Art. #1935, 13 pages, https://doi.org/10.15700/saje.v41n4a1935

The level of difficulty and discrimination power of the items of the National Senior Certificate Mathematics Examination

Nombuso P. Zondo, Temesgen Zewotir and Delia E. North
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa
zondon@ukzn.ac.za

South Africa's National Senior Certificate examination system was introduced in 2008 as a single national examination system, in order to facilitate fair and standardised assessment and to provide all learners with an equal chance of access to higher education. However, limited research has been done to investigate the discrimination power of the actual examination items and the spread difficulty level for learners from different school quintile types. The purpose of the study reported on here was to investigate differential performance of learners in the items of the 2009 National Senior Certificate mathematics examination. The dataset used in this study was from the Western Cape (WC) Education Department. From the analysis, the results show that the discrimination power of the different examination questions was not identical across different school quintiles. Further investigation of the data reflects a considerable range of category difficulty levels, with higher (above average) ability levels being tested for learners in the quintile 1 to quintile 4 schools, while only learners with average abilities were being tested in the quintile 5 and independent schools.

Keywords: baseline study; grade 12; Graded Response Model; Item Response Theory; latent trait; school quintile type

Introduction

Standardised school leaving examination is used as the conventional criterion and signal achievements of secondary school learners to universities and employers (Bishop, 1999; Brumwell, Deller & MacFarlane, 2017). The purpose of standardisation is to develop a benchmark for the setting of assessment and to ensure that learners with equal ability levels obtain equivalent results (Loock & Grobler, 2004). With the growth and increasing interdependence of the global economy, the national systems of standardised examinations tend to follow international and conventional standardised examination practices (see Wagner, Lockheed, Mullis, Martin, Kanjee, Gove & Dowd, 2012 and references therein). Accordingly, South Africa introduced a single national qualification examination in 2008, referred to as the National Senior Certificate (NSC), which serves as the key entry point to national higher education institutions and the working environment (Department of Basic Education [DBE], Republic of South Africa [RSA], 2021; Mahlobo, 2015; Sasman, 2011).

In South Africa, the school system is divided into public schools and independent (private) schools, where public schools receive government funding, while independent schools do not receive government funding. The public schools are subdivided into five national quintile (NQ) categories: NQ1, NQ2, NQ3, NQ4 and NQ5. The quintile of each school is assigned, based on its socioeconomic status, which is measured by the rates of income, unemployment and illiteracy within the school's catchment area. Schools in the poorest communities are classified as NQ1, with those in the wealthiest communities being classified as NQ5 (Ally & McLaren, 2016; Dass & Rinquest, 2017). A progressive school-fee system is applied in the public schools, which imposes a lower/free school-fee on low-income earning community schools, compared to those schools in higher income earning communities.

Our objective with this study was to investigate the level of difficulty and discrimination power of the mathematics examination questions (referred to as "items") for the learners from socioeconomic diversified schools. We trust that this study provides useful insights into the challenges of the fairness of the national examination items for assessing learners' ability and achievement, from both the lower and higher socioeconomic background community schools.

Consequently, if the results indicate that the items did not produce a fair discrimination and a tolerable difficulty level for diversified socioeconomic school learners, it provokes discussion and further research among policy makers, educationists and researchers on the standardisation process of adjusting aggregated raw marks by pre-specified criteria. Thereby, learners from lower socioeconomic backgrounds will have fair access to universities and employability, based on the sole national qualification criterion.

Literature Review

In the ongoing effort to improve learner performance in school, one of the aspects that still needs attention is narrowing the achievement gap for learners in the different socioeconomic school types. While the socioeconomic status of the school is not the only factor associated with learner performance, it remains one of the most important factors (Banerjee, 2016; Clotfelter, Ladd & Vigdor, 2007; Devine, Fawcett, Szűcs & Dowker, 2012; Osborne, 2001; Spaull & Kotze, 2015; Van der Berg & Louw, 2007). A study by Carnoy and Chisholm (2008) echoed that in the lower quintile schools, learner marks in mathematics were low, regardless of the teacher knowledge. Four years later, Reddy, Prinsloo, Arends, Visser, Winnaar, Feza, Rogers, Janse van Rensburg, Juan, Mthethwa, Ngema and Maja (2012), using the Trends in International Mathematics and Science Study (TIMSS) data, revealed that

learners in NQ1 and NQ2 schools performed at similar levels in mathematics, but lower than those of the relatively better resourced NQ3, NQ4 and NQ5 schools. By reviewing a number of studies, Spaull (2013) noted that the performance of learners from schools that fall in the wealthiest 20 to 25%, was much better than the larger education system, which catered for 70 to 80% of the learners in public schools. Similar findings were later reported in a study by Spaull and Kotze (2015), who found that the learning gap between the poorest 60% and the wealthiest 20% of learners in South Africa, was approximately four grade levels for learners at Grade 9 level.

Indeed, such disparities are not unique to South Africa; they are of concern on a global scale. In a study on learner performance in mathematics, focusing on the lower grade levels in the United States, Burnett and Farkas (2009) reported that the poverty status of learners, contributed to the reduced performance. These findings were supported by those of Maliki, Ngban and Ibu (2009) who reported that learners from independent schools in one of the states in Nigeria were better performers when compared to those from the public schools. On the contrary, 5 years earlier, Hanushek (2004) reported that schools in the United States with access to better resources did not necessarily produce better performances in the national examination. Similar findings were reported across the years by Birch and Miller (2007) and Smith and Naylor (2001) in Australia and by McNabb, Pal and Sloane (2002) in the United Kingdom (UK) who found that learners from non-government schools performed poorer and had lower ability levels in university later on than those from public government schools. In summary, these reviews highlight the importance of considering the effect of socioeconomic status on learner performance in the school national examination.

Grade 12 assessment in South African schooling

Since 1994, South Africa's highest schooling level exit examination system (Grade 12) developed from independent administration by different education departments, to a (single) national examination (DBE, RSA, 2012). The assessment of learners at this level is made up of the school-based assessment and a formal standardised external examination. The school-based assessment makes up 25% of the overall mark, while the formal external examination carries 75% of the final NSC mark for the different subjects areas of learning (Umalusi, 2013). The internal programme of assessment includes informal and formal assessments. The informal internal assessment informs the educators on the learners' progress towards achieving the assessments standards, while the formal internal assessment tools allow the teacher to differentiate between learners on a performance scale (Wedekind, 2013). However, the reliability of school-based assessment has been questioned in past research (Shapiro, Keller, Lutz, Santoro & Hintze, 2006; Stillman, 2001; Yong & Sam, 2008). The most common factors preventing the wider acceptance of school-based assessment, involve the problem of comparing performances of different learners from different schools and student assessment on a fair basis by the teachers (Biggs & Collis, 1989; Yung & Yung, 2001). Globally, the use of standardised external assessment is a common practice (Demski & Racherbäumer, 2017; García, 2003; Jacobs, Mhakure, Fray, Holtman & Julie, 2014; and references therein). In line with such international practices, in our study we focused on the performance of learners in the formal standardised external examination.

Aim, Justification and Problem Statement

The NSC mathematics examination is structured with a focus on the learners' ability to answer an examination item by solving problems using known knowledge, routine, non-routine and complex procedures. Past studies and reports often make use of average scores and overall pass rates as a measure of performance in mathematics (DBE, RSA, 2018; Hunt, Ntuli, Rankin, Schöer & Sebastiao, 2011; Makgato & Mji, 2006; Sasman, 2011). However, such scores and rates do not measure the underlying latent constructs, or latent variables, which are thought to influence the observable indicators (Embretson & Reise, 2000), that is, determining if the examination was adequately designed to measure the intended latent traits of the learners (Geremew, 2014). It is, therefore, useful to assess how well the mathematics examination items measure the learners latent construct (called ability level). It is also essential to study whether learners from the different school quintile types, differ in terms of mathematics ability level of achievement. Accordingly, with this study we strove to answer the following questions:

1) Are the NSC mathematics examination items equally discriminating for the learners from the different school quintile types?
2) Are the NSC mathematics examination items equally difficult for the learners from the different school quintile types?
3) Is the level of difficulty and discrimination power of the NSC mathematics examination items identical for the low and high ability learners?

To address these research enquiries, we used the 2009 NSC mathematics examination results data from the WC Education Department. Unlike any other province in South Africa, the WC Education Department captured every 2009 examinee score, for each mathematics NSC examination item. Besides this data being key to shedding light on the three questions above, the data structure unique to this study can also serve as a baseline study for such

data capturing for future NSC item analysis for other provinces, and, therefore, also nationally.

Theoretical Framework

In this research, we sought to explore the level of difficulty and the discrimination power of the NSC mathematics examination items for the learners from different school quintile types. Such item analysis assists in estimating how learners of different ability levels would score in a specific item (Boopathiraj & Chellamani, 2013). Because the learner scores are influenced by the characteristics of the learner and by the characteristics of the item (Birnbaum, 1968; De Boeck & Wilson, 2004; Rasch, 1960), learners within the same classroom or same school, taught the same mathematical concepts, may not develop the same understanding of the content. Item Response Theory (IRT) provides procedures for obtaining information on learners and assessment items. Therefore, within the ambits of IRT, there are parameters describing the learners and those describing the examination items (see Lazarsfeld & Henry, 1968; Lord, 1980; Rasch, 1960).

Item difficulty

According to the most basic form of IRT, the likelihood that the learner will respond correctly to a particular test item is affected by two things: the learner's ability and the item's difficulty. Therefore, the probability that a particular learner will respond correctly to a given test item depends on his/her mathematical ability and the level of difficulty of the test item.

Item discrimination

Just as the test items differ in terms of their level of difficulty, they might also differ in terms of the degree to which they can differentiate between learners with high ability levels, compared to learners who have low ability levels. This item characteristic is called item discrimination.

Methodology

The data in this study consisted of a record of 19,927 Grade 12 NSC mathematics examinees from 389 schools from six different school quintile types, namely, NQ1, NQ2, NQ3, NQ4, NQ5 and independent schools. The NSC mathematics examination items were structured to correspond to one of six subject topics of the mathematics paper 1 curriculum indicated in Table 1. The exam targets learner ability to answer the items by solving problems using known knowledge, routine and non-routine procedures, as well as complex procedures. The NSC mathematics examination consists of 13 items, each of different weighting. The data were presented as the raw learner mark captured for each item at the individual learner level.

As a preparation of the item analysis we recoded the raw captured marks into polytomously ordered scoring categories, R_i . The polytomous ordered scoring was performed by taking the different item weights into consideration. In order to ensure that the polytomously scored items were of uniform range, the data were rescaled, as shown in Table 2, where, for example, a learner who answered correctly less than 10% of item i would have been recorded to score in category 1, while a learner who answered correctly 10 to less than 20% of item i, would have scored in category 2. The rescaling is a standardisation approach in line with the guideline of Embretson and Reise (2000).

Table 1 NSC mathematics 2009 items by topic in the school curriculum

Item	Item topic
1	Item1: Algebra and equations
2	Item 2: Patterns and sequences
3	Item 3: Patterns and sequences
4	Item 4: Patterns and sequences
5	Item 5: Patterns and sequences
6	Item 6: Functions and graphs
7	Item 7: Functions and graphs
8	Item 8: Functions and graphs
9	Item 9: Annuities and equations
10	Item 10: Calculus
11	Item 11: Calculus
12	Item 12: Calculus
13	Item 13: Linear programming

Table 2 Scoring of the different categories

Graded score category	Exam item scoring range
1	0–10
2	10–20
3	20–30
4	30–40
5	40–50
6	50–60
7	60–70
8	70–80
9	80–90
10	90–100

Zondo, Zewotir and North (2020) studied the same dataset, using quantile regression and found that school type and the education district where the school was located were significant factors associated with learner performance in mathematics in the Western Cape province. However, the characteristics of the questions that made up the mathematics examination were not examined. Therefore, the IRT approach attempts to measure the latent construct.

IRT (Embretson & Reise, 2000; Subjex, 2019) is commonly used to model the probability of a correct response to an item by relating certain item characteristics to learner characteristics (Hambleton, Swaminathan & Rogers, 1991; Reckase, 2009). That is, IRT rests on the postulate that a learner's score in item i, is influenced by both learner and item characteristics. IRT can be divided into two branches, namely, unidimensional and multidimensional. In brief, unidimensional IRT requires the items (questions) to measure one latent

construct, while multidimensional IRT models describe more than one latent construct (De Boeck & Wilson, 2004). IRT can be used to model both dichotomous and polytomous items. Dichotomous items can be scored as correct/incorrect or true/false and such models are typically used to model items such as those that include the one-parameter logistic model (1PLM), two-parameter logistic model (2PLM) and three-parameter logistic model (3PLM), among others (Hays, Morales & Reise, 2000; Rasch, 1960). Polytomous IRT models accommodate items with more than two response/scoring categories, neither of which can be referred to as correct or incorrect. The polytomous IRT models include the Graded Response Model (GRM) (Samejima, 1969, 2016), Partial Credit Model (PCM) (Masters, 1982) and Nominal Response Model (NRM) (Embretson & Reise, 2000).

The GRM was used in this study since it is appropriate for items with ordered response/scoring categories and items that need not have an equal number of categories. In the GRM, each item i is described by one discrimination (slope) parameter, α_i, and $j = 1, \dots, m_i$ between category threshold parameters, β_{ij}, where $R_i = m_i + 1$ is the total number of scoring categories for item i. Basically, item i is treated as a series of $m_i = R_i - 1$ dichotomies, that is, for an item with $k = 3$ scoring categories, the dichotomies would be $0\ vs.\ 1{,}2$ and $0{,}1\ vs.\ 3$ (2 between threshold categories). The GRM is an extension of the 2PLM (Cagnone & Ricci, 2005; De Boeck & Wilson, 2004; Geremew, 2014; Samejima, 1969) for items with polytomous response categories, one discrimination parameter, and multiple between threshold parameters (Subjex, 2019). One goal of fitting the GRM is to determine the location of the threshold (β_{ij}) parameters on the latent trait scale. These threshold parameters represent the trait level a learner needs to score in or above threshold j with a probability of 0.5. Samejima's (1969) GRM is given by the following equation:

$$P_{ir}(\theta) = P^*_{ir}(\theta) - P^*_{i(r+1)}(\theta)$$

where

$$P^*_{ir}(\theta) = \frac{e^{\alpha_i(\theta - \beta_{ij})}}{1 + e^{\alpha_i(\theta - \beta_{ij})}}$$

are the m_i curves describing the probability of a learner scoring $(r = j = 1, \dots, m_i)$ in or above category threshold j for an item i, conditional on the learners' latent construct.

In the GRM, better performance in an item is represented by higher scores and is dependent on the learners' latent trait θ (in our case ability) and the parameters that characterise the item (α_i and β_{ij}). Both the learner and item character parameters are unknown, while the learner scores for each item in the examination are known. The GRM parameters are then estimated, using the Marginal Maximum Likelihood (MML) or the Joint Maximum Likelihood (JML) estimation technique. In addition to the estimation of the discrimination and threshold parameters, IRT allows for the modelling of the Operating Characteristic Curves (OCCs), Item Information Curves (IIC) and Test Information Functions (TIF).

IRT rests on the assumption of unidimensionality of the latent traits and local independence. Unidimensionality of the latent traits implies that the examination items collectively measure only one latent trait that influences the learner scores, with other factors being treated as random errors (DeMars, 2010). The local independence assumption indicates that if the assumption of unidimensionality holds, a learner's score in one item will be independent of their score in another item. That is, after controlling for the latent trait θ, no relationship will exist between the items. Various methods of assumption testing are proposed in the IRT literature, and the common approaches include: (i) analysis of eigenvalues (ii) factor analysis and (iii) multi-trait multi-method approach, which utilises the inter-item correlation and item-total correlation. None of the methods for verifying the assumptions are discussed in this work, but the process of testing for the assumption is illustrated.

The underlying assumption of rescaling is unidimensionality (Embretson & Reise, 2000). That is, the correlation among the items can be explained by a single latent variable. This assumption can be checked by examining the eigenvalues of standardised values of all items (An & Yung, 2014). From the scree plots it can be noted that the first eigenvalue of the polytomous item correlation matrix is much larger than the others, suggesting that a unidimensional model was sensible for the data. Furthermore, the Cronbach's alpha coefficient of reliability or internal consistency of the latent construct was used. The results show that the items measured the unidimensional latent variable adequately with Cronbach alpha coefficients that were much higher than 0.7.

Results

The 2009 NSC mathematics examination consisted of 13 items, all of which collectively assessed the learners on six topics. The GRM was fitted, which considered the whole distribution range of the latent construct (ability) across school quintile type. Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to assess the GRM fit, with and without school quintile type. The GRM with school quintile type was a better fit, that is, smaller AIC/BIC, indicating that there was a difference in learner ability (latent construct) between the different school quintile types. In evaluating the item parameter estimates, the study attempted to address the research questions.

Discrimination Power of the NSC Mathematics Examination Items for Learners from the Different School Quintile Types

The item discrimination parameters from the results ranged from 0.68 to 3.18, labelled by Baker (2001:34) as "moderate" and "very high" discriminations respectively. The item discrimination estimates are summarised and presented in Table 3 and Figure 1. From Table 3 it can be noted that, in general, all items discriminated the learners in NQ5 and independent schools. Item 12 (last question on calculus) had the lowest discrimination estimate for learners in all the different school quintile types (public and independent schools). Item 11 (second question on calculus), item 1 (algebra and equations) and item 10 (first question on calculus) had the largest discrimination parameter estimates for the learners in NQ1, NQ2, NQ3 and NQ4 schools. Item 11 (second question on calculus), item 6 (first question on functions and graphs) and item 8 (last question on functions and graphs) had the largest discrimination parameter estimates for NQ5 and independent schools.

Table 3 Discrimination parameter estimates for each item

Item	NQ1	NQ2	NQ3	NQ4	NQ5	INDEP
Item 1: Algebra and equations	2.59	2.89	3	2.6	2.89	2.5
Item 2: Patterns and sequences	1.66	1.42	1.69	1.58	2.09	1.98
Item 3: Patterns and sequences	1.66	1.31	1.58	1.59	2.03	1.78
Item 4: Patterns and sequences	2.12	1.94	2.22	1.9	2.29	2.3
Item 5: Patterns and sequences	1.44	1.01	1.38	1.69	2.12	2.28
Item 6: Functions and graphs	2.22	1.91	2.13	1.88	3.06	3.06
Item 7: Functions and graphs	1.13	1.4	1.28	1.35	1.82	1.88
Item 8: Functions and graphs	1.86	1.65	1.69	1.8	2.88	3.07
Item 9: Annuities and equations	1.65	1.58	1.69	1.62	2.48	2.42
Item 10: Calculus	2.77	2.52	2.73	2.27	2.43	1.88
Item 11: Calculus	2.84	3.18	3.16	2.89	3.13	2.7
Item 12: Calculus	0.98	0.68	0.83	0.9	1.77	1.74
Item 13: Linear programming	1.84	2.06	1.99	1.67	2.4	2.25

Visual details on item discrimination for the learners in the different school quintile types are shown in Figure 1. Figure 1 also shows that discrimination estimates for NQ5 and independent schools were similar and grouped together in all items, with the exception of items 1, 3, 10 and 11, which discriminated less for learners from the independent schools, as compared to those from NQ5 schools. Whether the discrimination parameters were identical across the different school quintile types was also examined. According to Cumming (2009), when there is an overlap of less than half the length of one arm, in the confidence intervals of two point estimates, then the point estimates are said to be significantly different from each other with $p = 0.05$. If the two confidence intervals "just touch", then there is a significant effect of $p \approx 0.01$. Furthermore, if there is a no overlap in the confidence intervals of the two point estimates, then the two-tailed *p*-value is less than 0.01, that is, the two-point estimates are statistically significantly different from each other. This is known as "the rule of eye" (Cumming, 2009:206). Figure 2 presents the 95% confidence intervals for the coefficient estimates of the different school quintile types. The extent of the overlap of the confidence intervals was thoroughly assessed, as shown in Figure 2. Accordingly, the intervals for the NQ5 and independent schools revealed an overlap for most items in the mathematics examination. Similar observations were made for the intervals for NQ1 to NQ4 schools.

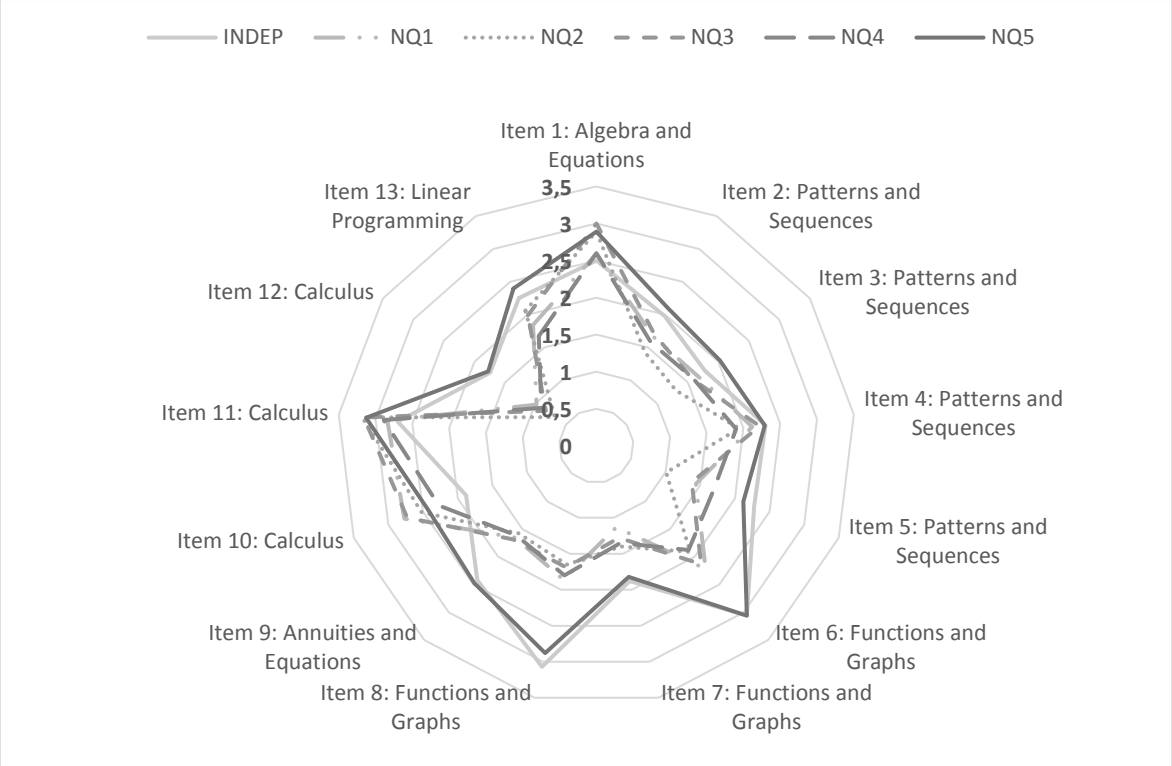

Figure 1 Item discrimination parameter estimates by school quintile type

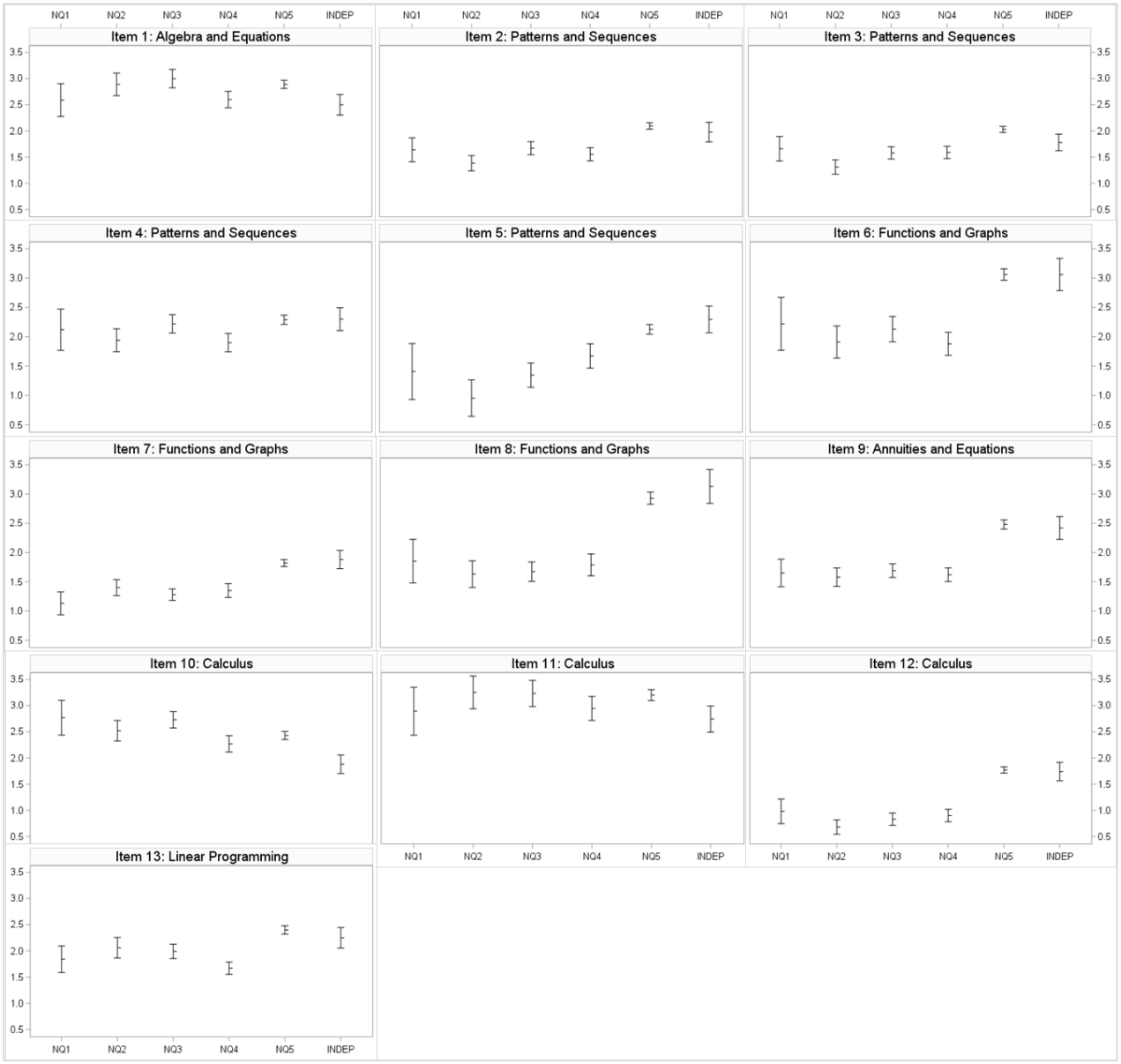

Figure 2 95% confidence limits for the discrimination parameter estimates by different school quintile types

Difficulty Level of the NSC Mathematics Examination Items for the Learners from the Different School Quintile Types

The threshold parameter estimates ranged between -0.91 and 7.4; -0.94 and 10.47; -0.68 and 10.25; -1.24 and 9.34; -1.98 to 3.44; -2.54 to 2.9; and -2.54 to 2.9 for NQ1, NQ2, NQ3, NQ4, NQ5 and independent schools respectively. The NQ2, NQ3 and NQ4 estimates covered the widest range. The values of these threshold parameters can be interpreted as the trait (ability) level a learner needs to score in or above threshold with the probability of 0.5. For item 1 (algebra and equations), the thresholds for all school quintile types were evenly spread in the range of the ability scale θ. For learners who wrote the exam in NQ1 to NQ3 schools, a score in the fifth or sixth category in item 1 was the most probable attainment for learners just above the zero latent trait level. Scoring in the highest two categories for learners from NQ1, NQ2 and NQ3 schools was unlikely, as the threshold parameter estimates were more than two standard deviations above the mean. For learners from NQ5 and independent schools, in item 1, the threshold parameter estimates spanned the negative section of the trait, and a score of eight was more probable for the average learner (zero trait level). Moreover, for the same item, learners from the NQ5 and independent schools were likely to score in the top two categories. The threshold parameter estimates for item 5 (last question on patterns and sequences) were spread to the positive side of the latent scale, that is, higher ability levels were required from learners in all school quintile types in this question. This was a harder item and the average learner was less likely to score in the higher categories. The

thresholds for item 12 (last question on calculus) were further spread to the right of the latent trait scale, therefore, less likely to receive a concentration of high scores, as it demanded the highest ability scores relative to all the other items. It is worth noting that no learner from the NQ1 schools scored in or above category eight, for items 5, 8 and 12.

Is the Level of Difficulty and Discrimination Power of the NSC Mathematics Examination Items Identical for the Low and High Ability Learners?

The discrimination and between threshold parameter estimates are summarised by the OCCs given in Figure 3. For the GRM, the item discrimination and threshold parameter estimates determine the shape and the location of the OCCs. The discrimination parameters dictate the shape (the higher the estimate, the steeper the curves), while the threshold parameters dictate the location of the OCCs (Embretson & Reise, 2000; Sharkness, 2014). For some items, the threshold parameters are not evenly distributed, but are clustered in particular areas of the ability scale. Figure 3 shows that the between-category threshold parameters represent the point of the latent construct scale at which the learners had a 50% chance of scoring in or above a certain category. For instance, for item 1, learners who went to NQ1 and NQ2 schools with an average ability score ($\theta = 0$) had a probability of 0.5 of scoring above category 4 in the question for algebra and equations, while those from NQ5 with the same ability score, had at least a 50% chance of scoring above category 7. For learners from NQ2 schools with a higher ability score of 2, the probability of scoring in or above category 7 was 0.5. For item 11 (second question on calculus), the probability of scoring above category 2 for learners from NQ1, NQ2 and NQ3 schools with an average ability score of 0, was approximately 0.2. This probability increased to more than 0.8 when ability scores were increased to 1, for the same learners in question to score in or above category 2. In contrast, learners from NQ5 and independent schools with an ability level of 1, displayed a probability of 0.4 and 0.6 respectively of scoring in or above category 9 for the same item. In general, the probability of scoring in the lower categories was common for learners from the so called poor government schools. Some items had OCCs clustered in particular areas. Specifically, the thresholds for item 2 (first item on patterns and sequences) were able to differentiate between low and high trait learners, where the OCCs for the item indicate a clear distinction in the clustering of the lower between threshold categories (1, 2, 3, 4 and 5) and the upper between threshold categories (6, 7, 8 and 9), for all school quintile types. Items 5, 6, 7, 8, 9 and 12 required ability scores above zero (average) for learners in the poor government schools. The OCCs for these items were located more to the right of the ability scale. Items 5, 6, 8 and 12 were not able to provide much information about learners from poor government schools with average ability scores. From Figure 3, it is interesting to note that the OCCs shift to the left of the ability scale, moving from NQ1 to NQ5 and independent schools. This means that higher ability levels were needed for leaners in the "poor" government schools (NQ1 to NQ4) to perform well when compared to learners from the NQ5 and independent schools. Furthermore, moving from item 1 (first item) to item 13 (last item), it is interesting to observe the location of the OCCs on the ability scale for the different items. From the Test Information Curves (TICs) presented in Figure 4, it is evident that maximum information gained from NQ1, NQ2 and NQ3 schools, was from learners with ability scores around 2. For NQ4, maximum information was obtained from learners with an ability score of 1, and around 0 for learners who attended the NQ5 and independent school quintile types.

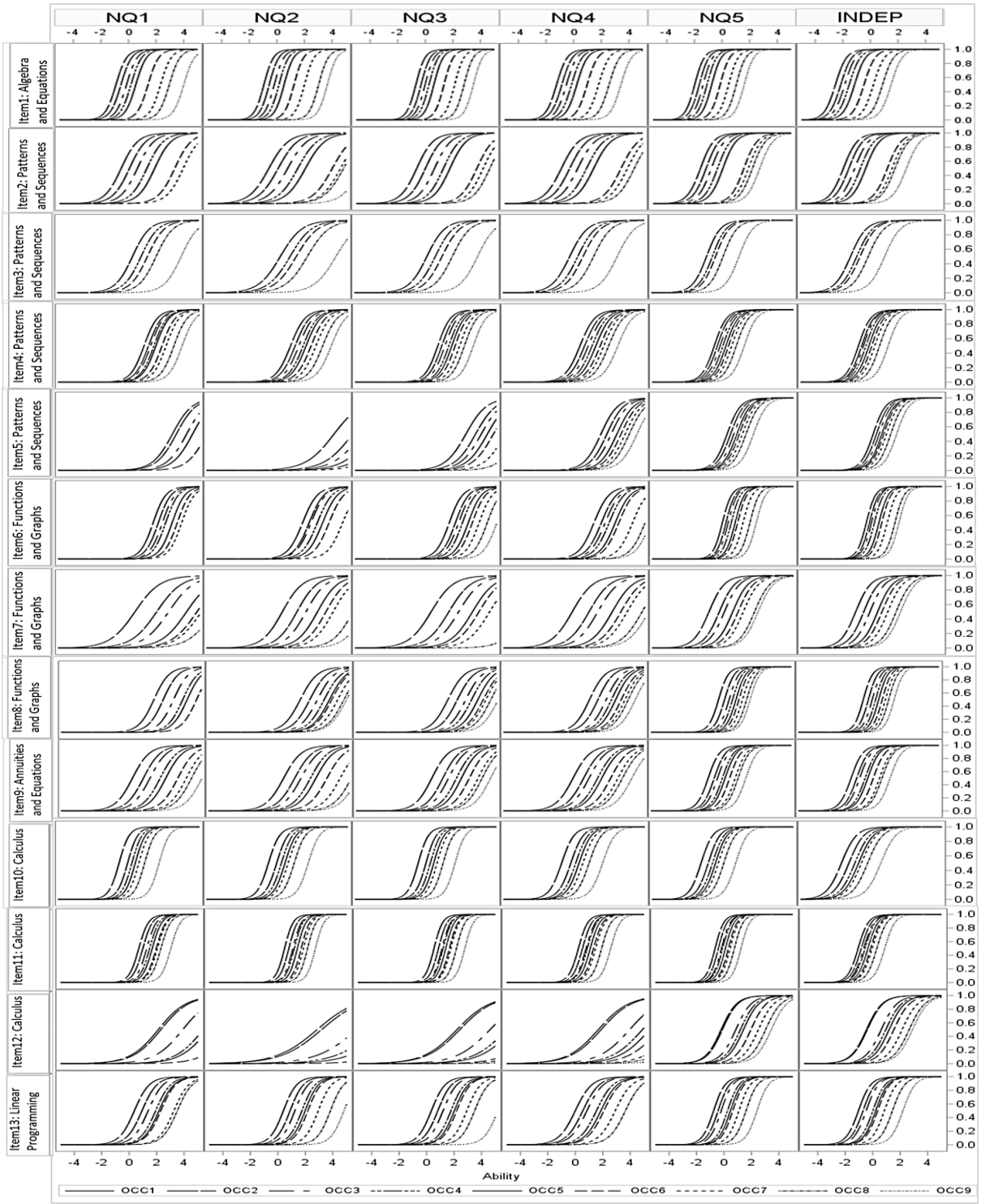

Figure 3 Operating Characteristic Curves (OCCs) for the 13 items of the NSC mathematics 2009 Paper 1 by school quintile type

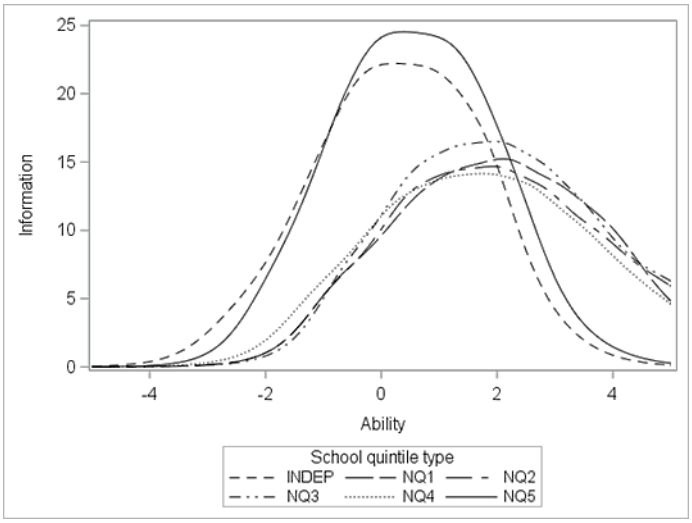

Figure 4 Test Information Curves (TICs) for the different school quintile types

Discussion

Discrimination Power of the NSC Mathematics Examination Items for Learners from the Different School Quintile Types

From our results, it is apparent that the items did not discriminate equally for the learners from the different socioeconomic school quintile types, as some pairwise comparisons were rejected, more specifically, the NQ5 and independent schools versus the NQ1, NQ2 and NQ3 schools. The overall values of the discrimination parameter were highest mostly for NQ5 and independent school quintile types, that is, the items discriminated more for the learners in NQ5 and independents schools.

Difficulty Level of the NSC Mathematics Examination Items for the Learners from the Different School Quintile Types

The difficulty level analysis showed that items 5, 6, 7, 8, 9 and 12 were the most challenging for learners in NQ1, NQ2, NQ3 and NQ4 schools. These items included the last question for patterns and sequences (item 5), all the questions on functions and graphs (items 6, 7 and 8) and the last question on calculus (item 12). Items 5, 6, 8 and 12 gave the least information on learners with average ability from the lower quintile schools, while most items gave much information on the average learner from the NQ5 and independent schools.

Past studies and reports on the NSC mathematics examination often use descriptive measures such as means, pass rates, or linear regression and correlation. Our study has significance for many scholars in developing countries in the context of measuring differences in mathematics performance by analysing the perceived level of item difficulty and item discrimination by learner abilities, for the learners in the different socioeconomic school quintile types. In research, it is rarely the case that the items of interest can be explained solely by the quantification of the item (or student) level measure for some latent construct (Briggs, 2008). We may rather explain group level difficulties among these measures (Briggs, 2008; De Boeck & Wilson, 2004; Mellenbergh, 1994). When no explanatory component is added, researchers should be aware of the effect that this would have on the results of their study.

Using the GRM approach, this analysis allowed us to model the relationship between learner ability score (latent construct) in mathematics and the pattern of responses to the items. The nobility of the approach is in systematically explaining the interrelationship with items explained by one latent variable. The reliability analysis further confirmed that the single latent construct measured the mathematics examination item scores with minimal loss of information. We aimed to investigate the level of difficulty and discrimination power of the mathematics examination items for the learners from socioeconomic diversified schools. The findings show that the extent to which the NSC mathematics examination discriminated between learners with low and high ability levels, varied for learners in the different school quintile types. Furthermore, the results indicate that the perceived level of difficulty varied across school quintile types. The identification of these differences shed light for further research and discussion for policy makers, measurement and evaluation researchers and educators, on how to account the effect of school quintile type on learner performance in the standardised mathematics examination and/or final examination marks.

Conclusion

A limitation of this research was that it relied on the 2009 dataset. However, as this was a year just after the introduction of the NSC, undoubtedly, the study serves as a baseline for how the item construction

and learner abilities change over the years as the NSC becomes a national norm; it shed light on how the first national examination was experienced in the different sectors mentioned. Despite this limitation, this study makes an important contribution to the literature and initiates insight and a national debate on how to evaluate whether the NSC mathematics items equally assess all Grade 12 learners from all school quintile types – it offered a unique opportunity to investigate the performance by topic area, across different quintile public and independent schools. Therefore, it will be helpful in developing effective assessment items and measures (Embretson & Reise, 2000; Fraley, Waller & Brennan, 2000; Kean & Reilly, 2014) for the NSC mathematics examination and potentially for other subject areas. Failure of such interventions will likely lead to a continuation of unequal performance in the final examinations and moreover, unfair access to higher institutions of learning for learners from the different school types, which clearly contradicts the purpose of a standardised single national assessment. In addition, this work further demonstrates the importance and need for detailed datasets for performance monitoring in the NSC examination system countrywide. Results could shed light on further studies that could be done, if more recent such data sets were collected.

Acknowledgements

The authors are grateful to Mr Brian Schreuden, Superintendent General, Western Cape Education Department for permission to use the data. Nombuso P. Zondo would also like to thank the National Research Foundation of South Africa and the University Capacity Development Programme (UCDP) for ongoing financial support.

Notes

i. This article is based on the doctoral thesis of Nombuso P. Zondo (corresponding author), under the research supervision of Professor Temesgen Zewotir and Professor Delia E. North.
ii. Published under a Creative Commons Attribution Licence.
iii. DATES: Received: 6 June 2019; Revised: 13 July 2020; Accepted: 25 September 2020; Published: 30 November 2021.

References

Ally N & McLaren D 2016. Fees are an issue at school too, not just university. *GroundUp*, 17 November. Available at https://www.groundup.org.za/article/fees-are-issue-school-too-not-just-university/. Accessed 31 January 2019.

An X & Yung YF 2014. *Item response theory: What it is and how you can use the IRT procedure to apply it*. Cary, NC: SAS Institute. Available at https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf. Accessed 30 November 2021.

Baker FB 2001. *The basics of item response theory* (2nd ed). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Available at https://files.eric.ed.gov/fulltext/ED458219.pdf. Accessed 30 November 2021.

Banerjee PA 2016. A systematic review of factors linked to poor academic performance of disadvantaged students in science and maths in schools. *Cogent Education*, 3(1):1178441. https://doi.org/10.1080/2331186X.2016.1178441

Biggs J & Collis K 1989. Towards a model of school-based curriculum development and assessment using the SOLO taxonomy. *Australian Journal of Education*, 33(2):151–163. https://doi.org/10.1177/168781408903300205

Birch ER & Miller PW 2007. The influence of type of high school attended on university performance. *Australian Economic Papers*, 46(1):1–17. https://doi.org/10.1111/j.1467-8454.2007.00302.x

Birnbaum AL 1968. Some latent trait models and their use in inferring an examinee's ability. In FM Lord & MR Novick (eds). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bishop JH 1999. Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review*, 6:349–398. Available at https://ecommons.cornell.edu/bitstream/handle/1813/75286/Bishop_16_Are_national_exit.pdf?sequence=1&isAllowed=y. Accessed 30 November 2021.

Boopathiraj C & Chellamani K 2013. Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research*, 2(2):189–193. Available at https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1080.8467&rep=rep1&type=pdf. Accessed 30 November 2021.

Briggs DC 2008. Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, 21(2):89–118. https://doi.org/10.1080/08957340801926086

Brumwell S, Deller F & MacFarlane A 2017. Why measurement matters: The learning outcomes approach–A case study from Canada [Special issue]. *Journal of Higher Education in Africa*, 15(1):5–22.

Burnett K & Farkas G 2009. Poverty and family structure effects on children's mathematics achievement: Estimates from random and fixed effects models. *The Social Science Journal*, 46(2):297–318. https://doi.org/10.1016/j.soscij.2008.12.009

Cagnone S & Ricci R 2005. Student ability assessment based on two IRT models. *Metodološki Zvezki*, 2(2):209–218. Available at http://old.stat-d.si/mz/mz2.1/cagnone.pdf. Accessed 30 November 2021.

Carnoy M & Chisholm L 2008. *Towards understanding student academic performance in South Africa: A pilot study of Grade 6 mathematics lessons in Gauteng province*. Pretoria, South Africa: HSRC. Available at https://repository.hsrc.ac.za/bitstream/handle/20.500.11910/5484/5199.pdf?sequence=1&isAllowed=y. Accessed 30 November 2021.

Clotfelter CT, Ladd HF & Vigdor JL 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6):673–682. https://doi.org/10.1016/j.econedurev.2007.10.002

Cumming G 2009. Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine*, 28(2):205–220. https://doi.org/10.1002/sim.3471

Dass S & Rinquest A 2017. School fees. In F Veriava, A Thom & TF Hodgson (eds). *Basic education rights handbook: Education rights in South Africa*. Johannesburg, South Africa: SECTION27. Available at http://section27.org.za/wp-content/uploads/2017/02/Chapter-7.pdf. Accessed 1 January 2019.

De Boeck P & Wilson M (eds.) 2004. *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

DeMars C 2010. *Item response theory*. Oxford, England: Oxford University Press.

Demski D & Racherbäumer K 2017. What data do practitioners use and why? Evidence from Germany comparing schools in different contexts. *Nordic Journal of Studies in Educational Policy*, 3(1):82–94. https://doi.org/10.1080/20020317.2017.1320934

Department of Basic Education, Republic of South Africa 2012. *National Senior Certificate Examination: Technical Report 2012*. Pretoria: Author. Available at https://www.gov.za/sites/default/files/gcis_document/201409/national-senior-certificate-examination-technical-report-2012re.pdf. Accessed 30 November 2021.

Department of Basic Education, Republic of South Africa 2018. *The 2018 National Senior Certificate results: Schools performance report*. Pretoria: Author. Available at https://www.education.gov.za/Portals/0/Documents/Reports/NSC%202018%20School%20Performance%20Report%20WEB.pdf?ver=2019-01-03-084932-000. Accessed 30 November 2021.

Department of Basic Education, Republic of South Africa 2021. *NSC examinations*. Pretoria: Author. Available at https://www.education.gov.za/Curriculum/NationalSeniorCertificate(NSC)Examinations.aspx. Accessed 30 November 2021.

Devine A, Fawcett K, Szűcs D & Dowker A 2012. Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, 8:33. https://doi.org/10.1186/1744-9081-8-33

Embretson SE & Reise SP 2000. *Item response theory for psychologists*. Mahwah: NJ: Lawrence Erlbaum Associates.

Fraley RC, Waller NG & Brennan KA 2000. An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78(2):350–365. https://doi.org/10.1037/0022-3514.78.2.350

García P 2003. The use of high school exit examinations in four southwestern states. *Bilingual Research Journal*, 27(3):431–450. https://doi.org/10.1080/15235882.2003.10162602

Geremew NM 2014. Applying item response theory models to the Survey of Adult Skills (PIAAC). Master thesis. Stockholm, Sweden: Stockholm University. Available at http://www.statistics.su.se/polopoly_fs/1.178914.1401264507!/menu/standard/file/Master_thesis_Netsanet_Geremew%5B1%5D.pdf. Accessed 4 December 2018.

Hambleton RK, Swaminathan H & Rogers HJ 1991. *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hanushek EA 2004. *Some simple analytics of school quality* (National Bureau of Economic Research [NBER] Working Paper 10229). Cambridge, MA: NBER. Available at https://www.nber.org/system/files/working_papers/w10229/w10229.pdf. Accessed 30 November 2021.

Hays RD, Morales LS & Reise SP 2000. Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38(9 Suppl.), II28–II42. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1815384/pdf/nihms14476.pdf. Accessed 30 November 2021.

Hunt K, Ntuli M, Rankin N, Schöer V & Sebastiao C 2011. Comparability of NSC mathematics scores and former SC mathematics scores: How consistent is the signal across time? *Education as Change*, 15(1):3–16. https://doi.org/10.1080/16823206.2011.574097

Jacobs M, Mhakure D, Fray RL, Holtman L & Julie C 2014. Item difficulty analysis of a high-stakes mathematics examination using Rasch analysis. *Pythagoras*, 35(1):Art. #220, 7 pages. https://doi.org/10.4102/pythagoras.v35i1.220

Kean J & Reilly J 2014. Item response theory. In FM Hammond, JF Malec, TG Nick & RM Buschbacher (eds). *Handbook for clinical research: Design, statistics, and implementation*. New York, NY: Demos Medical Publishing.

Lazarsfeld PF & Henry NW 1968. *Latent structure analysis*. New York, NY: Houghton Mifflin.

Loock CF & Grobler BR 2004. Equating examinations as a prerequisite for ensuring standards in Centralised Senior Certificate (Matric) examinations in South Africa. *Education as Change*, 8(2):74–91. https://doi.org/10.1080/16823200409487092

Lord FM 1980. *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Mahlobo R 2015. National Benchmark Test as a benchmark tool. In LD Mogari (ed). *ISTE International Conference on Mathematics, Science and Technology Education Proceedings*. Pretoria, South Africa: Unisa Press. Available at https://www.researchgate.net/profile/Kwanele-Booi/publication/301690956_The_impact_of_knowledge_gaps_in_conceptualisation_and_drawing_skills_in_the_first_year_Life_Sciences_education/links/58bdc4cbaca27261e52e9523/The-impact-of-knowledge-gaps-in-conceptualisation-and-drawing-skills-in-the-first-year-Life-Sciences-education.pdf. Accessed 30 November 2021.

Makgato M & Mji A 2006. Factors associated with high school learners' poor performance: A spotlight on mathematics and physical science. *South African Journal of Education*, 26(2):253–266. Available at http://www.sajournalofeducation.co.za/index.php/saje/article/view/80/55. Accessed 30 November 2021.

Maliki AE, Ngban AN & Ibu JE 2009. Analysis of students' performance in junior secondary school mathematics Examination in Bayelsa State of Nigeria. *Studies on Home and Community Science*, 3(2):131–134. https://doi.org/10.1080/09737189.2009.11885288

Masters GN 1982. A Rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174. https://doi.org/10.1007/BF02296272

McNabb R, Pal S & Sloane P 2002. Gender differences in educational attainment: The case of university students in England and Wales. *Economica*, 69(275):481–503. https://doi.org/10.1111/1468-0335.00295

Mellenbergh GJ 1994. Generalized linear item response theory. *Psychological Bulletin*, 115(2):300–307. https://doi.org/10.1037/0033-2909.115.2.300

Osborne JW 2001. Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology*, 26(3):291–310. https://doi.org/10.1006/ceps.2000.1052

Rasch G 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase MD 2009. *Multidimensional item response theory*. Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-0-387-89976-3

Reddy V, Prinsloo C, Arends F, Visser M, Winnaar L, Feza N, Rogers S, Janse van Rensburg D, Juan A, Mthethwa M, Ngema M & Maja M 2012. *Highlights from TIMSS 2011: The South African perspective*. Pretoria, South Africa: Human Sciences Research Council. Available at https://repository.hsrc.ac.za/bitstream/handle/20.500.11910/2877/7830.pdf?sequence=1&isAllowed=y. Accessed 30 November 2021.

Samejima F 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2):100.

Samejima F 2016. Graded response models. In WJ van der Linden (ed). *Handbook of item response theory* (Vol. 1). Boca Raton, FL: CRC Press.

Sasman M 2011. Insights from NSC mathematics examinations. In H Venkat & AA Essien (eds). *Proceedings of the Seventeenth National Congress of the Association for Mathematics Education of South Africa (AMESA)* (Vol. 1). Johannesburg, South Africa: AMESA. Available at http://amesa.org.za/AMESA2011/Volume1.pdf#page=10. Accessed 30 November 2021.

Shapiro ES, Keller MA, Lutz JG, Santoro LE & Hintze JM 2006. Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, 24(1):19–35. https://doi.org/10.1177/0734282905285237

Sharkness J 2014. Item Response Theory: Overview, applications, and promise for institutional research [Special issue]. *New Directions for Institutional Research*, 2014(161):41–58. https://doi.org/10.1002/ir.20066

Smith J & Naylor R 2001. Determinants of degree performance in UK universities: a statistical analysis of the 1993 student cohort. *Oxford Bulletin of Economics and Statistics*, 63(1):29–60. https://doi.org/10.1111/1468-0084.00208

Spaull N 2013. *South Africa's education crisis: The quality of education in South Africa 1994-2011*. Johannesburg, South Africa: Centre for Development & Enterprise. Available at http://www.section27.org.za/wp-content/uploads/2013/10/Spaull-2013-CDE-report-South-Africas-Education-Crisis.pdf. Accessed 30 November 2021.

Spaull N & Kotze J 2015. Starting behind and staying behind in South Africa: The case of insurmountable learning deficits in mathematics. *International Journal of Educational Development*, 41:13–24. https://doi.org/10.1016/j.ijedudev.2015.01.002

Stillman G 2001. The impact of school-based assessment on the implementation of a modelling/applications-based curriculum: An Australian example. *Teaching Mathematics and Its Applications: International Journal of the IMA*, 20(3):101–108. https://doi.org/10.1093/teamat/20.3.101

Subjex 2019. *National Senior Certificate (NSC)*. Available at https://www.subjex.co.za/subjex-education-system/national-senior-certificate. Accessed 27 October 2019.

Umalusi 2013. *Directives for certification - National Senior Certifcate (schools)*. Pretoria, South Africa: Author. Available at https://www.umalusi.org.za/docs/directives/2013/directives_nsc.pdf. Accessed 30 November 2021.

Van der Berg S & Louw M 2007. *Lessons learnt from SACMEQII: South African student performance in regional context* (Stellenbosch Economic Working Papers: 16/07). Stellenbosch, South Africa: Department of Economics and Bureau for Economic Research, University of Stellenbosch. Available at https://ideas.repec.org/p/sza/wpaper/wpapers47.html. Accessed 30 November 2021.

Wagner DA, Lockheed M, Mullis I, Martin MO, Kanjee A, Gove A & Dowd AJ 2012. The debate on learning assessments in developing countries. *Compare: A Journal of Comparative and International Education*, 42(3):509–545. https://doi.org/10.1080/03057925.2012.670480

Wedekind V 2013. *NSC pass requirements*. Pretoria, South Africa: Umalusi. Available at https://www.umalusi.org.za/docs/research/2013/nsc_pass.pdf. Accessed 30 November 2021.

Yong HT & Sam LC 2008. Implementing school-based assessment: The mathematical thinking assessment (MATA) framework. In *Seminar Inovasi Pedagogi* [Innovation and Pedagogy Seminar]. Sarawak, Malaysia: Institute of Teacher Education. Available at http://www.ipbl.edu.my/portal/penyelidikan/seminarpapers/2008/5a%20Hwa.pdf. Accessed 30 November 2021.

Yung BHW & Yung BHW 2001. Three views of fairness in a school-based assessment scheme of practical work in biology. *International Journal of Science Education*, 23(10):985–1005. https://doi.org/10.1080/09500690010017129

Zondo NP, Zewotir T & North D 2020. *Insights into the final Grade 12 Mathematics examination marks: A quantile regression approach*. Unpublished.