Rasch Analysis of South Africa's Grade 6 Annual National Assessment

Charlotte Modzuka[a], **Caroline Long**[b], **and France Machaba**[c*]

[a] *University of Pretoria, South Africa*
[b] *University of Johannesburg, South Africa*
[c] *University of South Africa (UNISA), Pretoria, South Africa*

**Corresponding author. Email: emachamf@unisa.ac.za*

Abstract

The aim of this study was to investigate the quality of the Annual National Assessment (ANA) Grade 6 Mathematics assessment instrument, administered in the year 2012 guided by the main question: to what extent does the 2012 Grade 6 Mathematics assessment instrument provide meaningful and consistent information for making appropriate interpretations? The responses to 29 questions, comprising 56 items, from 546 learners' scripts were coded individually, recaptured and analysed. The analysis shows that the Annual National Assessment was not well targeted, that is, the mean of the item and the mean of the person measures were not approximately equivalent. Specifically, in relation to this cohort, and taking into account the many factors that influence performance, we recommend that for reliable inferences to be made from test results, the quality of the assessment instrument should first be established.

Keywords: *Annual National Assessment; Rasch Measurement Theory; Mathematics*

Introduction

The principle that test results alone are not enough to inform policy for improving classroom practice and the quality of teaching and learning is well recognised: test results should be considered in relation to other factors. We would like to allege that the notion of assessment for learning and as a means of learning has been distorted by the focus on external assessment (Kanjee & Moloi, 2014). The broad purpose underpinning systemic assessment is to improve classroom practice. Preceding the introduction of the Annual National Assessment (ANA), South Africa participated in international and regional assessments. There was an urgent need to establish the quality of the education system in the country, particularly given that there was no systemic overview of the educational quality of the schooling system available since 1994 (Howie, 2012).

In an attempt to improve teaching and learning in South African schools, the Department of Education introduced the Foundations For Learning campaign in 2008 (Department of Basic Education, 2011). The ANA emerged from this campaign. The Department of Education required Grades 3, 6 and 9 to be assessed in the key subjects, English and Mathematics. The research study on which this article is based investigated the ANA 2012 Grade 6 Mathematics. The Grade 6 ANA indicated low performance at national level (27.0% pass), at Gauteng provincial level (30.9% pass) and for the particular Gauteng district (27.2% pass). There are number of factors that might have contributed to this low performance. These factors might be associated with the teaching and learning of Mathematics in the prior grades, and also with access to resources and the management of assessment practices (Modzuka, 2017). Moreover, there is a gap between school-based assessment and the ANA results as noted in Bansilal's (2017) work showing that learners perform better in school-based assessment than in ANA. There may be different reasons for this poor performance in the ANA.

Our concern specific to this study is to investigate the quality of the ANA assessment instrument as a possible factor contributing to poor performance in the ANA. Whereas several international systemic assessments have been conducted, and the Western Cape has had a province-wide systemic assessment, in 2015 the teaching unions objected to ANA because they claimed that the information provided by the test did not assist the teaching and learning (Pretoria News, 2015, p. 2). Furthermore, several studies (Graven & Venkat, 2014; Graven, 2016; Kanjee & Moloi, 2014; Spaull & Kotze, 2015) have queried the testing procedures in the administration of ANA as they cause undue stress and are time consuming. However, these studies have not, so far, explored the discriminating (and thus diagnostic) quality of the ANA items. A new model of systemic assessment is being planned by the Department of Basic Education. Hence, the purpose of this article is to explore the assessment instrument itself as one of the possible contributing factors to poor performance.

Providing an international perspective, Matters (2009) indicates two major variables that contribute to learners' achievement, particularly in systemic tests such as the ANA. She argues that achievement is influenced by factors internal to students as well as those imposed by features of the assessment environment. Factors that are associated with a student, according to Matters (2009), are motivation, test anxiety and academic self-concept. Factors related to the assessment environment include the assessment instrument itself, the preparation involved with the assessment process and the conditions under which the assessment is applied.

While all of the above factors are critical to understanding the results of systemic-type testing and for the making of policy inferences, the research on which this article is based focussed on only one factor, which is the *assessment instrument*. In addition to the qualitative analysis based on professional judgement, the Rasch model (Smith, 2004) was applied for the purpose of investigating the functioning of the ANA assessment instrument. The following three sub-questions guided this process : (a) are the items on the Mathematics assessment instrument functioning as expected—specifically do the items fit the requirements of the Rasch model; (b) how well are the items distributed along the continuum of the variable; and (c) how well is the assessment instrument targeted to the abilities of the learners?

The Rasch analysis was conducted on a sample of scripts from a particular Gauteng district. Thus in this paper we argue that, taking into account the many factors that impact on performance, inferences from results of systemic tests generally should be considered in the light of the design and quality of the assessment instrument. The Rasch analysis is applied to verify the validity of the instrument within this particular frame of reference.

The Conceptual Framework

The Queensland Studies Authorities' (2009) Assessment Policy document was used as a base for developing a model for quality systemic assessment (see Figure 1) and for addressing the research questions for this study. These questions seek to explore the validity and reliability of the ANA instrument influencing learner attainment in Mathematics. The conceptual framework in Figure 1 guided this study in identifying aspects that were required in the quality and design of an assessment instrument, specifically related to an analysis of the instrument. In the conceptual framework, validity centres on the extent to which meaningful and appropriate conclusions are made based on the assessment instrument used in this study. Messick (1989) viewed the concept of validity as an 'overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores' (p. 10). The interpretations of assessment scores are 'valid to the extent that these interpretations are supported by appropriate evidence' (Scheerens, Glass & Thomas, 2003, p. 97). One facet of validity is the 'extent to which the content of the test is representative of the curriculum or construct that is being measured' (Anderson & Morgan, 2009, p. 16). In the conceptual framework, the 2012 ANA Grade 6 Mathematics assessment instrument

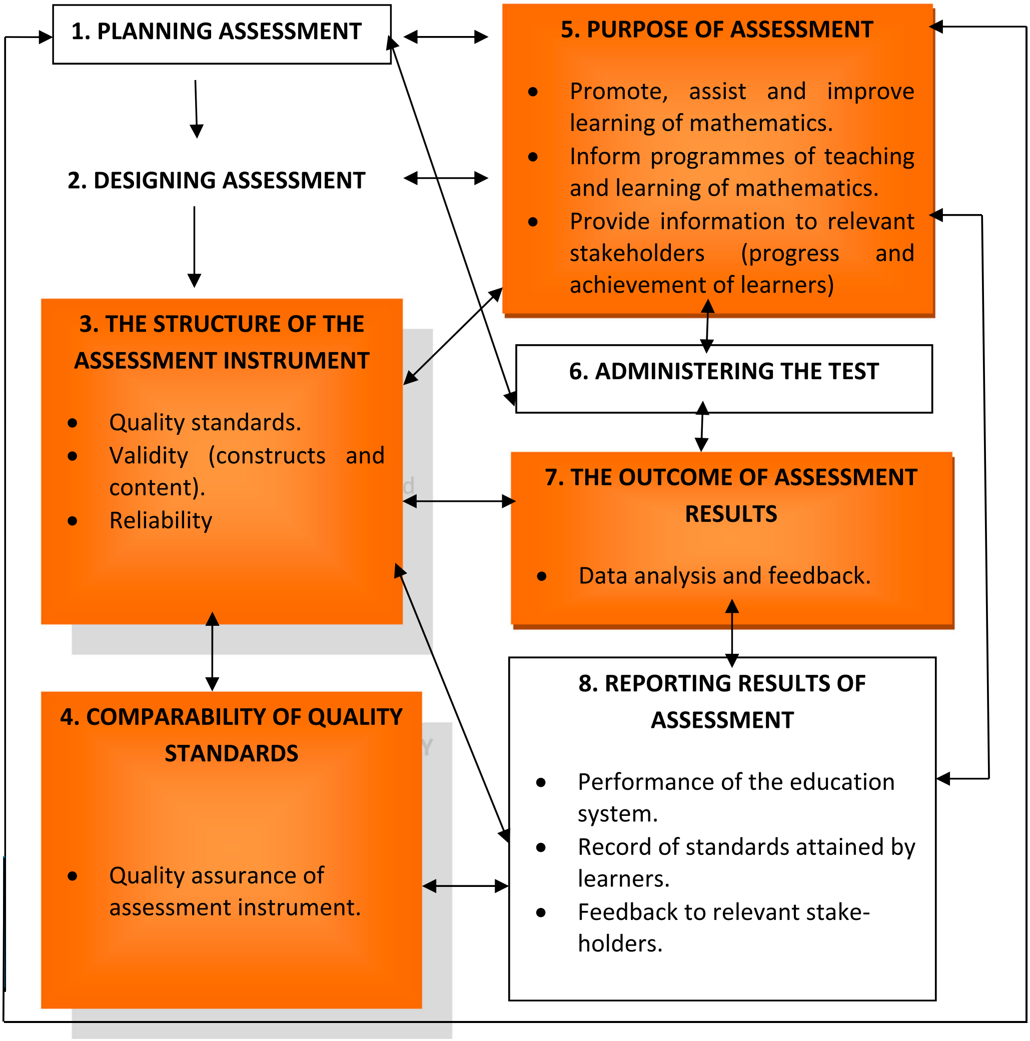

1. PLANNING ASSESSMENT
2. DESIGNING ASSESSMENT
3. THE STRUCTURE OF THE ASSESSMENT INSTRUMENT
Quality standards.
Validity (constructs and content).
Reliability
4. COMPARABILITY OF QUALITY STANDARDS
Quality assurance of assessment instrument.
5. PURPOSE OF ASSESSMENT
Promote, assist and improve learning of mathematics.
Inform programmes of teaching and learning of mathematics.
Provide information to relevant stakeholders (progress and achievement of learners)
6. ADMINISTERING THE TEST
7. THE OUTCOME OF ASSESSMENT RESULTS
Data analysis and feedback.
8. REPORTING RESULTS OF ASSESSMENT
Performance of the education system.
Record of standards attained by learners.
Feedback to relevant stake-holders.

Figure 1. A model for quality systemic assessment (adapted from Queensland Studies Authorities, 2009).

was analysed to check whether it measured what it was intended to measure. The 'construct validity of a national assessment test is checked by making sure that the test is measuring the construct it is supposed to measure, irrespective of the test format' (Scheerens et al., 2003, p. 20). Construct validity is based on the 'integration of any evidence that bears on the interpretation or meaning of the test scores' (Messick, 1989, p. 8). In contrast, reliability depends on the 'quality of test items, the test itself, the way the tests were administered, the characteristics of the group of learners (such as the effort they make while taking the national assessment tests), and the quality of scoring of the test items' (Anderson & Morgan, 2009, p. 76). Reliability is the extent to which a test measures consistently; it is 'an indicator of the consistency of the test results' (Anderson & Morgan, 2009, p. 74).

The model for quality systemic assessment (see Figure 1) has eight components. These are: planning of the assessment; the design of the assessment; the structure of the assessment instrument; comparability of quality standards; the purpose of assessment; administration of the test; the outcome of the assessment results; and reporting of the results of the assessment. These eight components are linked within the process of assessment and constitute the quality assurance of assessment instrument. However, the data gathered in the study focused on three components—the purpose of assessment, the structure of the assessment instrument and comparability of quality standards—because these components are related to the concepts of validity and reliability, central to this article.

Designing assessment involves making a decision on the form of assessment and the type of task needed for a particular content. The first major consideration in assessment design is how much assessment is needed and how much time should be spent on it (Anderson & Morgan, 2009). For this component, we looked at whether the assessment instrument was reasonable, and whether it catered equitably to learner difference. For an assessment to be fair, it should be unbiased and provide all learners with equal opportunities to demonstrate achievement (Meyer et al., 2010). The instrument should also be time-efficient and manageable.

With regard to the structure of the assessment instrument, the assessment instrument should be of an acceptable standard, meaning that the quality standards, validity, reliability, clarity of instructions and written language should be taken into consideration. The style and context of the questions used on the assessment instrument should be appealing to learners. The assessment instrument should be valid and reliable, cater for different cognitive styles and varied skills, and include a variety of questions in order to cater for these differences. The test items collectively should be a representation of the curriculum and the cognitive domains, and be language appropriate and relevant for learners (Kellaghan, Grenaney & Murray, 2009). The instructions and the language of the items should be clear, simple and appropriate to the level of development of learners. A Mathematics test should not contain so much language that the learner's performance depends on the language ability to read rather than on the Mathematics ability (Kellaghan et al., 2009). The assessment instrument should be a true representation of the curriculum and scope expected. This will ensure that learners have a fair opportunity to exhibit their knowledge and to complete the assessment task successfully. In order to ensure successful assessment, the assessment instrument should be balanced, comprehensive and varied (Dreyer, 2013).

In the education system emphasis is on quality assurance of instruments, which is necessary for monitoring the education system. Monitoring can be achieved by comparing the achievement of learners' performance with the national indicators of learner performance (Du Plessis et al., 2007). One factor central to quality assurance is the comparability of standards. In order for assessment results to monitor change over time, the instruments must be comparable (Kellaghan et al., 2009). The ideal situation would be that the same test is used every round and that it is kept secure between administrations (Anderson & Morgan, 2009).

Research Design and Methodology

The secondary analysis research design allowed the researcher to investigate the previously administered ANA Grade 6 Mathematics instrument. In order to gain an understanding of the low performance by learners in the year 2012, this study focused on the quality and the design of the ANA instrument and as such applied the Rasch measurement model. A statistical item analysis applying the Rasch model was utilised. The strong point concerning this design is that it allowed individual item analysis of the 2012 ANA Grade 6 Mathematics instrument. The design focused on the core principles provided by the Rasch measurement model to explore the quality of the instrument.

Data Collection and Research Procedures

The test responses from all (546) Grade 6 learners in one district of Gauteng Province were coded and recaptured into an Excel spreadsheet; the ANA test had 29 questions, some with several sub-questions. Question 1 of the ANA comprised six sub-questions which were captured as single items, which gave a total of six items. Fifty-two items were captured. The data was imported to the RUMM 2020 programme for Rasch measurement model analysis. The programme allowed a detailed investigation from different perspectives. The following outputs were provided:

- item statistics (the location, residual, chi-square and probability);
- person–item location distribution (the overview of the ANA test as a whole);
- person–item map (item difficulty and person ability).

The Rasch analysis aligns item difficulty and person proficiency on the same scale. This analysis provided insight into the functioning of each item and further addressed the three sub-research questions. This process was checked for any inconsistencies in entering the data.

Rasch Measurement Theory

The Rasch model was developed in the 1950s by George Rasch with the purpose of solving this educational dilemma posed by the Danish government, namely that of measuring the reading progress of learners over a number of years and using different assessments. His solution to this dilemma resulted in the alignment of all learners and all tests on the same scale through a procedure explained in his book (Rasch, 1980). This alignment allowed the assessment of learner proficiency across the entire cohort and provided information to the authorities.

Basic Requirements of the Rasch Model

The Rasch measurement model has basic principles that need to be adhered to in the process of analysing an assessment instrument. For the purpose of this article, we investigate whether the three requirements for the Rasch measurement model have been met. These are item fit, person–item distribution and item difficulty.

Item Fit

Rasch analysis provides fit statistics designed to aid the investigator in making a number of interrelated decisions about the assessment instrument data (Smith, 2004). This requirement implies that an ideal item function is calculated and the actual item performance is then compared with this ideal model to estimate how effectively the item is functioning at that particular level of difficulty. Significant disagreement between the model and the data should be a cause for concern, as this is an indication that the item is not performing as expected. These items are referred to as misfit items. Misfit items generally compromise test performance. Two types of misfit are over-discrimination and under-discrimination.

Over-discrimination

Over-discrimination occurs when learners with higher ability levels get the item right more often than the model predicts, but learners at lower ability levels get the item wrong more often than the model predicts. Items that have large negative residual values and a relatively steep curve indicate that the discrimination ability is too high. Although very high discrimination ability appears to be a desirable trait for an item to have, it could be an indication that the item may be disadvantaging low-ability learners unnecessarily. In other words, ability level alone is not enough to account for the differential performance of the item.

Under-discrimination

Under-discrimination occurs when learners with higher ability levels get a question wrong more often than the model predicts, but learners at the lower ability levels get it right more often than the model predicts. Under-discrimination is problematic since it implies that, as the ability of learner increases, the probability of gaining a higher score does not increase proportionally. Items that exhibit under-discrimination might have been guessed or possibly a construct is being tested that does not fit well into the overall framework because learner ability only partially accounts for item performance. Under-discrimination items are usually indicated by large positive residual values and a relatively flat curve.

Person–item distribution

The person–item map represents both item difficulty and person proficiency on the same scale. Likewise, on the left of the person–item map (see Figure 2 as an example), learners' ability is estimated in relation to item difficulty, with learners of greater ability or proficiency at the top of the scale and learners with lesser proficiency at the bottom of the scale.

Item Difficulty

In the Rasch measurement model, ‘performances are attributed relative importance in proportion to the position they hold on the measurement continuum’ (Bond & Fox, 2007, p. 120). The learner ability location is defined as the point at which learners have a 0.5 probability (50% chance) of responding correctly to all the items. The Rasch measurement model provides ‘indices that help the investigator determine whether there are enough items spread along the continuum, as opposed to clumps of them, and enough spread of ability (more difficult and easier) among persons’ (Bond & Fox, 2007, p. 40).

Smith (2004) describes the Rasch measurement model’s purpose as ‘providing a direct estimate of the modelled variance for each estimate of person’s ability and an item’s difficulty providing a quantification of the precision of every person measure and item difficulty which can be used to describe the range within which item’s true difficulty or person’s ability falls, that is, both person reliability and item reliability’ (p. 96).

Findings

In this article, we report the research findings following the sequence of the research questions. As mentioned earlier, three broad research questions guided this research:

- Are the items on the Mathematics instrument functioning as expected—specifically do the items fit the requirements of the Rasch measurement model?
- How well are the items distributed along the continuum of the variable?
- How well is the assessment instrument targeted to the abilities of the learners?

Each question will be regarded as a theme, emerging theme or category and will be discussed separately (but not in isolation) to the other questions.

The summary statistics from the Rasch analysis (Table 1), including item and person means and standard deviations, point to the appropriateness of the test for this cohort. We note that the item mean is set at zero. The difficulty of each of the items is then calibrated in relation to the mean. The person mean for this sample is −1.550 logits, which means that this test is too difficult to ascertain

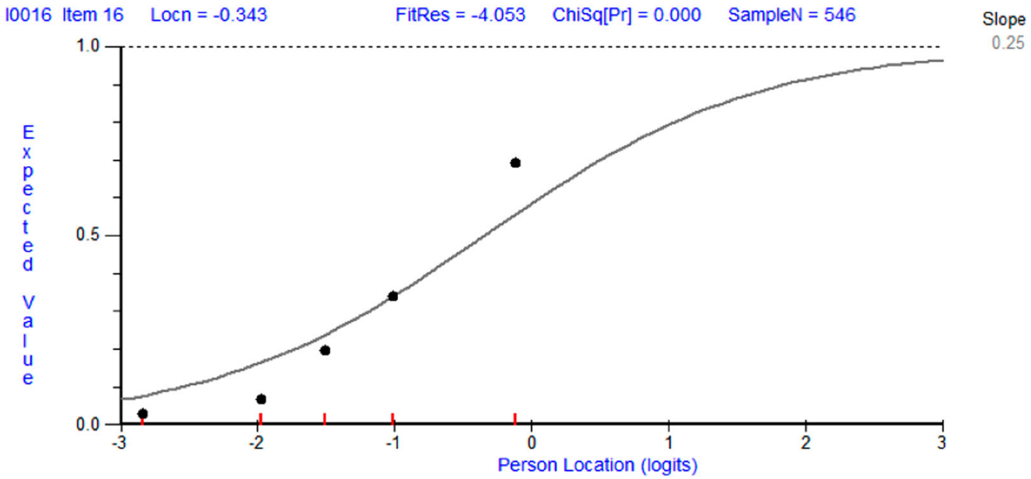


Figure 2. Item 16: over-discrimination for particular Gauteng District.

Table 1. The person separation index

	Item	Person
Mean	0.000	–1.550
Standard deviation	1.752	0.976
Person separation index	0.857	

an accurate measure of learner ability. The standard deviation of the items (1.752) shows the spread to be large. However the learner spread is closer together (0.976).

In Rasch measurement theory the person separation index (PSI) provides a measure of reliability, indicating the robustness of the test. The PSI, specific to the Rasch model, contrasts the variance among the proficiency estimates of the learner cohort as a whole relative to the error variance within each person (Bond & Fox, 2007). It provides a measure of internal consistency by providing an indicator of the separation of persons relative to the difficulty of the item. The equivalent in traditional test theory is the Kuder–Richardson 20, or Cronbach's α, which provides a measure of the internal consistency of the items, but does not provide a measure of person consistency relative to items (Bond & Fox, 2007). The PSI (0.857) shows the test to have good spread and therefore to be providing meaningful information (see Table 1). However this claim is mitigated when considering other factors such as targeting.

Are the Items on the Mathematics Instrument Functioning as Expected—Specifically do the Items Fit the Requirements of the Rasch Measurement Model?

The internal coherence of the ANA Grade 6 Mathematics assessment instrument was investigated by checking the alignment of item difficulty and the person proficiency. This alignment of person and item enabled an investigation of the instrument itself (see Table 1) and each of the items (in Table 2). Firstly, the person–item distribution was presented and discussed earlier. In the following section over-discriminating and under-discriminating items are discussed. Table 2, for example, presents five items—items 16, 20, 19, 29 and 50—that are listed as over-discriminating on the 2012 ANA Grade 6 Mathematics.

Over-discriminating Items

In this sub-section an example, i.e. item 16, of over-discriminating items is given in Table 3.

Figure 2 presents Item 16, an example of an over-discriminating item and its description. This statistic indicates the item is experienced very differently by high-proficiency and low-proficiency learners.

For item 16 learners at ability levels <–1.0 logits perform significantly poorer than the model predicts, while learners with ability levels >0 logits perform significantly higher. This item tests a difficult concept. Table 4 present five items listed as under-discriminating in the sample taken in this study.

Table 2. Over-discriminating items

Sequence	Item	Type	Location	SE	Residual	d.f.	χ^2	d.f.	Probability
16	10016	Poly	–0.343	0.104	–4.053	524.63	32.708	8	0.000069
20	10020	Poly	1.065	0.151	–3.283	530.49	24.425	8	0.001946
19	10019	Poly	0.38	0.122	–3.25	531.47	20.803	8	0.00769
29	10029	Poly	–0.221	0.106	–3.004	529.51	29.013	8	0.000315
50	10050	Poly	0.379	0.07	–2.982	530.49	16.713	8	0.033243

SE, Standard error; d.f., degrees of freedom.[T/S: footnote to table]

Table 3. Over-discriminating item and reasons

Item	Item on ANA Grade 6 Mathematics	Possible reasons
Item 16	Round off 29 702 to the nearest 5.	The instruction was clear. Learners with lower ability level were not competent with the skill of rounding off to the nearest 5.

Table 4. Under-discriminating items

Sequence	Item	Type	Location	SE	Residual	d.f.	χ^2	d.f.	Probability
43	10043	Poly	−1.469	0.059	6.595	508.02	82.415	8	0
28	10028	Poly	−1.606	0.094	5.113	530.49	45.147	8	0
24	10024	Poly	−0.976	0.097	4.369	517.79	60.671	8	0
6	10006	MC	−1.073	0.098	4.24	497.27	36.493	8	0.000014
7	10007	MC	0.114	0.117	3.601	493.36	55.244	8	0

Under-discriminating items

The under-discriminating items fail to adequately distinguish between high performers and low performers. Item 24 was identified as an under-discriminating item, see Table 5.

Under-discrimination

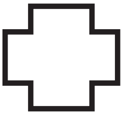
Item 24 was about lines of symmetry. This item might not be contributing in a valuable way to the test owing to higher-ability learners having a significantly lower probability of responding correctly to the item than the model predicts. Table 5 indicates the most under-discriminating item (item 24) in the test and its explanation.

Figure 3 indicates an example of an under-discriminating item label, item 24. Item 24 does not differentiate between learners of any ability level and even displays a trend of negative discrimination, where higher ability levels are associated with a decreased probability of responding correctly to the item.

How Well are the Items Distributed along the Continuum of the Variable?

This section is concerned with the reliability of the 2012 ANA Grade 6 Mathematics instrument as far as the distribution of items along the continuum was concerned. This question dealt with the overview of person–item proficiency and item difficulty. The person–item map in Figure 4 represents on the same scale both item difficulty and person proficiency. The item mean is set at 0. The difficulty of items is calibrated in relation to the mean, which is set at zero. Therefore, the item difficulty estimates are expressed in logits, in which a logit of 0 is arbitrarily set as the average, or mean, of the item difficulty estimates. Item difficulty is then calibrated in relation to the mean, with items of greater difficulty at the top of the scale, and items of lesser difficulty at the lower end. Learner ability is then estimated in

Table 5. Under-discriminating item labels and reasons

Item	Item on ANA Grade 6 Mathematics	Possible reasons
Item 24	*How many lines of symmetry does the diagram below have?* __________ 	The reason for low discrimination could be that this topic was not addressed in the curriculum; the high-ability learners were not familiar with the term ‘symmetry’, and neither were the low performers. Moreover, some learners may have arrived at the correct answer by guessing

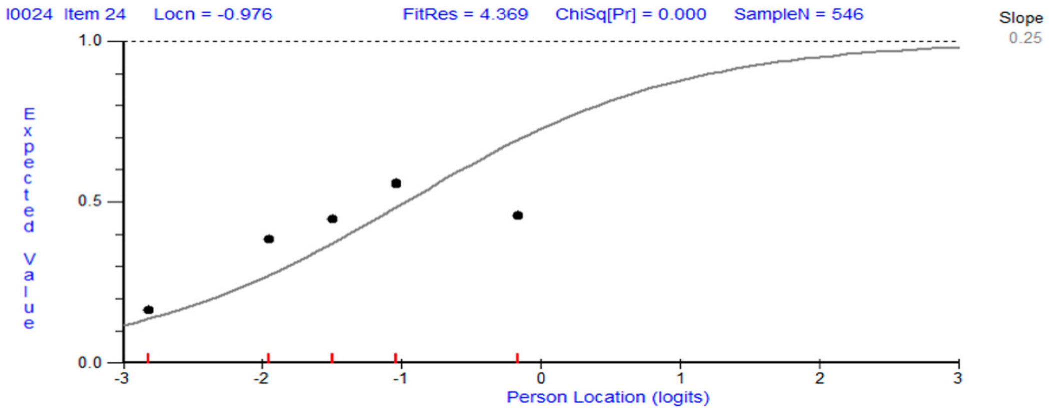


Figure 3. Item 24: under-discrimination.

relation to item difficulty. The learner ability location is defined as the point at which learners have a 0.5 probability (50% chance) of responding correctly to the item.

The person–item map in Figure 4 shows that there are no learners with ability levels high enough to justify an item as difficult as item 45 (located above 4 logits). Conversely, items 33 and 34 appear to be too easy, since the majority of learners have ability estimates above −3 logits, while these two items are both located below −3 logits. Items 51, 7 and 8 are close to the mean at zero. Items with a negative logit indicate that the item was relatively easy, for example items 34, 33 and 2 have negative logit scores. However, items with a positive logit are relatively more difficult; items 45, 38 and 36 have positive logit estimates.

How Well is the Assessment Instrument Targeted to the Abilities of the Learners?

Table 6 indicates that items 33 (−3.682 logits) and 34 (−3.385 logits) were the easiest items in the test, that is, these items were located at the extreme (<−3 logits). The next easiest item in the ANA test was item 2 (−2.166 logits), with item difficulty close to −3 logits. This table indicates that even learners with relatively low ability levels (<−3.5 logits) have a 0.5 probability (50% chance) of getting the easiest items correct. Table 7 shows one of the easiest items and the reasons for learner performance, item 2.

Table 8 presents the items 45, 38, 36, 37, 42, 17, 39, 46, 32 and 44 on the 2012 ANA Grade 6 Mathematics which were identified as the most difficult items. In Table 8 it can be seen that item 45 (4.843 logits), item 38 (3.77 logits) and item 36 (3.174 logits) were the extremely difficult items in the ANA test. Items at the extreme >3 logits will only be answered correctly by learners with very high mathematical abilities. Table 9 indicates an example of a difficult item, item label 45.

Conclusions and Discussions

Overview of the ANA

The broad and overarching purpose of the ANA is ‘to make a notable contribution towards better learning in schools, by serving broadly as a systemic tool and second, as a diagnostic tool in identifying areas of strength and weakness in teaching and learning’ (Department of Basic Education, 2013, p. 36). The ANA instrument appeared to have met its desired purposes to some extent. The investigation did not entirely confirm that the ANA Grade 6 Mathematics instrument met its purpose of assessment.

Within the model of quality systemic assessment which is the conceptual framework for this study, it is clearly indicated that the link between the stages of assessment must be taken into consideration in order to produce a quality assessment instrument that could be administered for different purposes.

```
LOCATION          PERSONS     ITEMS [locations]
----------------------------------------------------------------------------------------------
  6.0                      |
                           |
                           | I0045
                           |
                           |
  5.0                      |
                           |
                           |
                           |
                           |
  4.0                      |
                           |
                           | I0038
                           |
                           |
  3.0                      | I0036
                           |
                           |
                           | I0037
                           | I0017  I0042
  2.0                      |
                           |
                           | I0039
                           | I0049  I0032  I0046
                           | I0047  I0044
  1.0                    × |
                           | I0023  I0003  I0021  I0020
                         × |
                       ××× | I0022  I0009  I0011  I0052
                        ×× | I0031  I0019  I0050
  0.0                  ××× | I0008  I0007  I0051
                   ××××××× |
                     ××××× | I0035  I0013  I0029
                  ×××××××× | I0016
                ×××××××××× | I0004  I0001  I0014
 -1.0            ××××××××× | I0005  I0027  I0048
              ×××××××××××× | I0012  I0040  I0018  I0006  I0041  I0024
         ××××××××××××××××× |
          ×××××××××××××××× | I0043
        ×××××××××××××××××× | I0010  I0028
 -2.0    ××××××××××××××××× | I0015  I0030
                ×××××××××× | I0025  I0026
               ××××××××××× | I0002
               ××××××××××× |
                  ×××××××× |
 -3.0                   ×× |
                      ×××× |
                       ××× |
                       ××× | I0034
                         × | I0033
 -4.0                      |
                        ×× |
                           |
                           |
                           |             × = 3 Persons

 -5.0                      |
                           |
                           |
                           |
                           |
 -6.0                      |
----------------------------------------------------------------------------------------------
```

Figure 4. The Grade 6 person–item map for particular Gauteng District sample.

Reflection on the Conceptual Framework

The data gathered in the study was tested against the indicators in the framework. The criteria for a good assessment design are explicitly stipulated in the conceptual framework. Based on the research results, the ANA Grade 6 Mathematics instrument did not fully meet the criteria for a good assessment

Table 6. Ten easiest items on the ANA Grade 6 Mathematics

Sequence	Item	Type	Location	SE	Residual	d.f.	χ^2	d.f.	Probability
Display: 10 easiest items									
33	10033	Poly	–3.682	0.134	–0.55	529.51	15.405	8	0.051736
34	10036	Poly	–3.385	0.123	–0.228	531.47	33.133	8	0.000058
2	10003	MC	–2.166	0.098	2.397	525.60	17.826	8	0.022569
26	10026	Poly	–1.954	0.095	–0.561	532.44	8.22	8	0.412236
25	10025	Poly	–1.952	0.095	–1.819	532.44	16.692	8	0.033478
15	10015	Poly	–1.93	0.095	1.245	532.44	23.914	8	0.002371
30	10030	Poly	–1.926	0.095	–2.605	531.49	20.124	8	0.009876
10	10010	Poly	–1.669	0.094	–2.695	531.47	21.265	8	0.006477
28	10028	Poly	–1.606	0.094	5.113	531.49	45.147	8	0
43	10043	Poly	–1.469	0.059	6.595	508.02	82.415	8	0

Table 7. Easy item and reasons

Item	Item on ANA Grade 6 Mathematics	Possible reasons
Item 2	39 569 *was rounded off to* 40 000. *To which of the following numbers was it rounded off?* *A 5* *B 10* *C 100* *D 1 000*	Instruction clearly presented. Learners competent at rounding off numbers.

Table 8. Ten most difficult on the ANA Grade 6 Mathematics

Sequence	Item	Type	Location	SE	Residual	d.f.	χ^2	d.f.	Probability
Display 10 most difficult items									
45	10045	Poly	4.843	0.47	0.006	471.87	4.373	8	0.82203
38	10038	Poly	3.77	0.483	–1.205	528.53	6.826	8	0.555502
36	10036	Poly	3.174	0.363	0.051	531.47	9.143	8	0.330353
37	10037	Poly	2.595	0.279	–0.62	529.51	5.336	8	0.721145
42	10042	Poly	2.372	0.334	–0.082	515.83	6.762	8	0.562508
17	10017	Poly	2.298	0.244	–0.821	528.53	21.376	8	0.006214
39	10039	Poly	1.77	0.197	–1.434	523.65	21.323	8	0.006339
46	10046	Poly	1.668	0.219	–0.586	530.49	1.309	8	0.995446
32	10032	Poly	1.642	0.188	0.389	515.83	11.911	8	0.155243
44	10044	Poly	1.418	0.162	–0.503	521.7	5.77	8	0.672147

Table 9. Difficult item and reasons

Item	Item on ANA Grade 6 Mathematics	Possible reasons
Item 45	*Phiti has five numbered cards. How many different two-digit numbers can she make with these cards?* $\boxed{6}$ $\boxed{2}$ $\boxed{0}$ $\boxed{7}$ $\boxed{3}$	Learners may have found the numbers written on cards confusing. The topic of computations and permutations, a subset of the probability topic, may not have been covered in class.

design. When reflecting on the indicators of the conceptual framework we infer that a systemic assessment that is not designed properly will have a negative influence. With regard to the structure of the assessment instrument, the results show that the construct validity was achieved to some extent; content validity was fully achieved; and reliability was low on the ANA Grade 6 Mathematics instrument. Therefore, the ANA was not designed properly.

Moreover, the conceptual framework indicates that, in order to ensure a quality standard assessment instrument, the structure should be balanced, comprehensive and varied. Achievement and progress on assessment are recognised through the comparability of quality standards. The assessment instruments must be comparable in order to monitor learners' performance against the national indicators. The design of the existing ANA programme over successive years, although comparable in terms of the content tested on a similar model to the South African matric examination, does not have common items across years.

The Functioning of the Items on the Mathematics Assessment Instrument and the Requirements of the Rasch Measurement Model

In the ANA Grade 6 Mathematics 2012, there are items that were considered misfitting items, some items were categorised as under-discriminating and others were categorised as over-discriminating items. According to Matters (2009) a common question asked by teachers when examining aggregated data from standardised tests is: 'What made this multiple-choice item so difficult that only a small proportion of learners chose the correct answer?' (Matters, 2009, p. 214). She points out that some of the questions below might be useful in formulating an answer, even though these questions are composed for item formats other than multiple choice, i.e. for constructed response and open response formats:

> What kind of thinking was involved: concrete, conceptual or personal?; What abilities were required: verbal, numerical or spatial?; What emphases were placed on the treatment of the stimulus material: Did the learner need to absorb it, operate on it or transform it into something new?; Is it possible that a learner's perception of success on the item was influenced by features of the stimulus material such as context? (Matters, 2009, p. 215)

The factors referred to above might have affected learners' performance. Therefore, it is important for the assessor to answer those questions when setting a test or developing an assessment.

Are the Items Distributed Well Along the Continuum of the Variable?

The ANA Grade 6 Mathematics test was very difficult for the Gauteng District learners, hence the low performance, and therefore it does not have a good diagnostic ability. The ANA instrument seemed to have a reasonable person separation index, a measure of both item and person reliability, as it appeared to have items that were reasonably well distributed along the continuum of the variable, although with gaps at both the higher and the lower ends. In the ANA Grade 6 Mathematics 2012 it was found that items 51, 7 and 8 were close to the mean. Items 34, 33 and 2 were easy. However, items 45, 38 and 36 were difficult. Person ability is estimated in relation to the item difficulty estimates, for example, the more negative the value, the lower the learner's ability on this assessment, and the more positive the value, the higher the ability of learners (Bond & Fox, 2007). This conclusion casts doubt on the reliability of the 2012 ANA Grade 6 Mathematics as a systemic assessment instrument.

The Abilities of the Learners Targeted by the Assessment Instrument

If an item is too easy or too difficult, that particular item will compromise the test's ability to differentiate between low, moderate and high ability learners owing to its poor discrimination properties. If the overwhelming majority of students get a question right—or wrong—that particular item can be considered insensitive to learner ability, something that is obviously undesirable in a test. In the 2012 ANA Grade 6 Mathematics for the Gauteng District sample, the findings revealed that there were items that fell significantly beyond learners' abilities in terms of difficulty. Therefore the 2012 ANA Grade 6 instrument was considered to be limited in its ability to accurately measure those learners' abilities.

The learners' characteristics can have an effect on attainment. Learners from Gauteng District come from different backgrounds, are often of different abilities, go to different schools and have different levels of test preparedness. They also have had experience with different kinds of test items—'hard' compared with 'easy'; 'open' compared with 'closed'—as well as differences in their sources and levels of extrinsic and intrinsic motivation. Ability is not the same thing as achievement: 'Ability is the first of the learners' background characteristics of interest when looking for explanations of patterns, trends and relationships in data' (Matters, 2009, p. 213). Furthermore, Bond and Fox (2007) point out that a unique strength of the Rasch model is its requirement that the outcome of any interaction between person and item be solely determined by just two parameters, the ability of the person and the difficulty of the item. This requirement establishes a strong framework against which to assess the ANA data for the presence of anomalous behaviour that may influence the estimation of the item and person parameters (Bond & Fox, 2007).

Conclusion

The findings of the investigation indicated that the 2012 ANA Grade 6 Mathematics was not entirely appropriate but had room for improvement, including with respect to the overall reliability. The Rasch analysis proved to be useful in identifying problematic features of the assessment instrument, a strategy the Department of Basic Education will be able to incorporate. Therefore, the results can be used to establish the development of a theoretically based, empirically tested instrument to measure mathematical knowledge and skills of Grade 6 learners through its submission to the Rasch model. In addition, the data gathered from the instrument can be used to provide meaningful information to refine the ANA Grade 6 Mathematics instrument in terms of achieving its validity and high reliability to inform district-level interventions.

Disclosure Statement

No potential conflict of interest was reported by the authors.

REFERENCES

Anderson, P., & Morgan, G. (2009). *Developing tests and questionnaires for National Assessments of Educational Achievement*, Vol. 2. Washington, DC: World Bank.

Bansilal, S. (2017). The difficulty level of a national assessment of Grade 9 mathematics: The case of five schools. *South African Journal of Childhood Education*, *7*(1), a412.

Bond, T., & Fox, R. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Department of Basic Education. (2011). *Annual National Assessment Report of 2011*. Pretoria: Government Printing Works.

Department of Basic Education. (2013). *Report on the Annual National Assessment of 2013.* (2013). Retrieved from http://www.education.gov.za

Dreyer, J.M. (2013). *The Educator as Assessor.* Pretoria: Van Schaick Publishers.

Du Plessis, P., Conley, L. & Du Plessis, E. (2007). *Teaching and Learning in South African Schools.* Pretoria: Van Schaik.

Graven, M. (2016). When systemic interventions get in the way of localized mathematics reform. *For the Learning of Mathematics*, *36*(1), 8–13.

Graven, M., & Venkat, H. (2014). Primary teachers' experiences relating to the administration processes of high-stakes testing: The case of Mathematics Annual National Assessments. *African Journal of Research in Mathematics, Science and Technology Education*, *18*(3), 299–310.

Howie, S. (2012). *High Stakes Testing in South Africa: Friend or Foe? Assessment in Education Principles: Policy and Practice* (pp. 81–89). Oxford: Routledge.

Kanjee, A., & Moloi, Q. (2014). South African teachers' use of national assessment data. *South African Journal of Childhood Education*, *4*(2), 90–113.

Kellaghan, T., Grenaney, V. & Murray, S. (2009). *Using the results of a National Assessment of Educational Achievement*. New York: Worldbank.

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* (pp. 209–225). Dordrecht: Springer.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Meyer, L., Lombard, K., Warnich, P., and Wolhuter, C. (2010). *Outcome-Based Assessment for South African Teachers.* Van Schaik Publishers. Pretoria.

Modzuka, C.M (2017). *An investigation of the 2012 Annual National Assessment Grade 6 mathematics instrument*. Unpublished MEd thesis, University of Pretoria, South Africa.

Pretoria News (2015). *Teacher union queries differing results in school tests and ANAs*. Pretoria News, 17 October, p. 2.

Queensland Studies Authorities (2009). *P–12 Assessment Policy document*. Retrieved from https://www.qcaa.qld.edu.au

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

Scheerens, J., Glas, C., & Thomas, S.M. (2003). *Educational evaluation, assessment, and monitoring: A Systemic Approach*. London: Taylor & Francis.

Smith, E.V. (2004). Evidence for reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E.V. Smith & R.M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 93–122). Maple Grove, MN: JAM Press.

Spaull, N. & Kotze, J. (2015) Starting behind and staying behind in South Africa. *International Journal of Educational Development*, *41*, 13–24.