Anil Kanjee
Tshwane University of Technology. Email: KanjeeA@tut.ac.za

Qetelo Moloi
Tshwane University of Technology

DOI: http://dx.doi.org/10.18820/2519593X/pie.v34i4.3
ISSN 0258-2236
e-ISSN 2519-593X
Perspectives in Education
_2016 34(4): 29-51_

A standards-based approach for reporting assessment results in South Africa

Abstract

This article proposes the use of a standards-based approach to reporting results from large-scale assessment surveys in South Africa. The use of this approach is intended to address the key shortcomings observed in the current reporting framework prescribed in the national curriculum documents. Using the Angoff method and data from the Annual National Assessments, the article highlights how standard setting procedures should be conducted to develop meaningful reports that provide users with relevant information that can be effectively used to identify and develop appropriate interventions to address learning gaps. The findings of the study produced policy definitions and performance level descriptors that are proposed for use in enhancing the reporting of results for grade six English and mathematics. Moreover, the findings also indicate that the reporting of the Annual National Assessments using the national curriculum reporting categories overestimates the percentage of learners classified at the lowest performance levels and underestimates those in the next category. This finding has serious implications for the implementation of targeted interventions aimed at improving learning for all. The paper concludes by noting areas of further research for enhancing the use of results of large-scale assessment surveys and for supporting schools and teachers in addressing specific learning needs of all learners, especially the poor and marginalised.

Keywords: _Standard setting, Angoff method, performance levels, Annual National Assessments_

1. Introduction

Large-scale assessment surveys (LSAS) have been implemented in South Africa since the abolishment of the apartheid system and have evolved over time, changing in name, purpose, design, scope and frequency along the way (DoE, 1998, 2003, 2005, DBE, 2011b, 2013; Kanjee, 2007). The initial LSAS were administered on a sample basis in grades 3, 6 and 9, with the first survey administered in 2000. In 2011, the format was revised with LSAS being administered annually in all schools to determine the mathematics and language performance of all learners in grades 1 to 6 and later in grade 9. The primary purpose of this assessment, referred to as the Annual National Assessments or ANAs were to provide an objective picture of learners' competency levels, provide an analysis of difficulties experienced by learners and assist schools to design teaching programmes that are targeted at improving learning in classrooms. The policy document further notes that the ANAs should also assist in setting realistic improvement targets and help parents understand how their children are performing in nationally set tests (DBE, 2012).

The ANAs were administered in all schools from 2011 to 2014, with national learner performance and diagnostic reports produced by the DBE. The results of the ANAs are reported in rankings of learners based on the percentage correct responses in the administered test, as specified in the National Curriculum Statements (DoE, 2002) and the Curriculum and Assessment Policy Statement – CAPS (DBE, 2011a). Issues of whether ANA results are recorded, reported and disseminated in ways that make their meaning understandable so that they can be utilised optimally by targeted users have aroused growing interest into the role that LSAS in general and ANAs in particular, can play in improving the quality of teaching and learning in schools (Hoadley & Muller, 2014; Kanjee & Moloi, 2014). Reporting assessment results in raw scores such as percentage correct responses has received sharp criticism regarding its utility value as well as its lack of measurement accuracy and consistency (Braun & Kanjee, 2007; Bond & Fox, 2007; Dunne *et al*., *2012).*

Continued use of a reporting format that suffers from the aforementioned flaws of raw scores, viewed against the arguably rising stakes around the assessments, does not only compromise the usefulness of the results but it also sets serious limits to the possible impact that the assessment could make towards data-driven decision-making at national, provincial, district and school levels. This will affect the possible improvement of learning and teaching in South Africa. Meanwhile the results of national and international surveys continued to show that the performance levels of South African children were unacceptably low and, in the grades that participated in international studies, were lower than performance in neighbouring countries that invest significantly lower levels of resources in their education system (van der Berg, 2008). In their analysis on the use of standards in South Africa, Snow (2014) and Young (2014) noted that a key challenge facing educators pertains to the development of a culture of effective use of evidence from assessment data for identifying and addressing learning gaps

It is within this context that we explored the option of developing and using performance standards as a basis for reporting assessment results to improve their use for enhancing learning and teaching in schools. Specifically, this article reports on a study conducted to determine the similarities and differences of reporting assessment results using a standards-based approach versus the CAPS reporting specifications. It also investigates whether a standards-based approach can address the identified limitations that influence the CAPS reporting specifications.

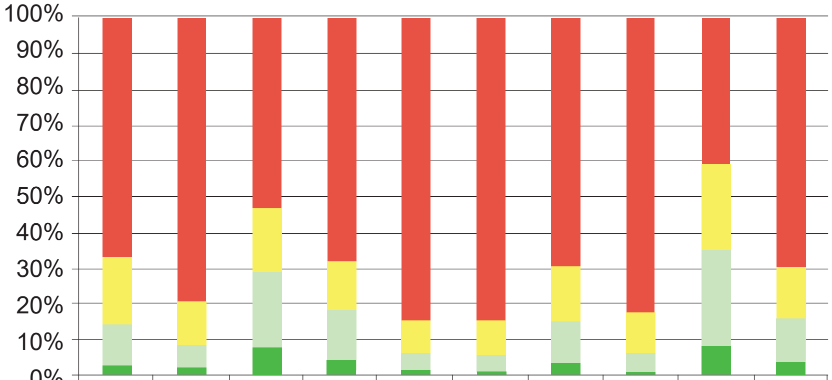
2. Curriculum specifications for reporting results

In the National Curriculum Statements (DoE, 2002) the specifications for reporting learner performance provide for a four-level scale numbered from 1 to 4, a description for each level and a score range associated with each level, as indicated in table 1 (DBE, 2011b). As noted in figure 1, the results of the 2011 ANAs for English First Additional Language (FAL) were reported using the four level rating scale.

Table 1: Levels and score-ranges in the DoE reporting framework

Level	Level description	Score range (%)
4	Outstanding	70-100
3	Achieved	50-69
2	Partly Achieved	35-49
1	Not Achieved	0-34

(Source: DBE, 2011b: 30)



	EC	FS	GP	KN	LP	MP	NC	NW	WC	SA
Not achieved	67	80	54	68	85	85	70	83	41	70
Partially achieved	19	12	17	14	9	9	15	11	24	15
Achieved	12	7	22	14	5	5	12	5	27	12
Outstanding	2	2	7	4	1	1	3	1	8	3

Figure 1: Percentage of grade 6 learners in achievement levels for English FAL by province for the 2011 ANAs (DBE, 2011: 31).

Following the Curriculum Review of 2009 (DoE, 2009), the DBE continued to report assessment results in raw scores using a revised reporting format that is specified in the latest CAPS documents (DBE, 2012). The revised reporting format specifies a seven-level rating scale numbered from 1 to 7, a description for each level and a score range associated with each level, as indicated in table 2.

Table 2: DBE performance levels

Rating codes	Description of competence	Percentage
7	Outstanding Achievement	80-100
6	Meritorious Achievement	70-79
5	Substantial Achievement	60-69
4	Adequate achievement	50-59
3	Moderate achievement	40-49
2	Elementary achievement	30-39
1	Not achieved	0-29

(Source: DBE, 2013: 57)

The results of the ANAs from 2012 onwards have been reported according to this framework prescribed in the CAPS document as indicated in figure 2.

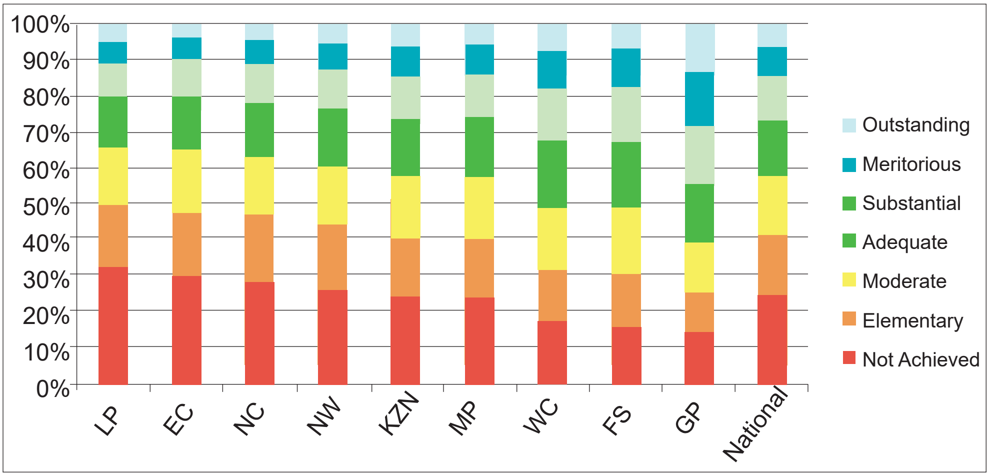


Figure 2: Percentage of grade 6 learners in achievement levels for English FAL by province for the 2013 ANAs (DBE, 2013: 79).

However, the curriculum reporting framework has a number of limitations that influence how assessment results are understood and used in practice. First, the rating codes as prescribed are not associated with any descriptions regarding what learners should know and be able to do. In the absence of such descriptions, the information from any assessment is limited for use in understanding and addressing learning gaps. For example, a score of 56% does not inform parents on what their child knows or can do nor does this score assist teachers in identifying and addressing the strengths and weaknesses of their learners. Second, the fixed raw scores that define the different levels do not account for variations in the difficulty of tests. Bond and Fox (2007) note that percentage correct raw scores depend on the difficulty of the test and the abilities of the test takers. For an easy test, percentage correct raw scores tend to be higher than on a difficult test while in the same test, it is expected that learners of higher ability score higher than their counterparts of lower ability. Therefore, a meritorious achievement in an easy test may not necessarily be so in a difficult test. In standards-based reporting, test difficulty is taken into consideration during the rating of test items and the setting of cut-scores.

Third, the use of raw scores is predicated on the assumption that the effort of improving one's score on the lower end of the ability continuum is the same on the upper end. Yet, studies based on the *latent-theory* model show that it is more difficult to improve from a score of, say, 80% to 85% than from 20% to 25% even though the interval (5%) is the same in both cases (Bond & Fox, 2007; Dunne *et al.*, 2012). Fourth, reporting performance against seven levels (DBE, 2012) could compromise the accuracy of information and meaningfulness of associated reports. An increase in the number of reporting levels will invariably always be accompanied by a decrease in the accuracy with which skills required for adjacent levels can be distinguished. For instance, distinguishing between skills and knowledge that characterise the "*Adequate Achievement*" and "*Moderate Achievement*" levels (Levels 2 and 3 respectively), could be less convincing than distinguishing between "*Not Achieved*" and "*Adequate Achievement*" levels. Similarly, a distinction between Levels 4 and 5, "*Adequate Achievement*" and "*Substantial Achievement*", respectively, could be equally less clear.

Fifth, notwithstanding the information limitations of percentage scores, a higher workload burden is placed on teachers and school leaders by expecting them to be able to record, report, categorise and address learner needs across seven levels of performance. Our view is that it is unrealistic to expect a teacher to keep track of and provide differentiated support across seven categories of learners in a class especially in the context marked by large class sizes, limited resources and high administrative workloads of teachers.

3. Practical implications of the CAPS reporting framework

In his review of studies related to the effective use of assessment data in decision-making by educational districts, Marsh (2012) concluded that one factor that influenced the success or failure of interventions to improve data use was the quality and usability of the focal data. Data that was organised into usable information, easy-to-understand and enabled users to make comparisons was reported to facilitate deep conversations and learning, which in turn led to comprehensive and valid understanding of problems and potential solutions (Marsh, 2012: 30-31).

In a survey (n=39) conducted to understand the perceptions of curriculum and assessment specialists at national and provincial levels on the value of reports that are based on the assessment policy reporting structure, Moloi (2016) notes that 82% of the respondents were of the view that the percentage-based reporting format was not useful. The typical motivation from this group was that the percentage-based reporting format "… does not in any way assist the process and cannot help anyone come up with intervention strategies to improve the system" (Moloi, 2016: 139). In his study on the use of the ANA results, Govender (2016) reported wide variations on how officials at provincial and district levels utilise and reported data in the two provinces that he sampled. Although the officials that Govender (2016) interviewed were aware of the utility value of the data, the majority reported that they lacked the technical capacity to analyse and interpret the data in meaningful ways. If officials, whose responsibility is to support teachers in the use of assessment data, are not entirely convinced about the value of the reporting format that they must promote, it is highly unlikely that they will be able to support teachers in making effective use of the data to improve the quality of teaching and learning.

4. Exploring a standards based approach

A common feature among education systems characterised by their enhanced use of assessment data pertains to the use of a standard setting process for reporting assessment results (Hambleton & Rogers, 1991; Goodman & Hambleton, 2004; Ravela, 2005). The purpose of setting performance standards is to provide a frame of reference in which policymakers, educators and the public can understand test results and provide more interpretive information regarding the meaning of learner scores (Hambleton, 2001). Cizek and Bunch (2007: 13) define the process of standard setting (SS) as “establishing one or more cut-score(s) on a test for [the] purposes of categorising test-takers according to the degree to which they demonstrate the expected knowledge and/or skills that are being tested”. By using a standards-based approach, learners are placed into ordered performance categories since performance standards make targeted and differentiated interventions possible as each learner or group of learners can be placed into specific performance categories that provide detailed information on what these learners know and can do.

In educational circles, a distinction is made between “content standards” and “performance standards” (Cizek & Bunch, 2007; Hambleton, 2001). Content standards are defined as that which learners need to learn. In the South African context, “content standards” are spelt out in the curriculum documents, i.e. CAPS by grade and by subject (DBE, 2011a), which specifies the nature and scope of content knowledge that a learner must acquire in a given grade. On the other hand, Hambleton (2001: 2) defines performance standards as,

> *Well-defined domains of content and skills and performance categories for test score interpretation (that) are fundamental concepts in educational assessment systems aimed at describing what examinees know and can do.*

4.1 Process of setting standards

The process of setting performance standards proceeds in two phases. Phase one comprises the specification of performance level descriptors (PLDs) that provide detailed information regarding the specific tasks, knowledge and skills as well as the degree to which these are expected to be mastered by learners at a given grade level (Hambleton, 2000; Cizek, Bunch & Koons, 2004). The PLDs describe in concrete terms what performance at a given level imply and help interpret what the learners at each level can do and potentially cannot do. According to Cizek, *et al.* (2004) it is highly desirable for PLDs to be developed in advance of standard setting by a separate committee for approval by the appropriate policymaking body. A separate panel of standard-setting participants then use these PLDs as a critical referent for their judgements in developing cut-scores. Separating the teams that develop PLDs from those that determine cut-scores serves as a validation measure because the panel that determines cut-scores independently also scrutinises the PLDs.

When developing PLDs, Perie (2008) notes that panellist should start with the policy definitions and expand these definitions in terms of specific knowledge, skills and abilities required at each level for each subject for each grade. Specifically, Perie (2008) proposes the following guidelines when developing PLDs. First, specify the number and name of the level, second, draft the policy definitions and then develop a written description for each level.

Phase 2 comprises the calculation of cut-scores to determine the minimum score required for learners to be categorised within each performance level. Cizek and Bunch (2007) note that the use of experts to rate the degrees of competency required for responding correctly to a specific set of items is an accepted approach for determining cut-scores. In this process, experts will be judging the percentage of items in a particular test that must be correct in order for performance to be classified as having reached a certain predefined standard. Alternatively, the experts may be required to indicate the percentage of learners at a certain performance level who are likely to get a particular item right. The two ways of judging items are expected to lead to the same result. The Angoff standard setting procedure, which takes the latter approach, was applied in this project.

The Angoff standard setting procedure is a research-based procedure used since the early 1970s and is the most commonly used standard setting method (Hambleton, 2001). It has undergone many modifications over the years and is often referred to as the Modified Angoff or Extended Angoff procedure. In the Angoff procedure, each question has to be evaluated by each panellist. For each question, the method is to request the panellist to make a judgement about the probability that a minimally qualified examinee at each performance level will answer the item correctly. A common modification is to ask, “If we had 100 barely proficient students in the room, how many of them would answer this question correctly?” The judgement made in this way by each panellist is referred to as the item rating and the ratings of panellists will then eventually be combined to arrive at a final item rating. Item ratings can then be summed up to arrive at a cut-score, which is the lowest mark required to be classified into a certain performance category in terms of a raw score. Although standard setting is a judgemental process, the standards set are not arbitrary as they represent the best estimate by informed experts.

The main advantages of the Angoff method of standard setting are that it is widely used in a number of fields such as education and medicine and is well supported by research evidence (Cizek & Bunch, 2007; George, Haque & Oyebode, 2006; Näsström & Nyström, 2008). The method involves discussions among panellists who are knowledgeable about the subject field and therefore, can generate performance standards that are rich in the content knowledge and expected skills at each performance level (Cizek & Bunch, 2007). However, the Angoff method can be quite labour intensive and time consuming (George *et al.,* 2006; De Lisle, 2015).

Hambleton (2001) and Cizek *et al.* (2004) propose the following guidelines for setting cut-scores. First, choose a standard-setting method, prepare training materials and a meeting agenda, second, select and train a large and representative panel, third develop descriptions of the minimally proficient learner, fourth allow panel members to implement rating procedures, fifth, review the process and final results and finally, report on the standard-setting process applied.

5. Methodology

The standard setting process was conducted over two separate workshops for the development of the PLDs and to determine the cut-scores. For both workshops, the focus was on grade 3 literacy and numeracy as well as grades 6 and 9 English Home Language, English FAL and mathematics. For this paper, only results for grades 3 literacy and numeracy and 6 English FAL and mathematics are reported as these are the subjects that are taken by the majority of the learners in schools, while the grade 9 results were based on pilot data that were still to be reviewed.

5.1. Selection of panel members

All panellists that participated in the standard setting process were drawn from qualified and experienced teachers who had been working for the DBE on developing items for the ANAs and thus had the requisite subject area expertise as well as experience in item writing and test development. They were from a diverse background and were broadly representative of the population of South African teachers. For the workshop on the development of performance levels, three panellists for each subject area and grade were selected. However, on the day of the workshop, one panel member was absent and thus only 14 panellists were involved in the process (see table 3).

Table 3: Number of panellist by grade and subject that participated in the PLD workshop

Grade and subject speciality	Number
Grade 3 Home Language	3
Grade 3 Numeracy	3
Grade 6 Home Language	3
Grade 6 First Additional Language	2
Grade 6 Mathematics	3

For the workshop on setting cut-scores, the plan was to select 32 panellists to create two groups of four raters for each grade level and subject area. However, only 24 panel members turned up to participate in the workshop. For each grade 6 subject, raters comprised of two groups while for each grade 3 subject, the raters comprised of one group (see table 4).

Table 4: Number of panellist by grade and subject that participated in the workshop on setting cut-scores

Instrument	Group 1	Group 2
Grade 3 Literacy	4	-
Grade 3 Numeracy	4	-
Grade 6 English FAL	4	4
Grade 6 Mathematics	4	4

5.2. Development of PLDs

The process of developing PLDs comprised five stages. First, panellists were informed of the purpose and intended objectives of the workshop. During this phase, views of panel members were solicited pertaining to the reporting of learner scores. There was unanimous agreement that the current forms of reporting, as specified in the curriculum documents were inadequate and did not provide any useful information to parents or teachers for addressing learner challenges in improving learning. Second, panel members were introduced to the stages in developing PLDs and trained on how to write and edit PLDs. This process comprises reviewing the purpose and process of developing PLDs as well as performance and PLDs from other countries. Third, panel members participated in determining the number of levels and policy definitions required to produce useful reports for teachers and education department officials.

Fourth, panel members were divided into groups to begin with the writing and initial editing of the draft PLDs. Participants were allocated to working groups according to phase and subject area speciality to ensure that their expertise and experience could be used in a more effective and efficient manner. In addition, it also allowed the teachers to more closely examine the horizontal alignment (within each grade level) and vertical alignment (between and among grade levels) of the PLDs. Four groups were established: foundation phase, English Home Language, English FAL and mathematics. In each group, panellists first worked to develop the PLDs for their grade and subject areas and second, to review the PLDs of their colleagues. For example, the grade 6 mathematics group first developed their PLDs and then reviewed the PLDs developed by the grade 3 group. Fifth, panellists within each group reviewed, revised and edited the PLDs. Thereafter, the PLDs were sent to a team of researchers and district officials with relevant expertise in the subject area for review and comment. These comments and suggestions were noted and relevant revisions were incorporated into the final version of the PLDs.

5.3. Determining cut-scores

The first stage in the process of determining cut-scores required the selection of a standard setting method. For this study, the Angoff method was identified as the most appropriate as noted above.

5.3.1. Training of panellists

Training was aimed at familiarising the panellists with the theory of standard setting, the test that they were going to work with and the actual rating of test items. Three researchers, one of whom was a standard setting (SS) expert, conducted the panellists' training. Key activities in the training of panellists included a review of the PLDs that had been previously developed, a PowerPoint presentation by the SS expert, questions from participants and explanatory answers by the SS expert. It also included test taking by panellists, leading panellists on how to use the rating forms, a practical exercise to give panellists experience on how to use the rating forms and an outline of the procedure to be followed in the rating process.

The participants endorsed the performance levels (PLs) and PLDs for each subject and grade and confirmed that they were still relevant, mainly because there had been no changes in the national curriculum between the two events. They also adopted the number of PLs, the policy definitions and the labels of the PLs. Panel members were introduced to key concepts such as cut-scores and how they are used to differentiate learners according to their competencies, performance level descriptors and how they help to identify learners who are able to demonstrate expected knowledge and skills from those who cannot and the possible influences that categorising learners may have for the education system. Particular attention and proportionally more time were spent on defining a "minimally competent" or "borderline" learner. A question-and-answer method, bolstered with different examples, was used to help participants internalise this definition, which was to guide the entire rating process. To simplify the conceptualisation of "minimally competent", participants were asked to make required estimations by answering the question "How many of ten 'minimally competent' learners in your class will get the right answer?" for each of the items that carried one mark. For items that carried more than one mark, the variation of the same question was "What will be the average mark of ten 'minimally competent' learners on this item?"

To enable participants to familiarise themselves with the test, they were asked to individually respond to each item in the test that they were going to work with. Upon completion of test taking, panellists were each given a rating sheet and asked to rate items from a different test with the same item structure as the target test, which was deliberately used for practise purposes. To limit variations in responses that required estimating average scores of "minimally competent" learners during the practise session, panellists were restrained to select from given values, which were limited to multiples of five (5).

5.3.2. Rating of test items

Different groups of panellists working with a specific grade and subject area separately conducted the rating of the test items. The Angoff method centres on estimating the likely achievement of learners who function just across the borderline into a specified PL. Such learners may be referred to as "Just achieving" learners. Ratings had to be done for each item at each performance level, i.e. one for each of the partly achieved, achieved and advanced levels. No such rating was required for the lowest category (not achieved) as this level was automatically defined as any score below the "Partly Achieved" level.

For each item counting one mark, each rater had to answer the following question. "How many of 10 just proficient learners will get this item right?" For each item counting more than one mark, s/he had to estimate the mean score of 10 just proficient learners at each level. For each group, the average rating for each level was calculated and then averaged across items and panellists to arrive at the cut-scores for the test. The process of rating items followed three (3) rounds. The next section presents an account of the purpose, the activities and the outcomes of each round.

Round 1: Panellists were given specially designed forms for rating items and were guided on the information that they were expected to record on the forms. They then worked individually to rate items at each performance level from partly achieved, achieved through to advanced and recorded their ratings on the specially designed form. Individual ratings were captured, the research team did the necessary calculations and feedback was presented on a printed Excel spreadsheet, which showed the *mean* rating of each item by the group and the *range* in individual ratings per item. In groups, panellists then discussed the feedback, particularly focusing on items where score ranges were in excess of 2 points.

Round 2: Panellists worked individually to rate items again, this time taking into consideration the arguments and counter-arguments from the group discussions. Individuals were given the options of either changing their scores, if they were convinced by motivating arguments or leaving their scores unchanged. Individual ratings were captured and feedback was again presented in printed Excel spreadsheets as described in Round 1. Learner scores, with calculated difficulty levels approximated by *p*-values were given as additional input for consideration. The definition of the *p*-value of an item as used in this context refers to the proportion of a well-defined population or sub-population of examinees that get the item correct in a test (Stage, 2003: 2).

Round 3: Panellists worked individually to rate items again, this time taking into consideration the learner responses and *p*-values of individual items as additional input to inform their ratings. For example, if the ratings so far indicated that a particular item was estimated to be more difficult than the learner scores or *p*-values showed, panellists were free to either re-consider their ratings or leave the ratings as is if they were not convinced otherwise.

5.3.3. Capturing of data

On completion of each round of ratings, a team of appointed data capturers collected and captured the rating sheets on an Excel spreadsheet. Each dataset was double-captured by two people. A senior researcher monitored the data capturing process and, when errors were identified, the affected data capturer was stopped and asked to check the entire sheet before proceeding. Cut-scores were then calculated from the summarised ratings.

5.3.4. Calculation of final cut-scores

The cut-scores for each grade and subject area were calculated by averaging ratings over items and panellists after the third round of iterations. The cut-scores were presented to the participants for comments and, when everybody was satisfied that the cut-scores were acceptable, the cut-scores were adopted and recommended to form part of the performance standards.

6. Analysis

The analysis conducted for this study was based on the 2011 grade 3 and 6 English FAL and mathematics ANA data (i.e. four datasets). The 2011 ANA data was used as this was deemed the most appropriate for highlighting the implications of using a standards-based approach in South Africa given that the 2011 results were reported using four levels and thus allowed for direct comparisons without any additional adjustments. Thus, the results of the DBE performance levels were obtained from the 2011 ANA report, which were calculated using the cut-scores noted in table 1 (DBE, 2011b). The results of the SS process were calculated by determining the percentage of learners that fell within the cut-scores calculated for each subject and grade (see table 8).

7. Results and discussion

This section presents the results emanating from each phase of the standard setting process, i.e. phase 1 in which the policy definitions and PLDS were developed and phase 2 in which the cut-scores were determined. In addition, this section also reports on the practical implications of using the standards-based approach at the national and provincial levels by comparing results obtained from the SS process to that reported by the DBE (DBE, 2011b).

7.1. Policy definitions

The final number of performance levels, names of each level as well as the level definitions that the panellist developed during the first workshop is listed in table 5. The policy definition should apply to all subjects and grade levels and should answer the question: how good is good enough? Perie (2008) notes that policy definitions facilitate the articulation of performance levels across grades by ensuring the same level of rigour at each level across each grade and allows a reader to interpret any specified level in a similar manner regardless of the subject assessed. To add more meaning to the policy definitions and to provide a basis for its use in practice, the authors added the two columns titled “Progression implications” and “Intervention implications”.

Table 5: Policy definitions and intervention implications developed by panellists

Level	Level definition	Progression implications	Intervention implications
Advanced	Performance at this level indicates that a learner demonstrates a **comprehensive understanding** of the knowledge and skills required to function at this grade level	Learner has a high likelihood of success in the next grade	Learner requires little or no academic intervention but needs to be provided with more challenging tasks to maximise their full potential
Achieved	Performance at this level indicates that a learner demonstrates **sufficient understanding** of the knowledge and skills required to function at this grade level	Learner has a reasonable likelihood of success in the next grade	Learner may require some assistance with complex concepts to progress to the advanced level
Partly Achieved	Performance at this level indicates that a learner demonstrates **partial understanding** of the knowledge and skills required to function at this grade level	Learner is unlikely to succeed in the next grade without support	Learner requires specific intervention to address knowledge gaps and additional support to progress to the required grade (achieved) level
Not Achieved	Performance at this level indicates that a learner demonstrates **very limited understanding** of the knowledge and skills required to function at this grade level	Learner is unlikely to succeed in the next grade without significant support	Learner requires specific intervention to address knowledge gaps, with extensive and continued support to progress to the required achieved level

7.2. Performance level descriptors

As an exemplar, tables 6 and 7 list the final PLDs developed by the panel members for grade 6 mathematics and English FAL (the grade 3 PLDs were not included due to space constraints). Cizek and Bunch (2007) note that PLDs describe in concrete terms what performance at a certain level implies and provides a basis for interpreting what the learners at each level can do and potentially cannot do. In practice, learners are categorised as functioning in one of the four levels, thus providing teachers and district officials specific information regarding the key knowledge and skills that have been learnt or which still need to be learnt. However, in order to categorise learners into any one of the levels, cut-scores need to be determined for each test.

Table 6: Grade 6 mathematics PLDs developed by panellists

Not Achieved	Partly Achieved	Achieved	Advanced level
A learner at this level can recognise basic number systems and can: • count forward only with whole numbers • count objects not exceeding 10 • add whole numbers up to 10 • draw simple pictures of objects • Measure length of lines • Name few SI units • measure area & perimeter of objects • draw simple bar graphs	A learner at this level can, in addition to skills and knowledge in the lower PL can: • count forward and backwards in decimals • recognise place value up to 9 digits • round off number up to 1000 • add and subtract up to 9 digits • do simple calculations using ordinary fractions and decimals • read digital and analogue time • measure using basic SI units • draw pictographs	A learner at this level can, in addition to skills and knowledge in the lower PLs: • count, recognise and do calculations using fractions & decimals • represent multiples, factors & prime numbers • find percentages of whole numbers • solve problems involving finances and measurement • compare rate and ratio • identify & describe numeric and geometric patterns • use and describe transformations • locate and describe movement on a grid • differentiate between sample & population • draw & interpret bar and pictographs	A learner at this level can, in addition to skills and knowledge in the lower PLs: • Critically read, interpret and analyse with awareness of sources of error and manipulation draw conclusions and make predictions • list possible outcomes for simple experiments including tossing a coin, rolling die and spinning a spinner • distinguish between volume, surface area & dimensions of rectangular prisms. • solve problems involving different time zones • estimate, record, compare & convert between SI units (including mass, temperature, distance and capacity) • organise & record data • calculate the median & mode of data • list possible outcomes & predict “likelihood” of events.

Table 7: Grade 6 English FAL PLDs developed by panellists

Not Achieved	Partially Achieved	Achieved	Advanced level
READING			
A learner at this level can read and can: • recognise words that were previously learnt • read forward and answer short questions based on the text • identify the key character in a text	A learner at this level can, in addition to skills and knowledge in the lower PL: • skim & scan some poetical elements • skim, scan and summarise texts • use a vocabulary of 1500 words. • write friendly letters • extract information directly from a short text • read short stanzas and answer literal questions	A learner at this level can: • identify characters and plot setting • identify ethical issues such as cultural/social diversity • highlight the moral lesson behind a story • infer information from a complex text • identify key elements of poetry • use dictionaries in building vocabulary of at least 3000 words • understand factual information from non-fictional documents	A learner at this level, in addition to skills and knowledge in the lower PLs, can: • read poetry effectively • evaluate texts • identify formal and informal texts • easily utilise vocabulary of at least 5000 words • critique texts through book reviews and reports. • analyse formal & informal documents • synthesise information from different parts of a text • demonstrate comprehension of inferred information from a text
WRITING			
A learner at this level can: • write single words and short sentences • represent ideas with drawings • spell few commonly used words correctly • write brief diaries	A learner at this level can, in addition to skills and knowledge in the lower PL: • write personal letters, diaries, news reports • create a book cover • identify similar & different texts • express cause and effect relationships • write simple sentences.	A learner at this level can, in addition to skills and knowledge in the lower PLs: • write for social purposes, e.g. frames, personal letters, diaries, dialogues & simple news reports • design a book cover • develop & edit key language structures.	A learner at this level can, in addition to skills and knowledge in the lower PLs: • write extensively for social purposes • develop news reports • design questionnaires and adverts. • integrate ideas by classifying information • solve problems • use relevant questioning styles to obtain information.

7.3. Recommended cut-scores

This section presents the cut-scores recommended from the standard setting exercise undertaken by panel members for grade 3 literacy and numeracy as well as grade 6 English FAL and mathematics. The final cut-scores recommended were calculated from the last set of ratings and appear in table 8 below. According to this table, the cut-score for partially achieved for grade 3 literacy is 22%. Learners getting less than 22% will fall in the not achieved category while those scoring 22% or between 22% and 46% will fall in the partially achieved category. Similar results can be generated for grade 3 numeracy as well as grade 6 English FAL and mathematics.

Table 8: Cut-scores derived from the standard setting process

Grade	Subject	Cut-scores in %		
		1	2	3
Grade 3	Literacy	22	46	72
	Numeracy	27	47	64
Grade 6	English FAL	18	50	67
	Mathematics	31	58	74

8. Practical implications of using the DBE vs SS cut-scores

In this section, the practical implications of using the cut-scores for reporting on the percentage of learners categorised into each of the four achievement levels are analysed by comparing the national and provincial results obtained from using the standard setting (SS) process to those results reported by the DBE (2011b).

8.1. Implications at national level

The DBE criteria and the standards setting exercise indicated that the majority of learners did not achieve the required learning outcomes. This general pattern held in grade 3 and grade 6 for both subjects. Specifically, for grade 3 literacy, only a third of the learners were functioning at the required grade level while the corresponding proportion for numeracy was one sixth. However, there were large variations between the two approaches regarding the percentages of learners classified at the lower performance levels. Specifically, the use of the DBE criteria resulted in a substantial over-estimation at the not achieved (NA) level and an under estimation at the partly achieved (PA) level across both subject areas and grade levels.

As noted in figures 3 and 4, for grade 3 literacy, 18% more learners were classified as NA and 14% less learners were classified as PA using the DBE criteria. Similarly, for grade 3 numeracy, 12% more learners were classified as NA and 10% less learners were classified as PA using the DBE criteria. No differences were noted for the achieved (Ach) level while for the advanced (Adv.) level, the difference was 2%.

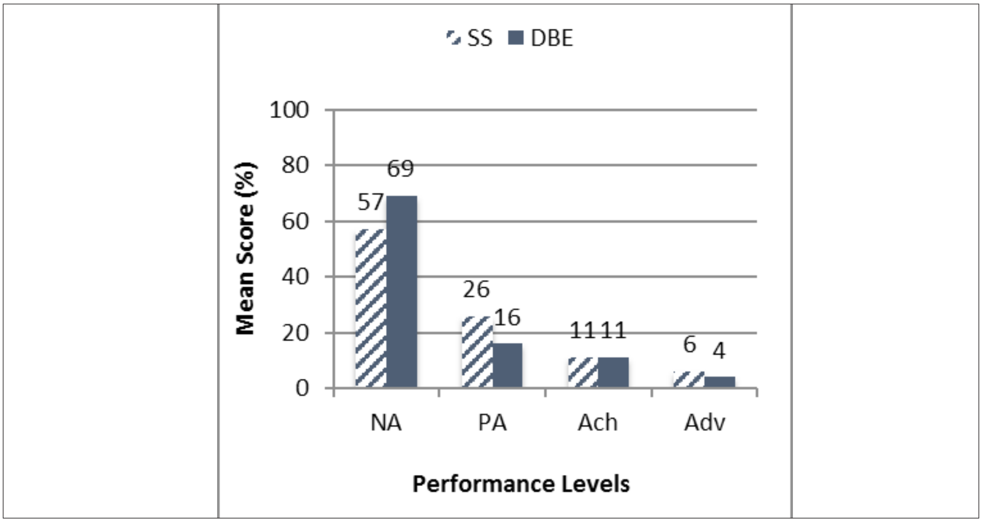


Figure 3: Grade 3 literacy performance using SS v DBE criteria

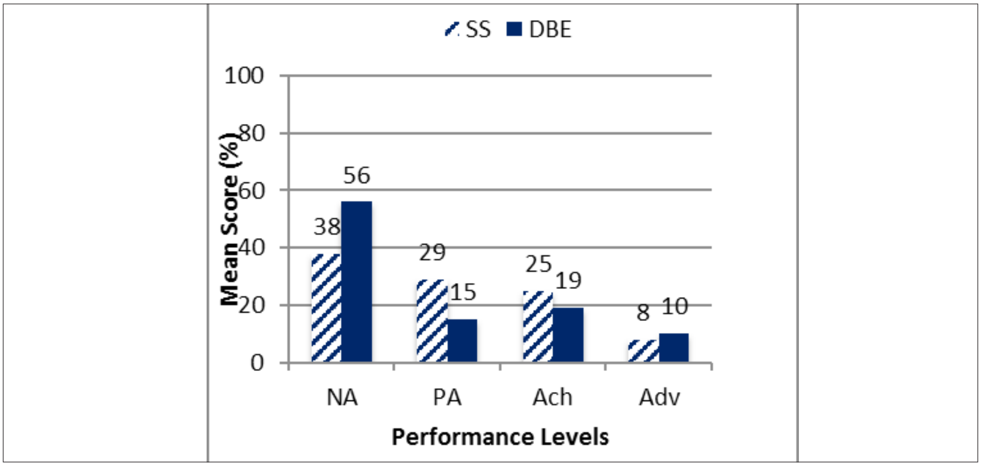


Figure 4: Grade 3 numeracy performance using SS v DBE criteria

The results for grade 6 English FAL and mathematics reveal similar trends. As noted in figure 5 and 6, for grade 6 English FAL 10% more learners were classified as NA and 17% fewer learners were classified as PA using the DBE criteria. Similarly, for grade 6 mathematics, 33% more learners were classified as NA and 19% less learners were classified as PA using the DBE criteria. For the Ach and Adv. levels, the differences for English FAL and mathematics were 3% and 1% respectively.

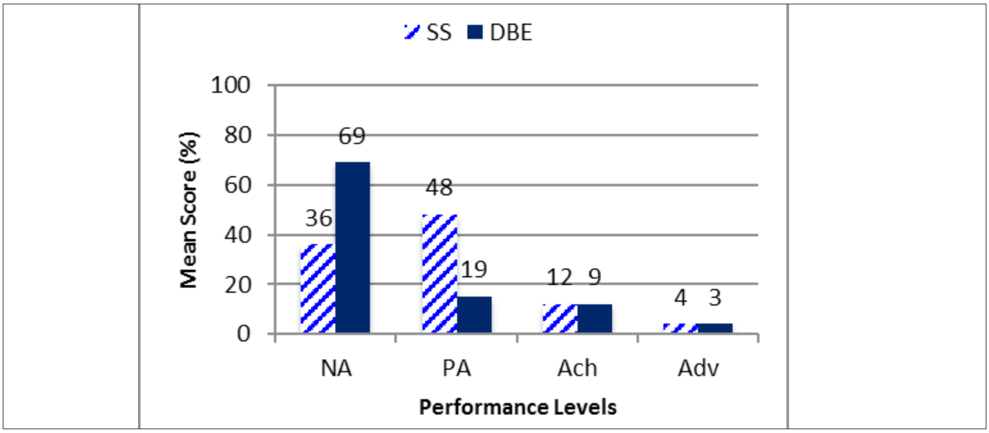


Figure 5: Grade 6 mathematics performance using SS v DBE criteria

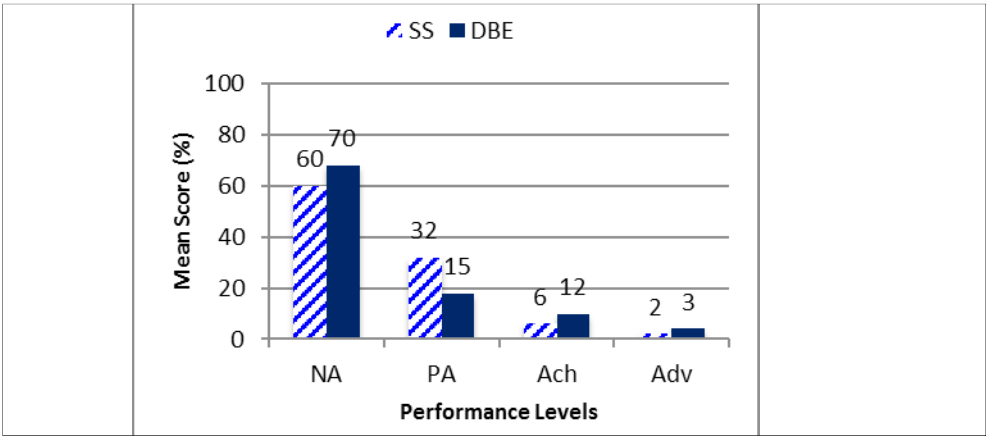


Figure 6: Grade 6 English FAL performance using SS v DBE criteria

8.2. Implications at the provincial level

While the overall performance of province X learners was higher compared to the rest of the country, similar trends were found compared to the national results. That is only a third of the grade 3 literacy and grade 6 English FAL learners while a fifth of the grade 3 numeracy and grade 6 mathematics learners were functioning at or above the required grade level. Moreover, as noted in figures 7 and 8, there were large discrepancies at the NA level and at the PA level between the SS and DBE cut-scores. For grade 3 literacy, 16% more learners were classified as NA and 12% fewer learners were classified as PA using the DBE criteria (Figure 7). Similarly, for grade 3 numeracy, 13% more learners were classified as NA and 10% fewer learners were classified as PA using the DBE criteria. The differences for the Ach and Adv. levels were all less than 5%.

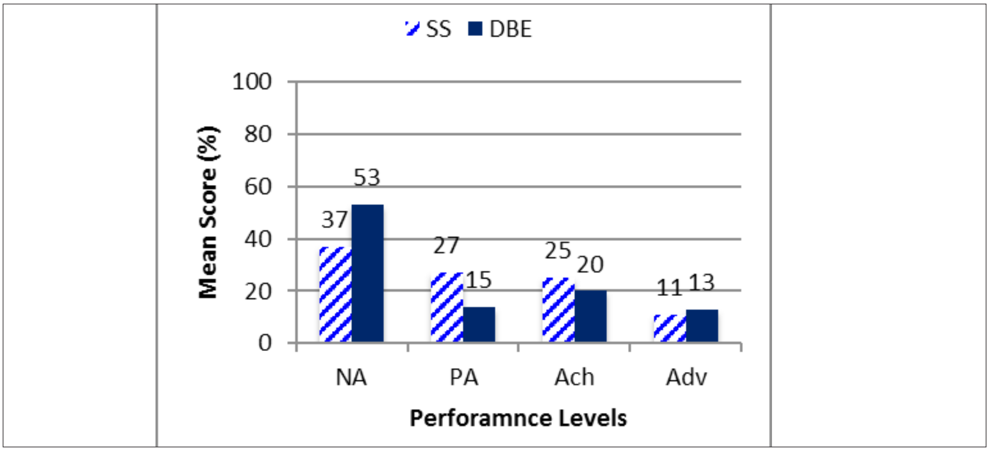


Figure 7: Grade 3 literacy performance using SS v DBE criteria for Province X

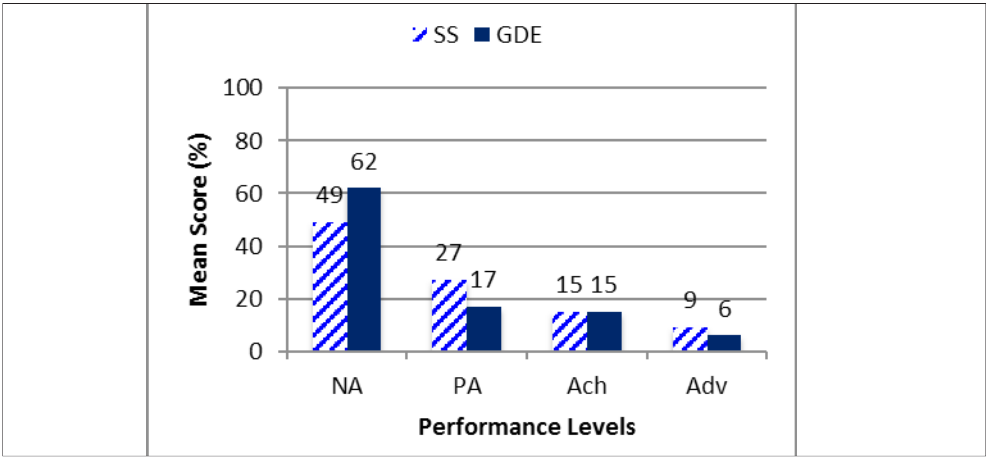


Figure 8: Grade 3 numeracy performance using SS v DBE criteria for Province X

The results for grade 6 English FAL and mathematics reveal similar trends. As noted in figure 9, for grade 6 English FAL, 28% more learners were classified as NA and 28% fewer learners were classified as PA using the DBE criteria. Similarly, for grade 6 mathematics, 8% more learners were classified as NA, 16% fewer learners were classified as PA and 9% fewer learners were classified as 'achieved' using the DBE criteria. For the other levels, only minor differences were noted.

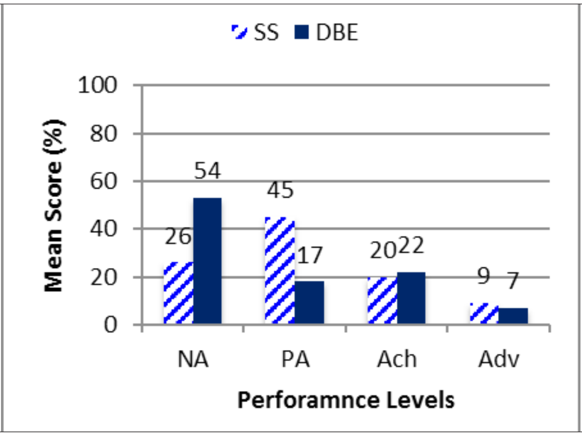


Figure 9: Grade 6 English FAL performance using SS v DBE criteria for Province X

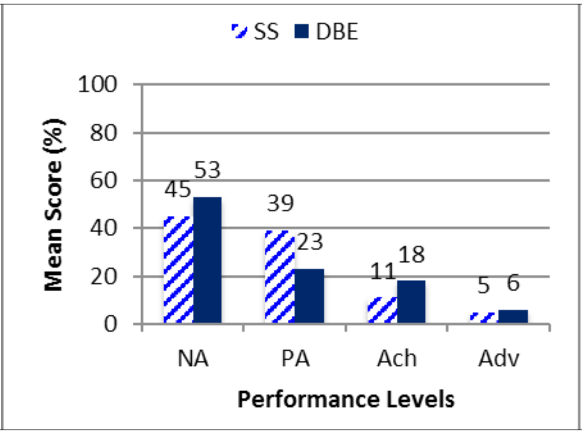


Figure 10: Grade 6 mathematics performance using SS v DBE criteria for Province X

9. Conclusion and way forward

The standard setting approach followed here produced relevant policy definitions and PLDs that could be used to improve the reporting of results from large-scale assessment surveys. In addition, the SS procedure also has the potential to produce well-founded cut-scores for the (ANA) tests, provided the test is appropriate, the standard setting panel is of the right mix and quality and the procedure is conducted meticulously with adequate time devoted to important aspects. The Angoff method used here to provide cut-scores generally yielded plausible cut-scores, with acceptable levels of agreement between raters and the agreement of ratings with *p*-values. Moreover, the use of this procedure can also be extended to results of common examinations that are administered at provincial and national level.

A key finding of the project is that the ANA results reported by the DBE do not accurately identify learners functioning at the lower performance levels. Specifically, the results indicate that the DBE results overestimate the percentage of learners classified at not achieved and underestimate those classified as partly achieved. This could be due to the manner in which the DBE cut-scores were calculated. However, no evidence was found on how the DBE cut-scores were determined. The standard setting cut-scores, on the other hand, were determined applying well known and internationally recognised procedures that account for the difficulty of each test item as well as the ability of learners to respond to that specific item.

Of greater concern, however, is that the inflation of learners in the lowest performance level (NA) and deflation of learners in the next lowest category (PA) can have a critical impact on implementing relevant intervention for improving learning. That is, addressing the needs of learners classified as PA and thus by definition, learners who have acquired some knowledge and skills pertaining to the subject area under concern, require less effort, time and resources as compared to developing interventions for learners who have acquired considerably less skills and knowledge. Given that the current teaching and learning context is marked by a high number of un- and under-qualified teachers as well as limited resources and facilities in the majority of schools in South Africa, the incorrect identification of learner needs can have significant consequences on the challenge of improving learning for all. While the findings from this study clearly demonstrate the value of the standards-based approach to reporting results, additional information is required on how best to report these results to different audiences within the South African education sector to enhance the use of data for use in addressing the challenge of improving learning for all (Kanjee & Sayed, 2013). In this context, we highlight three areas for additional research.

First, there is a need for the development of relevant guidelines, reporting tools and practical templates for supporting teachers and school leaders to enhance the use of results from common examination as well as large-scale assessment studies. Disaggregating data to the level of districts, schools and classrooms has the potential to provide valuable information that can be used to identify and address learning needs of learners in the subject areas assessed. What is still unclear, however, is the form and format for reporting these results and the manner in which teachers, school leaders and district officials will respond to the results based on the standard setting procedures. Moreover, it is unknown whether this information will be effectively used to identify learners in need of assistance and to develop relevant interventions for addressing their learning needs, especially those from poor and marginalised backgrounds.

A second area of research pertains to the use of appropriate standard setting methods for determining the cut-scores. In this study, we applied the Angoff method. However, research indicates that this method might not be the most appropriate nor cost effective (Moloi, 2016). Other approaches that should be explored include the Objective Standard Setting method (Stone, 2001) and the Objective Borderline Method (Shulruf *et al.*, 2015).

Third, there is a need for reports to focus on the issues of equity. The use of a standards-based approach highlights the specific learning needs of learners at the lowest performance levels and provides teachers, school leaders and district officials with specific information for developing relevant information that address the learning needs of these learners, most of whom come from poor and marginalised backgrounds. Moreover, the standards-based approach can also provide more detail and accurate information on the percentages of learners at the lower performance levels, at the district, school and classroom levels. This information can be used for setting targets for reducing the percentage of learners at this level. In this context, the key challenge pertains to how relevant information is reported, adequate support is provided and effective monitoring is conducted to develop and promote a culture of data use for improving learning for all.

References

Bond, T.G. & Fox, C.M. 2007*. Applying the Rasch model: Fundamental measurement in the human sciences*, 2nd ed. London: Lawrence Erlbaum.

Braun, H. & Kanjee, A. 2007. Using assessment to improve education in developing nations. In J.E. Cohen, D.E. Bloom & M.B. Malin (Eds.). *Educating all children: A global agenda*. Cambridge: MIT, MA. pp. 303-353.

Cizek, G.J. & Bunch, M.B. 2007. *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781412985918

Cizek, G. J., Bunch, M. B. & Koons, H. 2004. Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23(4), 31-50. https://doi.org/10.1111/j.1745-3992.2004.tb00166.x

De Lisle, J. 2015. Installing a system of performance standards for national assessments in the Republic of Trinidad and Tobago: Issues and challenges, *Applied Measurement in Education*, 28(4), 308-329. https://doi.org/10.1080/08957347.2015.1062765

Department of Education (DoE). 1998. *Assessment policy in the general education and training Band: Grades R to 9 and ABET*. Pretoria: Government Printer.

Department of Education (DoE). 2002. *Revised national curriculum statement for grade R-12*. Pretoria: Government Printer.

Department of Education (DoE). 2003. *Systemic evaluation: Foundation phase*. Pretoria: Government Printer.

Department of Education (DoE). 2005. *Intermediate phase systemic evaluation Report*. Pretoria: Government Printer.

Department of Education (DoE). 2009. *Report of the task team for the review of the implementation of the national curriculum statement*. Pretoria: DoE.

Department of Basic Education (DBE). 2011a. *Curriculum and assessment policy statement (CAPS).* Pretoria: Government Printer.

Department of Basic Education (DBE). 2011b. *Report of the Annual National Assessment 2011*. Pretoria: Government Printer

Department of Basic Education (DBE). 2012*. Action plan to 2014: Towards the realisation of schooling 2025* – Full version. Pretoria: Department of Basic Education.

Department of Basic Education (DBE). 2013. *Report of the Annual National Assessment 2013: Grades 1 to 6 & 9*. Pretoria: Government Printer.

Dunne, T., Long, C., Craig, T. & Venter, E. 2012. Meeting the requirements of both classroom-based and systemic assessment of mathematics proficiency: the potential of Rasch measurement theory. *Pythagoras*, 33(3), 1-16. https://doi.org/10.4102/pythagoras.v33i3.19

George, S., Haque, M.S. & Oyebode, F. 2006. Standard setting: Comparison of two methods. *BMC Medical Education*, 6(1), 4-52. https://doi.org/10.1186/1472-6920-6-46

Hambleton, R.K. 2001. Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates. pp. 89-116.

Hambleton, R.K. & Rodgers, J.H. 1991. Advances in criterion-referenced measurement. In R.K. Hambleton & J.N Zall (Eds.). *Advances in educational and psychological testing*. Boston: Kluwer Academic. pp. 3-44. https://doi.org/10.1007/978-94-009-2195-5_1

Hoadley, U. & Muller, J. 2014. *Testing, testing: Investigating the epistemic potential of systemic tests*. Mimeograph. Cape Town: University of Cape Town.

Goodman, D.P. & Hambleton, R.K. 2004. Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education,* 17(2), 145-220. https://doi.org/10.1207/s15324818ame1702_3

Govender, D.A. 2016. The use of Annual National Assessments by provinces and districts to improve teaching and learning. Unpublished PhD thesis. Pretoria: Tshwane University of Technology.

Kanjee, A. 2007. Improving learner achievement in schools: Applications of national assessments in South Africa*.* In S. Buhlungu, J. Daniel, R. Southall & J Lutchman (Eds.). *State of the nation: South Africa 2007*. Pretoria: HSRC Press, 470-499.

Kanjee, A. & Moloi, M.Q. 2014. South African teachers' use of national assessment data. *South African Journal of Child Education,* 4(2), 90-113. https://doi.org/10.4102/sajce.v4i2.206

Kanjee, A. & Sayed, Y. 2013. Assessment policy in post-apartheid South Africa: Challenges for improving education quality and learning. *Assessment in Education: Principles, Policy & Practice*, 20(4), 442-469. https://doi.org/10.1080/0969594X.2013.838541

Marsh, J.A. 2012. Interventions promoting teachers' use of data: Research insights and gaps. *Teachers College Record,* 14(11), 1-48.

Moloi, M.Q. 2016. A national framework for reporting the results of large-scale assessment studies in South Africa. Unpublished PhD thesis. Pretoria: Tshwane University of Technology.

Näsström, G. & Nyström, P. 2008. A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research & Evaluation*, 13(9), 1-12.

Perie, M. 2008. A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, 27(4), 15-29. https://doi.org/10.1111/j.1745-3992.2008.00135.x

Ravela, P. 2005. Evaluating students' achievement, a formative approach to national assessments: The case of Uruguay. *Prospects*, 35(1), 21-43. https://doi.org/10.1007/s11125-005-6816-x

Shulruf, B., Poole, P., Jones, P. & Wilkinson, T. 2015. The objective borderline method: A probabilistic method for standard setting. *Assessment & Evaluation in Higher Education*, 40(3), 420-438. https://doi.org/10.1080/02602938.2014.918088

Snow, C.E. 2014. Language, literacy, and the needs of the multilingual child. *Perspectives in Education,* 32(1), 7-16.

Stage, C. 2003. Classical test theory or item response theory: The Swedish experience. *Educational Measurement*, 42, 1-29.

Stone, G. E. 2001. Objective standard setting: or truth in advertising. *Journal of Applied Measurement*, 2(2), 187-201.

Van der Berg, S. 2008. How effective are poor schools? Poverty and educational outcomes in South Africa. *Studies in Educational Evaluation*, 34(3), 145-154. https://doi.org/10.1016/j.stueduc.2008.07.005

Young, M. 2014. Standards and standard setting and the post school curriculum. *Perspectives in Education* 32(1), 17-29.