7)

Alignment of item difficulty and person proficiency on same scale

We have noted that a key feature of the Rasch model is that the difficulty of items is located on the same scale as the ability of the persons attempting those items, precisely because the construct of interest underpins both the design of the items and the proficiency of learners. The focus of the model is on the interaction between a person and an item and is premised on the probability that a person v with an ability β_v will answer correctly, or partially correctly, an item i of difficulty δ_i. The equation that relates the ability of learners and the difficulty of items is given by the logistic function:

$$P\{X_{vi} = x\} = \frac{e^{x(\beta_v - \delta_i)}}{1 + e^{x(\beta_v - \delta_i)}} \qquad \text{[Eqn 1]}$$

This function expresses the probability of a person v with ability β_v responding successfully on a dichotomous item i with two ordered categories, designated as 0 and 1. Here P is the probability, X_{vi} is the item score variable allocated to a response of person v on dichotomous item i, x is the response, either 0 or 1, β_v is the ability of person v and δ_i is the difficulty of item i (Dunne et al., 2012).

Applying Equation 1, we can see that if a person v is placed at the same location on the scale as an item i, then $\beta_v = \delta_i$, that is, $\beta_v - \delta_i = 0$, and the probability in Equation 1 is thus equal to 0.5 or 50%. Thus, any person will have a 50% chance of achieving a correct response to an item whose difficulty level is at the same location as the person's ability level. Similarly, if an item difficulty is above a person's ability location, then the person has a less than 50% chance of obtaining a correct response on that item, whilst for an item whose difficulty level is below that of the person's ability the person would have a greater than 50% chance of producing the correct response. In Figure 1, the person location is represented on the horizontal axis, with the probability of a correct response located on the vertical axis.

Dichotomous and polytomous item responses

The Rasch model was initially developed for the analysis of dichotomously scored test items.[3] However, in many cases, tests require items that are scored at graded levels of performance. Rasch (1960/1980) extended the model for dichotomously scored items to include a model for test items with more than two response categories, with possible scores of 0, 1, 2, … m[4]. These items are termed polytomous items.

....
3. See Dunne et al. (2012) for details of the analysis of dichotomous items.

4. In this study we use the Rasch partial credit model, which is the default model in the RUMM 2030 software.
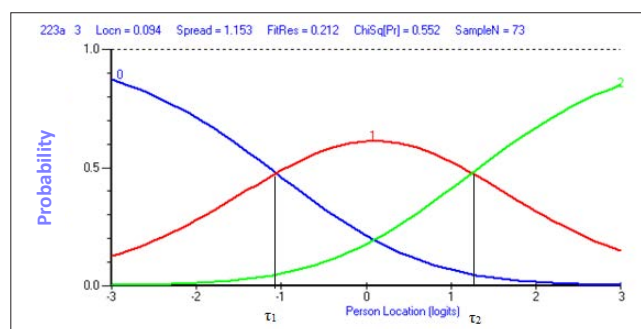
Note: Item 2.2.3 was one of the rescored items. The item characteristic curve (ICC) depicts the rescored item.

FIGURE 1: Category probability curves (Item 2.2.3).

Figure 1 models the conditional probability of a score of 0, 1 or 2 for a polytomous item (Item 2.2.3a) with three categories. As the person ability increases, the conditional probability of a score of 0 decreases. By contrast, as ability increases the probability of obtaining a maximum score of 2 increases. Also on this graph is the curve that shows the probability of a score of 1. In summary, this curve shows that when a person has very low ability relative to the item's location, then the probability of a response score of 0 is most likely; when a person is of moderate ability relative to the item's location, then the most likely score is 1 and when a person has an ability much greater than the item's location, then the most likely response score is 2 (see also Van Wyke & Andrich, 2006, p. 14).

In Figure 1, the thresholds,[5] and the categories they define, are naturally ordered in the sense that the threshold defining the two higher categories of achievement is of greater difficulty than the threshold defining the two lower categories of achievement. The first threshold (τ_1), which represents the point where a score of 1 becomes more likely than a score of 0, is about −1.10 logits. The second threshold, where a score of 2 becomes more likely than a score of 1, is approximately 1.25 logits. These thresholds show that progressively more ability is required to score a 0, 1 or 2 respectively on this item (Van Wyke & Andrich, 2006, pp. 13–14).

Requirements of the model

We have noted that the central proposition for the Rasch model[6] is that the response of a learner to a dichotomous item is a function of both the item difficulty and the person ability and nothing else. The probability of a person achieving success on a particular item is entirely determined by the difference between the difficulty of the item and the learner's ability.

The principle underlying the Rasch model is:

[A] person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one. (Rasch, 1960/1980, p. 117)

....
5. The term threshold defines the transition point between two adjacent categories, for example scoring 0 and 1, or scoring 1 and 2.

6. The discussion here will concern the dichotomous model. Extensions of the model have been derived from this model for partial credit scoring by Masters (1982) and rating scales by Andrich (1978).

In RMT it is expected that the data will accord well with the model. The notion of 'fit', that is, accord with the model, is defined as 'the correspondence between a data set and a statistical model' (Douglas, 1982, p. 32). The model provides indicators that alert researchers to where this principle of invariance of comparisons is not being met, which may result in item misfit. The fit residual is a measure of the difference between the observed response of each person to each item and that predicted by the model. The analysis process, whether showing a degree of conformity with the model or not, inevitably leads to greater understanding of the construct in question.

Item misfit

As noted, it may be observed that the items are working well and are a good indicator of the learners' proficiency. It may also be the case however that some items are highlighted as problematic. In subsequent sections we refer to particular examples where we focus on item functioning and the scoring rubrics. In some of the examples, the Rasch model analysis confirms that the scoring rubric is working as required by the model. In other items, the analysis discloses that the scoring rubric is not working in an ordinal way.

In a Rasch analysis test of fit, the learners are placed into class intervals of approximately equal size. We have used four groups. The mean ability of the four groups becomes the horizontal coordinate of points in the diagrams, depicting the probability of answering correctly.

Where the data conform to the model, the theoretical curve (the expected frequencies) and the observed proportions (the empirically established average of the actual item scores in the four chosen groups) are in alignment. Figure 2 shows the theoretical curve as expected by the model and the observed proportions, represented by black dots.

Where the theoretical curve and the observed proportions are in alignment we assume fit to the model, but where the theoretical curve and the observed proportions deviate substantially we are alerted to some kind of misfit between the data and the model. There are four broad categorisations that describe how the observed proportions might relate to the theoretical expectation. In this section we describe a selection of items that fall into the categories of *fairly good fit*, *under-discrimination*, *over-discrimination* and *haphazard misfit*.

Firstly, the observed proportions may align with the theoretical curve, in which case there is a good fit to the theoretical requirement. Figure 2 shows the item characteristic curve (ICC) for Item 2.1.3, in which the observed proportions are aligned fairly well with the theoretical curve. Note that the fit residual, 0.601, is relatively small tending towards zero and within an acceptable range of good fit (–2.5 to +2.5). This relatively small residual means that the difference between the observed response of each person to each item and the expected response is small.

A second phenomenon may be that the observed proportions are flatter than the theoretical curve, in which case the

item does not discriminate enough. The pattern is labelled under-discrimination or underfit and is illustrated in Figure 3. This unexpected pattern indicates that learners of lower proficiency appear to perform better than expected on this item and consequently, because of the interactive nature of item difficulty and learner ability, the high proficiency learners are falsely estimated to respond to the item as if the item was easier than it really was. The qualitative analysis suggests that a possible explanation lies with the marking rubric, which gives scores between 0 and 3. The scoring rubric allocates an arbitrary method mark and an additional mark for presenting information provided in the instruction. The allocation of marks appears to be more generous for the poorly performing learners and too constrained for the higher performing learners.

Figure 3 presents the ICC for Item 3.2.2, which shows the observed proportions to be flatter than the expected theoretical curve. The fit residual indicating difference between observed response and that expected by the model is relatively high at 2.410.

A third general category occurs when the observed proportions are steeper than the theoretical curves, in which case the discrimination is greater than expected, as shown in Figure 4. Over-discrimination in an item may unduly
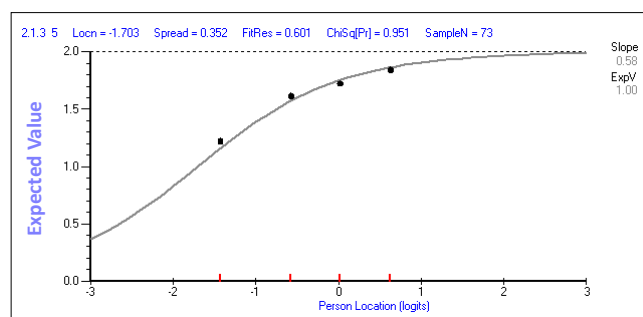

FIGURE 2: Item characteristic curve for Item 2.1.3, indicating fairly good fit.
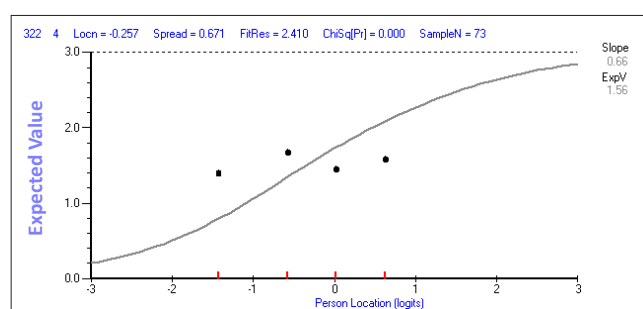

FIGURE 3: Item characteristic curve for Item 3.2.2, indicating under-discrimination.
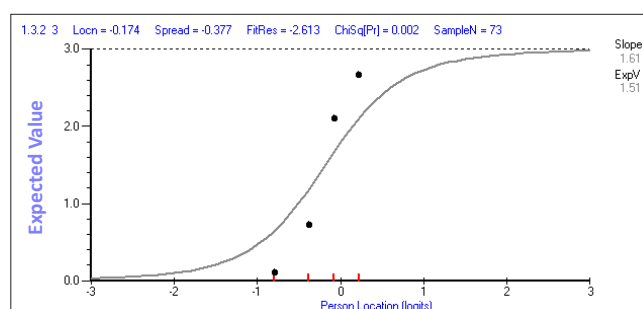

FIGURE 4: Item characteristic curve for Item 1.3.2, indicating over-discrimination.

TABLE 1: Summary statistics prior to rescoring.

Statistic	Items [N = 51]		Persons [N = 73]	
	Location	Fit residual	Location	Fit residual
Mean	0.0000	0.1136	−0.2537	0.0026
SD	1.1378	1.0727	0.3988	0.8333
	Cronbach's alpha = 0.8845		Person separation index = 0.8851	

N = number.

advantage high proficiency learners, whilst disadvantaging learners of lower proficiency. Whilst traditional test theory asserts the greater the item discrimination the better, the case in RMT is that highly discriminating items provoke a concern that there is a marked dependence amongst responses in one form or another. An example of such a misfit is that of Item 1.3.2, shown in Figure 4. Again we note that the fit residual is relatively high at −2.613. Both a high negative fit residual and a high positive fit residual signal poor fit to the model.

The fourth general category occurs when the observed proportions are haphazardly but substantially different from the theoretical requirement, as in Figure 5. This pattern demands specific investigation of the construct, an examination of the suitability of the item or the identification of another educational explanation. After analysis, Item 4.1.1 was deleted from the test on the grounds of its misfit. This excision is discussed in the section *Refinement of the instrument*. Note that here the fit residual is 3.321, indicating a fairly large deviation from the model that should be investigated.

Results from initial analysis

From the initial Rasch analysis, the summary statistics (Table 1), person-item location distribution (Figure 6) and person-item threshold distribution (Figure 7) were generated. Table 1 presents the initial summary statistics, which report the item mean as 0 (as set by the model) and the person mean as −0.2537. The standard deviation for the item location is 1.1378, whilst the standard deviation of the person location is just 0.3988. This contrast suggests that the spread of the items is high whilst the person locations are clustered together. Cronbach's alpha and the person separation index both indicate internal consistency reliability.

Figure 6 illustrates the person-item location distribution (PILD). The item location mean is set at zero; the person location mean is calibrated at −0.254. The item locations range from −2.2035 to 4.565 logits. The person locations are estimated between −1.414 and 0.441. The fact that the person location mean is lower than the item location mean suggests that the test was difficult for this particular learner cohort. Reasons for the mismatch[7] may be posited, for example the test questions might have been more easily answered if the cohort had been afforded more experience in basic algebra. Discussion of explanatory conditions and factors may be found in Debba (2012).

The person-item threshold distribution (PITD) (Figure 7) shows that the spread of the item threshold location ranges

7.This mismatch is in itself not a problem; however, more information could be gleaned from a test situation that is better targeted.
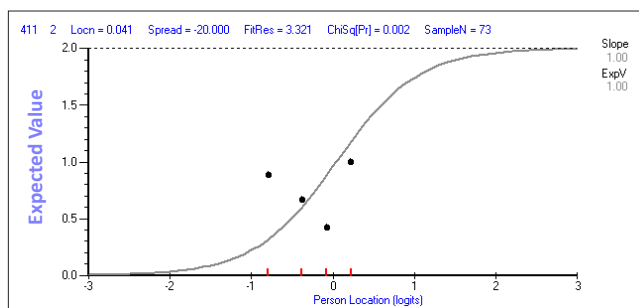
FIGURE 5: Item characteristic curve for Item 4.1.1, indicating haphazard misfit.
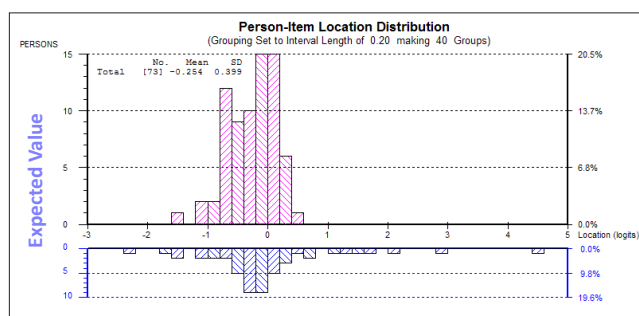
FIGURE 6: The person-item location distribution prior to rescoring.
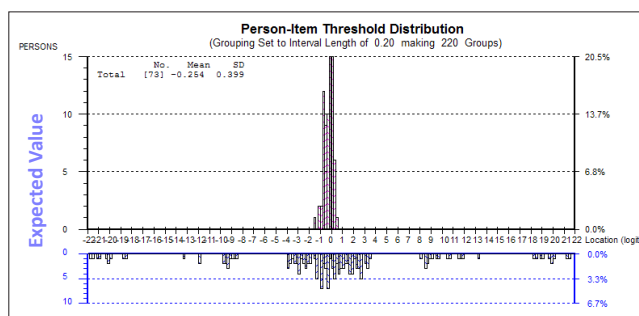
FIGURE 7: The person-item location distribution prior to rescoring.

from −22 logits up to 22 logits, whilst the person location is from −1.4 to 0.4. The PITD representation indicates the distribution of the various categories in the items, for example an item that was weighted at 3 marks will have three thresholds. This wide distribution suggests that some of the 51 items may have been awarded too high a score. This is supported by the fact that for many items several of their possible score values between 0, 1, 2 … m, were not observed in this class of 73 learners or observed only once. There are at least 40 problematic thresholds. See the frequency chart (Appendix 3).

Clearly this odd arrangement of thresholds and persons is out of alignment with what is expected of a balanced assessment. The items as originally conceptualised are not distinguishing the intended range of proficiency levels in

this particular set of learners. The detailed discussion of thresholds, and the disordering of thresholds is not presented here. See Andrich (2012) for a detailed discussion.

Both the summary statistics (see Table 1) and the PITD (see Figure 7) indicate some disjuncture between the items and the persons suggesting that there are some anomalies in the data. Further investigation is required for both items and persons in terms of fit to the model[8]. In this article, and in the next section, we focus only on the possible anomalies that arise due to the scoring of items.

Individual item analysis

In this section we present a short discussion about problems identified in particular items. We describe how the problems were highlighted and how the qualitative verification of measurement problems prompted rescoring. Three items are discussed, firstly Item 3.2.3, an example of an item that showed *haphazard misfit* (see Figure 5), Item 4.1.4, which shows how the *allocation of two marks* is not warranted, and Item 1.3.2, which shows *disordered thresholds.*

Item 3.2.3

Item 3.2.3a forms part of a question with a farming context (see Appendix 2). The task is to determine whether doubling the dimensions of the bucket will double the volume of the bucket. The question requires a yes or no answer, and in requiring this response may not gauge the understanding of dimensions or of volume:

Item 3.2.3: Farming context	
3.2.3.	Sipho decides to reduce the time spent walking to carry the milk to the farmer's house by first pouring milk from each cow into a larger second bucket. The dimensions of the second bucket are double the dimensions of the first bucket.
3.2.3a	Using this second bucket, do you think Sipho will double the amount of milk he takes to the farmer's house per trip? (1)

The ICC for Item 3.2.3a (Figure 8) shows a *haphazard misfit*, with learners of lowest proficiency on the test as a whole having an almost equal chance of obtaining a correct answer as learners of high proficiency on the test as a whole. Information provided here and the qualitative investigation suggests a revision of wording of this question. It is possible that learners at the lower end of the proficiency scale could have randomly chosen yes or no.

Item 4.1.4c

Item 4.1.4c requires that the learner give two causes for the observed relative change in a child's weight. The scoring rubric *allocated two marks per reason*. It was found that no learner obtained 1 mark without obtaining 2 marks and similarly no learner obtained 3 marks without obtaining 4 marks. The flat

category curves in Figure 9 show that the categories 1 and 3 are not functioning at all and category 2 is not functioning well. This outcome reflects the fact that the learners either got 0 marks (for providing no reason) or 4 marks (for providing two reasons). Only rarely did a learner offer only one reason.

Item 4.1.4

4.1.4	Peter's weight at birth was at 50th percentile.[1] When he was 3 years, he weighed 13 kg as indicated on the chart provided.
4.1.4c	Provide any two causes for this relative change in Peter's mass from 33 months to 36 months. (4)

Item 1.3.2

In Item 1.3.2 the learner is required to calculate the deduction to his wages. This item was part of a broader problem context (Item 1.3), which required calculating John's wages, where the hourly pay was provided, the number of hours worked per day and the percentage deducted.

Item 1.3: Wages context	
1.3	John was paid R11,50 per hour. In one week, he worked 9 hours a day, Monday to Friday. His employer deducted (subtracted) from his gross wages, 1% for sick leave and 1,2% for the Unemployment Insurance Fund (UIF).
1.3.1	Workers are allowed to work 40 hours a week at the normal rate and, above that, additional hours are regarded as overtime. They must be paid overtime at a rate of R15,75 per hour. Calculate John's gross wages for this week.
1.3.2	Determine the deduction from John's wages toward his sick leave and UIF in this week. (3)

Details of marking scheme for 1.3.2:

$$\text{Total deductions} = 1\% + 1,2\% \qquad 1 = M$$
$$= 2,2\% \; \checkmark$$
$$\therefore \text{Deduction} = \frac{2,2}{100} \times \frac{538,75}{1} \qquad 1 = CA$$
$$= R11,85\checkmark \qquad 1 = CA$$

A = Answer/accuracy, M = Method, CA = Consistent accuracy

The item itself is problematic as it depends on the learner answering the previous item (1.3.1) correctly. In addition, the scoring of part marks is odd, as once the learner identifies the 2,2% total deduction, they are likely with the help of a calculator to obtain a correct answer.
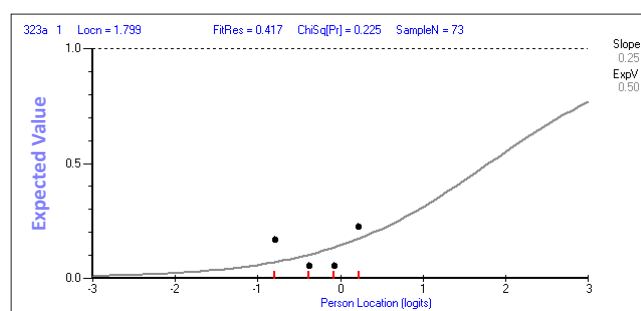

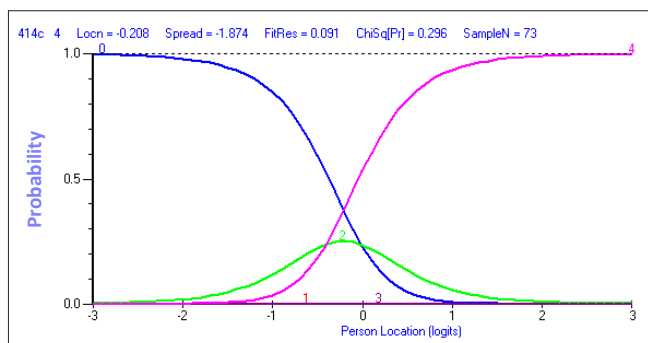**FIGURE 8:** Item characteristic curve (Item 3.2.3a).

8.In addition, the investigation of factors such as response dependency and differential item functioning is demanded in the interests of valid measurement.

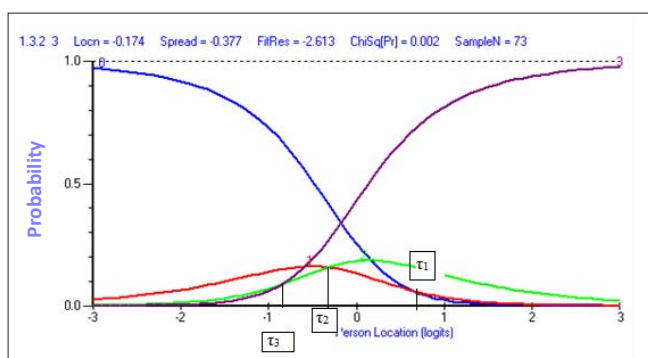**FIGURE 9:** Category probability curves for Item 4.1.4.c.

FIGURE 10: Initial category probably curves for Item 1.3.2.

The category probability curves for Item 1.3.2 (Figure 10) are used to illustrate *disordered thresholds*. Figure 10 shows that the location of the first threshold (τ_1) (the intersection of the curves for score 0 and score 1) has a difficulty of approximately 0.83 logits. However, the location of the second threshold (τ_2) (the intersection of the curves for a score 1 and score 2) has a difficulty of approximately –0.15 logits. The location of the third threshold (τ_3) (at the intersection of the curves of scores 2 and score 3) is approximately –0.68 logits.

The problem is that the location of the first threshold is greater than the location of the second threshold, whilst the location of the second threshold is also greater than the location of the third threshold. These reversed thresholds are due to the failure of the two middle categories, corresponding to scores of 1 and 2, to function as intended. At no point on the horizontal axis is a score of 1 most likely; neither is there an interval or point where a score of 2 is most likely. Although persons with low ability relative to the item's difficulty are still most likely to respond incorrectly and score 0 and persons with high ability relative to the same item's difficulty are still most likely to respond correctly and score 3, persons with ability in the range –0.68 logits to 0.83 logits, where a score of 1 or 2 should be most likely, are more likely to score either 0 or 3. This disordering is evident where the high middle group is more likely to score 1 and where the low middle group is more likely to score 2.

The disordering of the thresholds confounds the idea that thresholds between higher level categories are more difficult to attain than thresholds between lower order categories.

These initial analyses help us to identify possible anomalies and inconsistencies in the scoring rubrics, which can alert us to possibilities that should be considered when devising scoring rubrics. There are strategies we can use post hoc to adjust the item scoring in order to exhibit ordered thresholds as shown in Item 2.2.3a (Figure 1). The verification that scoring rubrics are functioning as expected, and subsequent revision where necessary, contribute to the reliability of the ML examination paper in the context of this set of 73 learners.

Refinement of the instrument

In addition to looking at misfit statistics, we studied the associated category probability curves for each item, to further explore anomalies in the data. When we investigated the category probability curves for each item, we found that in most cases the thresholds were disordered.

In the process of refining the instrument, iterative adjustments took place. The first step was to identify items that showed severe misfit according to the fit residual statistics and the chi square probability. In addition the ICCs were investigated. Where there were indications of anomalies we checked the item itself and the scoring rubric to identify any problems from both a mathematics education perspective and an assessment perspective. Where the qualitative investigation confirmed the anomaly and it was deemed proper to adjust the scoring of the item, this step was taken. The item statistics were then reinvestigated by reanalysing and by rechecking the fit to the model with the revised scores. If the fit had not improved we would reverse the change. If there was no theoretical reason to support the rescoring of an item, then no rescoring took place. The details of the various items, together with the marking memorandum provided by the Department of Education, and details of the rescoring appear in Appendix 2.

First round of rescoring

The scoring for the two dichotomous items, Item 2.2.1b and Item 3.2.3a, was retained. All other items were rescored according to the process identified above.

One of the items, Item 4.1.1 (Figure 7), showed extreme misfit when investigating the ICC. In addition the fit residual statistic (3.321) was outside the generally acceptable interval of between –2.5 and 2.5 and the chi-square probability (0.002) showed that the expected and the observed outcome were statistically significantly different. This item warranted further investigation to see whether the problem lay with the scoring and whether rescoring may resolve the misfit. After studying the evidence and finding that there were problems with the item which resulted in the item not contributing to the test, we deleted the item from the test. In Item 4.1.1 learners were given four graphs from a growth chart used by parents and health workers to monitor the weight gain of infants. They were asked to find the normal weight of a baby boy at birth. Two curves represented the weight of boys whilst two represented the weight of girls. The poor quality of the graphs and unclear titles contributed to the difficulties with this item. In addition, it was not clear how

these graphs could be used to provide information about a 'normal weight'. The assessment task did not make sense in the real-life context. This confusion of meanings illustrates a fundamental tension that exists between the intentions of assessment task designers and learner participants in the real-life context. The questions posed in the examination paper may not be the ones that are posed in the context of health workers who use growth charts to identify children whose health is at risk.

Results of first round of rescoring where the rescoring worked well

After refining the scoring, we found that in several cases the rescoring improved the fit, whilst in few other cases it did not. The rescoring of Item 1.1.3 (see Box 1) resulted in the improved fit, the adjustment more closely approaching measurement principles. We provide the educational rationale for a change in scoring and also show the category curves both before rescoring and after rescoring to show how the rescoring helped improve the functioning of the categories.

Figure 11 shows that before the rescoring most of the categories were not working as they should, that is the allocation of scores 2 and 3 appeared somewhat redundant. Furthermore, the ICC for Item 1.1.3 in Figure 12 shows unduly high discrimination, labelled overfit. This problematic outcome may be explained as follows: learners obtaining the item's answer correctly are unduly advantaged by scoring 4 points, whereas those answering incorrectly are unduly disadvantaged by 'losing' 4 possible points.

The item, was rescored, moving from a five-category item (scoring 0, 1, 2, 3, 4) to a four-category item (scoring 0, 1, 2, 3). Category 1 remained the same, Category 2 and Category 3 were recoded as 2, whilst category 4 was recoded as 3. A qualitative investigation in this case indicates that scoring 0, 1, 2, 3 can be justified.

The details of the item and rescoring appear in Box 1. In each case where the categories were not working according to the model, we applied multiple criteria before revising the scoring, keeping in mind principles of best test design (see Van Wyke & Andrich, 2006; Wright & Stone, 1979). After this rescoring process, both the category curves (Figure 5) and the ICC (Figure 6) indicated better fit according to the model.

BOX 1: Details for Item 1.1.3.

Extraneous details of context omitted.	
International system	**Metric system**
1 pint (pt)	569 m
1 foot (ft)	0,3048 mL
1 inch	2,54 cm
1 foot = 12 inches	

1.1.3 You need ¾ of a 7 ft iron bar. How many cm is the iron bar that is needed?	
Original scoring of 1.1.3:	*Rescoring of 1.1.3*:
¾ of 7ft = 5,25 ft √	¾ of 7 ft = 5,25 ft √
= 5,25×12×2,54 √√	= 5,25×12×2,54 √
= 160 cm √	= 160 cm √
1 = A, 1 = C, 1 = A, 1 = CA	*We allocated 1 mark for the second step instead of 2*

A, answer/accuracy; M, method; CA, consistent accuracy.

Comparing Figure 11 and Figure 13, one can see that the categories are now working more appropriately. Figure 14 shows the ICC for Item 1.1.3 after the rescore. When comparing Figure 12 and Figure 14, one can see that the fit between the observed and expected means for the persons in the four class intervals is much improved. A final check of the fit residual statistic and corresponding chi square values indicates that initially the fit residual statistic was –0.794 (chi square probability = 0.4317), whilst the final
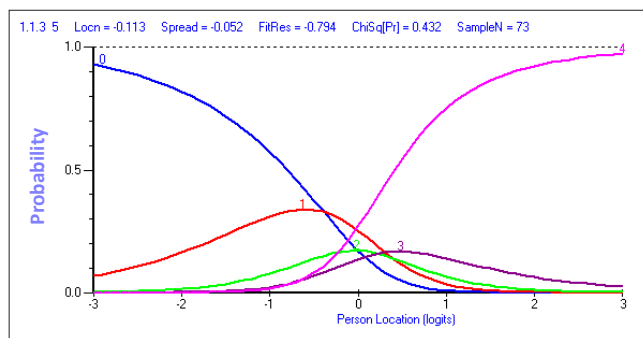

FIGURE 11: Initial category probability curve for Item 1.1.3.
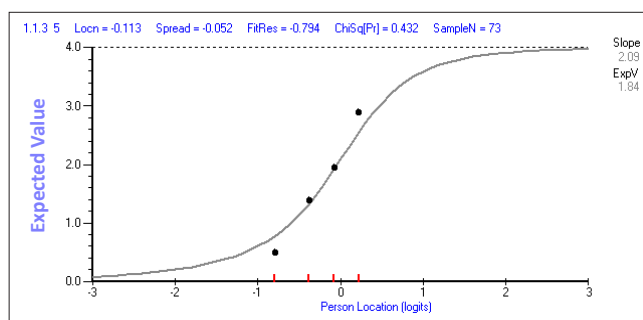

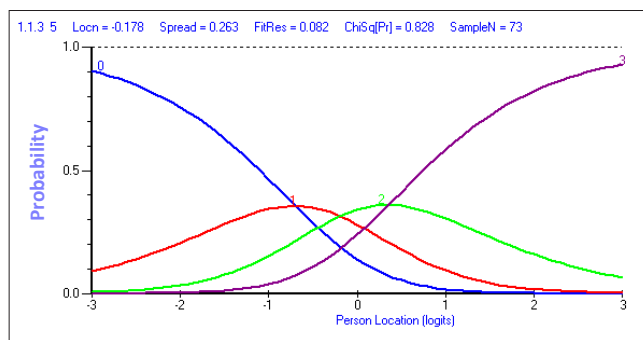FIGURE 12: Initial item characteristic curve for item 1.1.3.

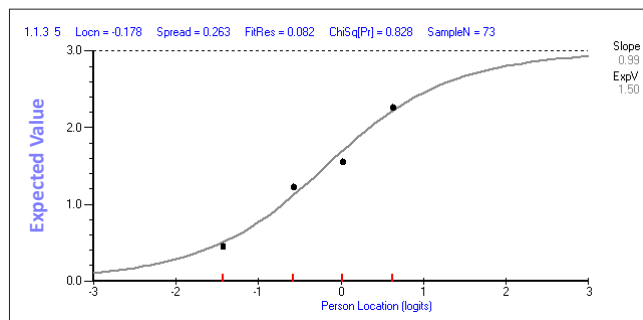FIGURE 13: Final category probability curve for Item 1.1.3

FIGURE 14: Final item characteristic curve for Item 1.1.3.

fit residual statistic is 0.082 (chi square probability = 0.8281), showing that the change affected the fit statistics positively. Also note that with RMT, a *p*-value higher than 0.01 for the fit statistics indicates that the difference between the observed and expected is not statistically discernible and hence that the model is working well.

Results of first round of rescoring where the rescoring did not work well

As explained earlier, Item 4.1.1 was rejected as the fit was extreme and in addition it was judged to be a very poor indicator of mathematical proficiency. In addition, Item 3.2.2, after rescoring, was now also identified as having extreme fit statistics outside of the acceptable interval.

In a second process we investigated each ICC again, and then the category probability curves, and found that the thresholds were disordered for Items 1.2.1b, 1.2.2, 1.3.1, 1.3.2, 3.2.1, 3.2.3b, 4.1.4b, 4.1.4c, 5.1.4 and 5.3. (See Appendix 2 for details of these items.) At this stage, we re-examined the questions to see whether we could justify rescoring these items again. We also examined the fit residual statistics to help us decide whether a rescoring was necessary or not. A summary of these processes and decisions appears in Appendix 1.

After resolving anomalies

After the refinement of the scoring using information from both the qualitative investigation and the Rasch analysis, we now examine the final summary statistics in Table 2.

The figures in Table 2 when compared to those of Table 1 show improvement across most of the statistics. For example, Cronbach's alpha, a reliability index, has increased from 0.8845 to 0.8887. The standard deviation for the item location has increased to 1.615, showing that the items now have a larger relative spread.

We note as well that the differences observed in the new person-item location distribution (PILD) (Figure 15) are an improvement on the initial PILD (Figure 2).

The distribution represented in Figure 15 indicates a better balance than the original distribution represented in Figure 2. We can see that the standard deviation for person locations has increased in the person-item location distribution from 0.399 to 0.808 (see Figure 6 and Figure 7). The person location spread was initially between −1.414 and 0.441. After the rescoring the spread of learner locations ranged between −2.460 and 1.176, thus providing greater discrimination of learner proficiencies. We judge the post-

rescoring estimates of learner proficiencies to be more accurately reflected by the refined scoring.

Furthermore, one can see that the most difficult item, Item 5.1.3, was far too difficult for this group of learners. Its estimated difficulty level, calibrated at 5.448 logits, is much higher than the ability level of the most proficient learner in this test, who was estimated just above 1 logit. The items 4.1.2, 4.1.3, 4.2.1, 4.2.2, 5.1.4, 5.2, 2.1.2 and 3.2.3a were also beyond the ability level of the learners, but the difference $\delta_i - \beta_v$ for this group of items was not as large as it was in the case of Item 5.1.3.

Item 5.1.3 required learners to work out the price of an item before VAT, given the final price including VAT. The usual solution involves setting up an equation of the form *A + 14% of A = Final price,* and they were required to find A. This item required the setup and solution of an equation and involves applying algebraic techniques. The ML curriculum document makes it clear that skill in such algebraic manipulation is not a focus of ML. The document states: 'As a rule of thumb, if the required calculations cannot be performed using a basic four-function calculator, then the calculation is in all likelihood not appropriate for Mathematical Literacy'(Department of Basic Education, 2011, p. 8).

Given such a stipulation, and noting that the item was far beyond the ability of this group of learners, it is recommended that a greater use of basic algebraic manipulation skills should be encouraged, to enable them to solve items such as Item 5.1.3. We note that this item could also be solved by informal methods based on reasoning about percentage change or using proportion and perhaps opportunities for such reasoning should also be encouraged.

On the other hand, three items, 4.2.3, 3.1.1 and 2.2.1a, were lower than the ability level of all the learners in the class, but the location of the items was less than 2 logits lower than the person with the lowest total score. This difference contrasts with the finding that in the higher ability ranking, one item
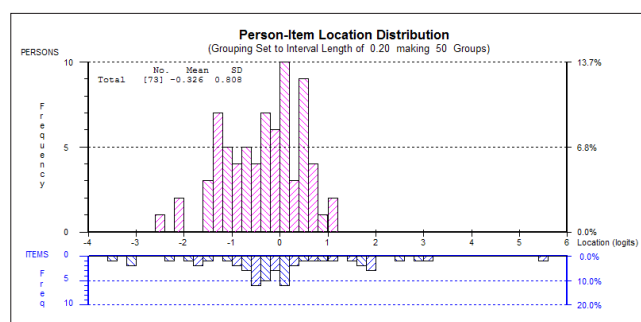
FIGURE 15: Person-item location distribution after refinement to instrument.

TABLE 2: Final summary statistics.

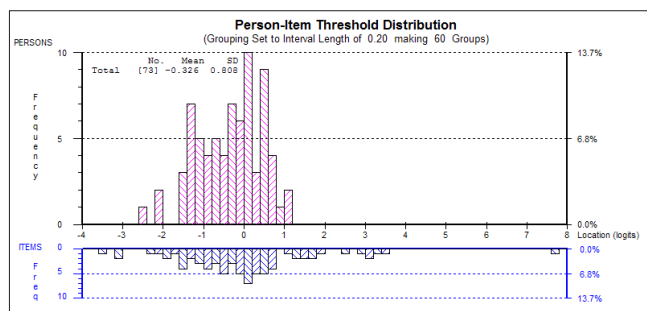| Statistic	Items [*N* = 51]			Persons [*N* = 73]	
	Location	Fit residual		Location	Fit residual
Mean	0.0000	0.00126		−0.3260	−0.0846
SD	1.6195	1.046		0.8083	0.9013
	Cronbach's alpha = 0.88873			Person separation index = 0.886	

N = number.

FIGURE 16: Person item threshold distribution post-rescoring.

was almost 5 logits higher than the location of the person with the highest total score, with two items being 2 logits or more higher and six other items being higher than the location of the person with the highest total score. We now present the person-item threshold distribution that was generated after the rescoring process.

There is a marked difference between the initial PITD (Figure 7) and the final PITD (Figure 16). The distribution of the thresholds prior to rescoring ranged between –22 and +22 logits (see Figure 7). Now after rescoring the threshold distribution has been limited to between –4 and 8 logits. The number of thresholds has been reduced from 150 to 75 by the rescoring processes (see Appendix 3).

Implications for assessment in Mathematical Literacy

In this study we examined the responses of 73 students to each of 51 items by using tools provided by the RUMM2030 software (Andrich, Sheridan & Luo, 2011). The analysis revealed some important issues, which are discussed below.

Allocation of marks for accuracy and method

We have discussed the case of Item 1.3.2, where marks were allocated for method and for consistent accuracy. This item was just one instance of many where the method and accuracy system did not work well. In fact, it unduly disadvantaged the person who did not address the question correctly. Most learners who identified the method were able to obtain the correct answer because it involved just entering the numbers into the calculator and reading off the answer. The learners who answered correctly were thus unduly advantaged by getting additional marks. There were very few learners (too few as revealed by their respective category probability curves in Figure 10) who achieved the method mark without achieving the accuracy mark.

Another related issue was in the case of Item 4.1.4c, which asked for two causes of an observed change in the data. The scoring rubric allocated 2 marks per reason. The category curves in Figure 9 revealed that the categories 1 and 3 were redundant. No learner attained the first mark without getting the second mark and similarly no learner attained 3 marks without getting the fourth mark.

Guessing

Questions that allocated a mark for answering yes or no to a question and then asked for a reason resulted in a haphazard response that could indicate guessing at the lower proficiency level. This pattern was revealed in some items by the ICC where there was a haphazard misfit, such as in Item 3.2.3 (shown in Figure 10) and Item 1.3.3 (not shown). We make a distinction here between the learners who know the answer to the question and the group who guess. It is the latter group who have a 50% chance of answering correctly if they guess.

Item response dependency

In some cases, the answer to a previous item influenced the probability of success of the learner in a following item, as in the case of Item 2.1.2a and Item 2.1.2b as well as Item 1.2.1a and Item 1.2.1b. Those learners who were able to get the first item correct were likely to get the second one correct. None of the learners who answered the first part of Item 2.1.2a incorrectly achieved full marks in the second part of the item, Item 2.1.2b, and none of the learners who answered Item 1.2.1a incorrectly was able to obtain full marks in Item 1.2.1b. The dependency of a subsequent item on an earlier item is regarded as unfavourable test practice. It is the independence of items that offers greater precision.

Concluding remarks

In this article we illustrated how the Rasch model could be used in conjunction with professional judgment to check the validity of the assessment, by using the responses of 73 Grade 12 ML learners to their preliminary examination. The process described in this article involved identifying items that did not fit the model. We described the items and the original scoring rubric. The Rasch output was provided and the anomalies and inconsistencies were discussed. Before initiating any changes in the scoring we sought educational reasons that warranted rescoring. Thus, the Rasch analysis and rescoring processes were guided at all times by the qualitative analysis that was conducted by an experienced ML teacher.

An important advantage of the Rasch analysis is that item difficulty and person proficiency are located on the same scale. Without checking individual item validity, we cannot take the results at face value; instead, we need to verify the item fit. Now that we have subjected the scoring of each item to Rasch analysis and qualitatively investigated the structure of the scoring rubrics, we argue that the new scoring procedures for the instrument allow greater precision than the original instrument scoring. It must be emphasised that the scoring rubrics were not the only threats to validity in this test.

The process that was conducted, which involved systematically rescoring each of the assessment items, has revealed the important role of scoring rubrics in contributing to the validity of assessments. Examiners need to ensure that each mark that is allocated can be justified educationally. Marks should not be allocated on the basis of time that is required to be spent during the examination. Neither should

marks be allocated for guessing. If these checks are not taken into account, then the total score has diminished meaning. If the total score is indicative of a position on a unidimensional scale, then differences between the score must reflect differences between proficiency levels of learners.

A further process that could be followed is to study which items were experienced as more difficult and which items were experienced as easier in order to obtain a description of items on different levels of the scale. It may then be possible to identify different demands of the various questions. A study of a similar nature has been conducted by Long (2011) who presented a comprehensive application of RMT to the multiplicative conceptual field. In her study, Long used the Rasch model to develop clusters of proficiency zones along the continuum representing the alignment of person proficiency and item difficulty.

It may be the case that particular items are mathematically demanding because they require sophisticated use of algebraic tools (as in Item 5.1.3). Perhaps the increasing demand of the questions may be related to the contexts that were used. These analyses will be conducted and reported in a follow-up article.

Acknowledgements

Competing interests

We declare that we have no financial or personal relationship(s) that might have inappropriately influenced us in writing this article.

Authors' contributions

The data was collected by R.D. (University of KwaZulu-Natal) as part of his post-graduate studies. C.L. (University of Pretoria) led the process of the Rasch analysis, helped by S.B. (University of KwaZulu-Natal) and R.D. The write-up was done by all three authors, led by C.L.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. http://dx.doi.org/10.1007/BF02293814

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: SAGE Publications.

Andrich, D. (2012). An expanded derivation of the threshold structure of the Polytomous Rasch Model that dispels any "threshold disorder controversy". *Educational and Psychological Measurement, 73*(1), 78–124. http://dx.doi.org/10.1177/0013164412450877

Andrich, D., Sheridan, B., & Luo, G. (2011). *RUMM2030 software and manuals*. Perth, Australia: University of Western Australia. Available from http://www.rummlab.com.au/

Debba, R. (2012). *An exploration of the strategies used by Grade 12 Mathematical Literacy learners when answering mathematical literacy examination questions based on a variety of real-life contexts*. Unpublished master's thesis. University of KwaZulu-Natal, Durban, South Africa.

Department of Basic Education (DBE). (2011). *Curriculum and assessment policy statement (CAPS). Mathematics Grades 10–12*. Pretoria: DBE.

Department of Education (DOE). (2003) *National curriculum statements Grades R–9 (schools).* Pretoria: DOE.

Douglas, G. (1982). Issues in the fit of data to psychometric models. *Education Research and Perspectives*, *9*(1), 32–43.

Dunne, T., Long, C., Craig, T., & Venter, E. (2012). Meeting the requirements of both classroom-based and systemic assessment of mathematics proficiency: The potential of Rasch measurement theory. *Pythagoras, 33*(3), Art. #19, 16 pages. http://dx.doi.org/10.4102/pythagoras.v33i3.19

Humphry, S.M. (2005). *Maintaining a common arbitrary unit in social measurement.* Unpublished doctoral dissertation. Murdoch University, Perth, Australia.

Humphry, S.M., & Andrich, D. (2008). Understanding the unit in the Rasch Model. *Journal of Applied Measurement, 9*(3), 249–264.

Long, C. (2011). *Mathematical, cognitive and didactic elements of the multiplicative conceptual field investigated within a Rasch assessment and measurement framework*. Unpublished doctoral dissertation. University of Cape Town, Cape Town, South Africa. Available from http://web.up.ac.za/sitefiles/file/43/314/Long,_M__C__(2011)__The_multiplicative_conceptual_field_investigated_within_a_Rasch_measurement_framework_.PDF

Masters G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. http://dx.doi.org/10.1007/BF02296272

Matters, G. (2009). A problematic leap in the use of test data: From performance to inference. In C. Wyatt-Smith, & J.J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 209–225). Dordrecht: Springer. http://dx.doi.org/10.1007/978-1-4020-9964-9_11

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests (expanded edn. with foreword and afterword by B.D. Wright). Chicago, IL: University of Chicago Press. (Original work published 1960)

Van Wyke, J., & Andrich, D. (2006). *A typology of polytomously scored items disclosed by the Rasch model: Implications for constructing a continuum of achievement*. Perth: Murdoch University.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, Winter, 33–45.

Wright, B.D., & Stone, M.H. (1979). The measurement model. In B.D Wright, & M.H Stone (Eds.), *Best test design* (pp. 1–17). Chicago, IL: MESA Press.

Wright, B.D., & Stone, M.H. (1999). *Measurement essentials.* Wilmington, DE: WIDE Range, Inc.

APPENDIX 1

TABLE 1–A: Decisions after results of first round of rescoring.

Item	Problem	Comment
1.2.1b	Poor fit observed in the item characteristic curve (ICC)	Rescore again, after confirming with qualitative data of scoring rubric that this action was justified.
1.2.2	Poor fit in ICC	After studying residual fit statistics, it was decided to leave in present form.
1.3.1	Poor fit in ICC	Rescore again, after confirming with an investigation of scoring rubric that it was justified.
1.3.2	Poor fit in ICC	By studying the fit residual statistics, we decided to leave in present form, since the rescoring exacerbated the fit residual.
3.2.1	Poor fit in ICC	After studying fit residual, it was decided to leave in present form as the qualitative analysis did not support a change in the scoring rubric. The poor fit may be due to the small sample size, with too few responses in some categories.
3.2.2	Extreme fit residual statistics, haphazard misfit, but category curves working well	After studying fit residual, it was decided to leave in present form as the qualitative analysis did not support a rescore.
3.2.3b	Haphazard misfit in ICC	Rescore again, after confirming with qualitative analysis of scoring rubric that it was reasonable. Perhaps if we had a larger sample, the category currently not functioning as expected may work better.
4.1.1	Extreme fit residual statistics, disordered thresholds, poor fit	After studying the different evidence, it was clear that the item was not working as intended and we decided to delete it from the test.
4.1.4 b	Poor fit in ICC	Re-score again, after confirming with qualitative data of scoring rubric that the rescore was justified.
4.1.4c	The ICC did not suggest a rescoring	However, after a study of the category probability curves, it was also decided to rescore this item.
5.1.4	Disordered thresholds, poor fit in ICC	Rescore again, after confirming with qualitative data of scoring rubric that it was reasonable.
5.3	Disordered thresholds, poor fit in ICC	Based on the fit residual it was decided to leave as is.

ICC, item characteristic curve.

APPENDIX 2

Questions, marking guide and rescoring details

The wording of the items and the marking guide are as they appear in the original documents (KwaZulu-Natal Department of Education, Mathematical Literacy Grade 12 trial examination paper 2, 2009).

TABLE 1: Explanation of symbols in the marking guide.	
Symbol	**Explanation**
A	Answer/accuracy
M	Method
CA	Consistent accuracy
MA	Method with accuracy
SF	Substitution
O	Opinion
R	Rounding off/reason
C	Conversion

Item 1.1.3

Extraneous details of context omitted

International system	**Metric system**
1 pint (pt)	569 m
1 foot(ft)	0,3048 mL
1 inch	2,54 cm
1 foot – 12 inches	

1.1.3 You need ¾ of a 7 ft iron bar. How many cm is the iron bar that is needed?

Marking scheme			**Rescoring details**
Item 1.1.3: ¾ of 7ft	= 5,25 ft √	1 = A	1→1
	= 5,25×12×2,54 √√	1 = C, 1 = A	2,3→2
	= 160 cm √	1=CA	4→3

Item 1.2

1.2 The company makes and paints nuts, better known as sockets like in Fig. A. Fig. C shows the sockets (hole of nut) at the bottom and Fig. B the hole at the top. The hole at the bottom is cylindrical. The hole at the top is cubic. The socket has a hole at the top and a hold at the bottom.
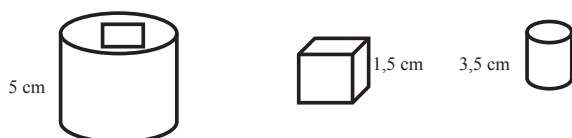
1.2.1 (a) The nut has a diameter of 1.5 cm.
Calculate the minimum circumference of the iron that could be used to make this nut. Use the formula C= 2 × π × r   (2)
(b) Calculate the surface area of the nut before the holes for the sockets are removed.
$SA = 2 \times \pi^2 \times r^2 + \pi \times d \times h$ (4)

1.2.2 The whole nut must be painted grey, then decorated with white paint at the top and red at the base/bottom (holes are not painted). To paint one nut, you need 25 mL of grey paint. How many 500 mL tins of paint will be needed to paint 1 200 nuts?   (4)

Marking scheme		**Rescoring details**
Item 1.2.1		
1.2.1(a) $C = 2\pi \times r$		
$= 2 \times \pi \times \frac{1,5}{2}$ cm √	1 = SF	1.2.1a1,2→1
$= 4,71$ cm √	1 = A	
1.2.1(b) SA of cylinder		
$= 2 \pi \times r^2 + \pi \times d \times h$		
$= 2(\pi)(\frac{1,5}{2} m)^2 \sqrt{} + \pi \times 1,5$ cm $\times 5$ cm√	2 = SF	
$= 3,53$ cm² $+ 23,55$ cm² √	1 = A	1.2.1b
(grey surface) $= 27,08$ cm² √	1 = CA	
		1,2→1
NB: Error on Q: $(\pi)^2$ 2 marks subst. + 2 marks for error		3,4→3
Wrong answer $= 11,09 + 23,55 = 34,64$ cm²		
Item 1.2.2		
1.2.2 One socket = 25 mL		
1200 sockets $= 25 \times 1200$ √	1 = MA	
$= 30\ 000$ mL √	1 = A	
No. of tins $= \frac{30\ 000\ mL}{500\ mL}$ √	1 = MA	1,2→1
$= 60$ tins √	1 = CA	3,4→2

Item 1.3

1.3 John was paid R11,50 per hour. In one week, he worked 9 hours a day, Monday to Friday. His employer deducted (subtracted) from his gross wages, 1% for sick leave and 1,2% for the Unemployment Insurance Fund (UIF).

1.3.1 Workers are allowed to work 40 hours a week at the normal rate, and above that, additional hours are regarded as overtime. They must be paid overtime at a rate of R15,75 per hour. Calculate John's gross wages for this week. (5)

1.3.2 Determine the deduction from John's wages toward his sick leave and UIF in this week. (3)

Marking scheme			Rescoring
Item 1.3.2 Total deductions	$= 1\% + 1,2\%$ $= 2,2\%$ √		
	$= \dfrac{2,2}{100} \times \dfrac{538,75}{1}$ √	1 = M 1 = CA	$1 \rightarrow 1$ $2,3 \rightarrow 2$
Deduction	$= R11,85$ √	1 = CA	

Item 2.1

2.1 The amount of work left at the stadium at the beginning of September 2009 can be completed in exactly 4 months by 200 workers. The number of workers employed and the time taken to complete the stadium is given in the table below.

TABLE 2: Number of workers and time taken.							
No. of workers	400	320	**B**	200	**C**	125	100
No. of months	2	**A**	3,2	4	5	6,4	8

2.1.1 Calculate the numerical values of A, B and C. (5)

2.1.2 (a)Write down a formula which can be used to determine the number of workers and months needed to complete the stadium. (2)

2.1.3 Draw a graph which illustrates the data in the table. (3)

Marking scheme	Rescoring details
Item 2.1.3 3M plotting points, 1M joining points, 1CA overall shape	$1,2,3 \rightarrow 1$ $4,5 \rightarrow 2$

Item 2.2

2.2 FIFA awards 3 points for a **W**in, 1 point for a **D**raw and no point for a **L**ose. The log table shown in the table is based on the following results

Match 1: Spain 3, South Africa 3
Match 2: South Korea 2, USA 2
Match 3: South Africa 1, USA 1
Match 4: Spain 4, South Korea 2

TABLE 3: Log table.				
COUNTRY	WIN	LOSE	DRAW	POINTS
Spain	1	0	1	**4**
South Africa	0	0	2	D
USA	0	0	2	2
South Korea	0	1	1	E

2.2.1 Spain, South Africa, South Korea and USA are in the same group. Using the FIFA awards system, calculate the numerical values.
(a) D   (2)   (b) E   (1)

2.2.2 Some soccer fanatics are proposing a new format of awarding points, based on the following:

Proposed Format:

Win (by a margin of 1 goal (e.g 1-0, or 2-1, or 3-2)	= 3 points
Win (by a margin of 2 or more goals (e.g. 2-0 or 4-1)	= 3+ 1 bonus point
Draw of less than 2 goals (0-0 or 1-1)	= 1 point
Draw of 2 or more goals	= 1+1 bonus point
Lose by a margin less than 2 goals	= 1 bonus point
Lose by 2 or more goals	= no points

(a) Using the new format, how many points will a team have if it has won one match (4-2) and drew another match (2-2)? (3)

(b) Determine the values of D and E in Table 3 if the points awarded were based on the Soccer fanatic's new proposed scoring method (4)

2.2.3 The last round robin matches will be Spain against USA, and South Africa against South Korea.

(a) Using the FIFA scoring format, is it going to be possible for South Korea to overtake Spain? Give a reason for your answer. (3)

Marking scheme		Rescoring details
Item 2.2.1 2.2.1 (a) D = 2 √√ 2.2.1 (b) E = 1 √	2 = A 1 = A	2.2.1a 1,2→1 2.2.1b left as is
Item 2.2.2 2.2.2 (a) 3√ + 1 √ + 1 + 1 = 6 √ 2.2.2 (b) (1) D = 1 + 1 + 1√ = 3 √ (2) E = 1 + 1√ = 2√	1 = M 2 = A 2 = CA 2 = CA	2.2.2a 1,2→1 3→2 2.2.2b(1) 1,2→1 2.2.2b(2) 1,2→1
Item 2.2.3 a No√, because according to the FIFA format, a team gets 3 points for a win√ so it will end up with 4 points.√	1 = No 2 = R	1→1 2,3→2

Item 3.2

Extraneous details of context omitted…

3.2   As part of his duties, Sipho has to collect 10 cows from the veld and milk them before releasing them at 08:00. He must use a bucket with a round base of radius 10 cm and a height of 25 cm for each cow. It takes him 15 min to milk each cow and pour the milk into a bigger container in the farmer's house.

3.2.1 At what time must Sipho start milking the cows to finish milking all the cows before 08:00?   (3)

3.2.2 Determine the maximum capacity of the bucket (in litres) used by Sipho if

    1 m³ = 1000 litres and **Volume of cylinder= $\pi \times r^2 \times h$**   (4)

3.2.3 Sipho decides to reduce the time spent walking to carry the milk to the farmer's house by first pouring milk from each cow into a larger second bucket. The dimensions of the second bucket are double the dimensions of the first bucket.

    (a) Using this second bucket, do you think Sipho will double the amount of milk he takes to the farmer's house per trip?   (1)

    (b) Prove your answer in (a) by calculating the ratio of the volume of the new bucket to the volume of the old bucket.   (5)

Marking scheme		Rescoring details
Item 3.2.1 Time need = 10 × 15 minutes = 150 minutes √ At 05:30 √√	1 = M 2 = A	1→1 2,3→2

Item 3.2.2
Capacity of bucket = $\pi \times r^2 \times h$
              = $\pi \times (10\ cm)^2 \times 25$ cm
              = 7853.98 cm² √ √
 	
    Litres of milk: 1 m³ = 1 000 litres √
    0,00785 m³ = litres
    Litres = 0,00785 × 1 000
           = 7,85 litres √
	1 = M
	1 = vol. of bucket
	
	1 = using scale
	
	1 = A
		1→1
		2,3→2
		4→3

Item 3.2.3 3.2.3(a) No √	1 = A	no change
Item 3.2.3(b) Capacity of new bucket = $\pi \times r^2 \times h$ = $\pi \times (20\ cm)^2 \times 50$ cm √√ = 62 831,85 √ Ratio volume of new bucket : volume of old bucket 62 831,85 : 7853,98 √ = 8: 1 √ Therefore, new capacity is more than double the old bucket	2 = doubling dimensions 1 = vol of bucket 1 = M 1 = C	1→1 2,3→2 4,5→3

Item 4.1

Extraneous details of context omitted

4.1   The graph in […] shows the percentiles in babies' growth chart.

    4.1.1 What is the normal weight of the baby boy at birth?   (2)

    4.1.2 Determine the ranges of baby boys' masses at birth.   (2)

    4.1.3 What will be the mass of a healthy 2 year-old girl who weighs 2.5 kg at birth?   (2)

    4.1.4 Peter weighed at 50th percentile at birth. When he was 3 years, he weighed 13 kg as indicated.

        (a) What does 50th percentile mean?   (2)

        (b) Is Peter's weight normal as compared to other 3-year-old boys? Give a reason for your answer.   (3)
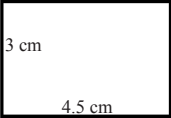
        (c) Provide any TWO causes for this change in Peter's mass from 33 months to 36 months.   (4)

Marking scheme		Rescoring details
Item 4.1.1 3.2 kg (accept any answer between 3 kg and 3,5 kg) √√	2 = R	1,2→1
Item 4.1.2 Range = 4,5 kg − 2,2 kg √ =2,3 kg√	1 = M 1 = CA	1,2→1
Item 4.1.3 9,5 kg √√	2 = R	1,2→1
Item 4.1.4 (a) It means half of boys normally weigh that particular kg at that particular age√√	2 = R	1,2→1
Item 4.1.4(b) No√, since it is mentioned in the card that if the graph is that direction, there is something wrong.√√	1 = No 2 = R	1,2 →1 3→2
Item 4.1.4(c) He was sick and did not like food √√ He started walking and as a result, lost some weight √√ 2 × 2 = 4 His mother gone back to work and does not have mothers milk OR any other valid reason	2 *marks per valid point*	1,2→1 3, 4→2

Item 5

5.1 Your brother came to you to ask for assistance in identifying the shop which has a better price.

    5.1.1 Calculate the discount offered by Blue Gum. (2)

    5.1.2 If the discount offered by Casanova is R324,75, calculate the original price of the DVD/VCR combo player.

    5.1.3 Determine the price of the Blue Gum CD player excluding VAT before the discount is calculated. (3)

    5.1.4 If your brother had the money to buy a CD player, at which store would he purchase the CD player? Give a reason for your answer. (3)

5.3 If the size of the front face of the CD player is 45cm × 30cm, use the following scale to **draw** the space taken by its picture in the advert of Doggy Sounds. **Scale 1:10** (3)

Marking scheme		Rescoring details
Item 5.1.1 Discount $\quad = \frac{19}{100}$ of R1 299 √ $\qquad\qquad = $ R246,81 √	1 = MA 1 = A	1,2→1
Item 5.1.2 Original price $\;= 4 \times$ discount √ $\qquad\qquad\quad = 4 \times$ R324,75 √ $\qquad\qquad\quad = $ R1 299,00 √	1 = M 1 = A 1 = CA	1,2,3→1
Item 5.1.3 Price before VAT $= 100/114$√ $\times$ R1 299 √ $\qquad\qquad\qquad = $ R1 139,47√	2 = M 1 = A	0,1→0 2→1 3→2
Item 5.1.4 He must buy from Casanova. √ Casanova charges = 3 × R324,75 = R974,25 √ which is the lowest price. √ OR any suitable calculation	1 = A 2 = R	1,2→1 3→2
Item 5.3 3 cm 4.5 cm 1 = exactly 3 cm breadth 1 = exactly 4.5 cm length 1 = right angles		1,2→1 3→2

Appendix 3 starts on the next page →

APPENDIX 3

Category frequencies for each of the items

Item	Frequencies of scores in each category per item						Initial number of thresholds
	0	1	2	3	4	5	
1.1.	23	2	48				2
1.1.2	43	0	30				2
1.1.3	20	20	11	7	15		4
1.2.1a	18	0	55				2
1.2.1b	14	34	6	0	19		4
1.2.2	26	3	4	1	39		4
1.3.1	27	2	6	5	3	30	5
1.3.2	29	9	11	24			3
1.3.3a	52	6	15				2
1.3.3b	3	33	2	35			3
2.1.1	36	2	12	3	5	15	5
2.1.2a	63	4	6				2
2.1.2b	63	3	0	7			3
2.1.3	5	0	4	15	11	38	5
2.1.4a	41	0	32				2
2.1.4b	34	0	39				2
2.1.5	33	2	0	38			3
2.2.1a	4	0	69				2
2.2.1b	11	62					1
2.2.2a	18	4	21	30			3
2.2.2b(1)	45	1	27				2
2.2.2b(2)	36	1	36				2
2.2.3a	22	39	3	9			3
2.2.3b	28	3	29	1	12		4
3.1.1	7	2	64				2
3.1.2	27	2	44				2
3.1.3	18	21	11	23			3
3.2.1	25	10	1	37			3
3.2.2	9	24	32	1	7		4
3.2.3a	64	9					1
3.2.3b	44	8	13	2	1	5	5
3.3.1	26	1	46				2
3.3.2	33	9	4	6	21		4
3.3.3	23	0	24	0	26		4
4.1.1	46	0	27				2
4.1.2	70	0	3				2
4.1.3	69	0	4				2
4.1.4a	48	8	17				2
4.1.4b	18	11	3	41			3
4.1.4c	29	0	16	0	28		4
4.1.5	11	5	0	57			3
4.2.1	43	11	7	11	1	0	5
4.2.2	66	6	1				2
4.2.3	5	0	68				2
4.2.4	32	0	23	1	17		4
5.1.1	16	0	57				2
5.1.2	48	0	1	24			3
5.1.3	71	0	2	0			3
5.1.4	54	0	6	13			3
5.2	47	0	24	1	0	1	5
5.3	38	7	4	24			3
Total number of thresholds							150

Item	Frequencies of scores in each category per item				Number of thresholds after rescoring
	0	1	2	3	
1.1.	25	38	48		1
1.1.2	43	30			1
1.1.3	20	20	18	15	4
1.2.1a	18	55			1
1.2.1b	14	40	19		2
1.2.2	26	7	40		2
1.3.1	27	16	30		2
1.3.2	29	9	35		2
1.3.3a	52	21			1
1.3.3b	36	37			2
2.1.1	36	22	15		2
2.1.2a	63	10			1
2.1.2b	63	10			1
2.1.3	5	19	49		2
2.1.4a	41	32			1
2.1.4b	34	39			1
2.1.5	35	38			1
2.2.1a	4	69			1
2.2.1b	11	62			1
2.2.2a	18	25	30		2
2.2.2.b(1)	45	28			1
2.2.2.b(2)	36	37			1
2.2.3a	22	39	12		2
2.2.3b	28	32	13		2
3.1.1	7	66			1
3.1.2	27	46			1
3.1.3	18	32	23		2
3.2.1	25	10	38		2
3.2.2	9	24	33	7	3
3.2.3a	64	9			1
3.2.3b	44	23	6		2
3.3.1	27	46			2
3.3.2	33	19	21		2
3.3.3	23	24	26		2
4.1.1	Deleted				
4.1.2	70	3			1
4.1.3	69	4			1
4.1.4a	56	17			1
4.1.4b	29	44			1
4.1.4c	45	28			1
4.1.5	16	57			1
4.2.1	43	29	1		2
4.2.2	66	7			1
4.2.3	5	68			1
4.2.4	32	23	18		2
5.1.1	16	57			1
5.1.2	48	25			1
5.1.3	71	2			1
5.1.4	60	13			1
5.2	47	25	1		2
5.3	38	11	24		2
Total number of thresholds					75