NATIONAL STANDARDISED ACHIEVEMENT TESTS: A SOURCE OF RELIABLE FORMATIVE FEEDBACK FOR TEACHERS

Simson N. Shaakumeni and Simon P. Mupupa[8]

ABSTRACT

Namibia has been implementing the national Standardized Achievement Tests (SATs) since 2009. These tests (SATs) as a national assessment are aimed at overcoming constraints in the assessment system, especially at the primary school level as well as to improve the quality of education through learner assessment. The Namibian SATs are low-stakes in nature; hence they are not used for promotional purposes but rather to provide diagnostic feedback to schools, decision makers and other education stakeholders. Six years down the line, the SATs results have been showing improvement in learner performance at primary school level. A perception survey conducted amongst primary school teachers in Namibia (N = 130) revealed a largely positive perception about SATs with majority of respondents agreeing with the relevance of this assessment and its continuation in the education system. The purpose of this paper is to present an overview of Namibian SATs as well as to highlight teachers' perceptions about this national assessment.

KEY WORDS: Standardised Achievement Tests, Assessment

[8] **Mr. Simson N. Shaakumeni** has over 14 years of experience in the basic education sector. He holds a Master of Education (Science Education). He is currently, a Senior Education Officer for Research and Development in the Directorate of National Examinations and Assessment, Ministry of Education, Arts and Culture.
Mr. Simon P. Mupupa has 16 years of experience in the basic education sector and is a Master of Education candidate at the University of Namibia. He is currently, a Chief Education Officer for Research and Development in the Directorate of National Examinations and Assessment, Ministry of Education, Arts and Culture.

INTRODUCTION

Namibia has been implementing the national Standardized Achievement Tests (SATs) since 2009. These tests (SATs) are aimed at overcoming constraints in the assessment system especially at the primary school level as was unravelled by the World Bank's study titled *Namibia Human Capital and Knowledge Development for Economic Growth with Equity* (Marope, 2005). The SATs also attempt to improve the quality of education through learner assessment. The Namibian SATs are low-stakes in nature; hence they are not used for promotional purposes but rather to provide diagnostic feedback to schools as form of formative assessment; to decision makers and other education stakeholders in education. Six years down the line, the SATs results have been showing improvement in learner performance at primary school level. The purpose of this paper is to present an overview of Namibian SATs as well as to highlight teachers' perceptions about this national assessment. This paper is divided into two sections. The first section provides background information about the SATs including the objectives of the tests as well as the highlight of the test results since their inception in 2009. The second section provides findings on primary school teachers' perceptions of this national assessment.

SATS background

The World Bank study found the following barriers to learner assessment in the Namibian education system:

- Namibia has too few mechanisms to measure the levels of achievement and performance of the school system, particularly at primary level.
- The mechanisms to provide information to judge the performance of individual schools is not effective.

- The mechanisms to identify teaching and learning difficulties and provide feedback and assistance to individual schools about mastering key skills and competences are insufficient.

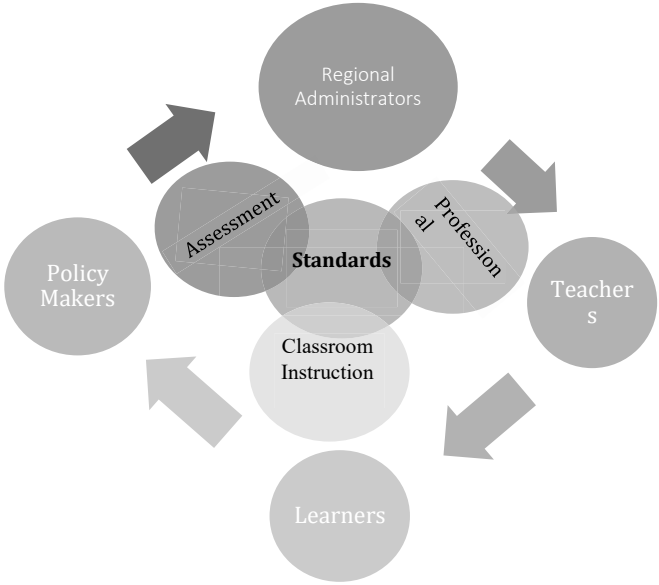
To overcome the above mentioned constraints in the assessment system, especially at the primary level, the Directorate of National Examinations and Assessment (DNEA) in Namibia was tasked with developing a sustainable, long-term assessment system, the Standardized Achievement Tests (SATs). The SATs were developed based on the following objectives:

- Monitor learners' acquisition of identified skills and competencies in key subject areas.
- Set baseline and subsequently monitors the progress of learners at individual schools.
- Disseminate diagnostic feedback from test results to schools and advisory/ inspection services in each of the regions.

Standardized assessment framework

The design of the standardized assessment framework for Namibia was done after reviewing and utilized the best practices of high performing education around the world e.g., Australia, Finland, Singapore, Sweden, and the United Kingdom (Hammond & Wentworth, 2010; Ferrer, 2006; Ravela, 2005). It was revealed from the review that a national assessment system can be made very effective when all stakeholders in the education enterprise participate equally and feel strong ownership.

Figure 1: Standardized Assessment Framework in Namibia



An integrated assessment framework was developed to not only serves as a performance measurement system for monitoring schools' performance and providing feedback but also as a motivational system that serves a number of socio-political or symbolic purposes in communicating to educators, administrators what is expected and in insisting on high expectations for all learners.

Figure 1 shows one view of how the standardized assessment system works in Namibia. The standards are the foundation on which the whole system sits, and these standards establish clear, reasonable, and important goals (NIED, 2008) for what learners are expected to learn (i.e., content standards) and how they should be performed (i.e., performance standards). For learners to attain standards, regions and

schools take action to improve learners' learning opportunities in what and how well learners are taught in classrooms, through supplemental services and programs (e.g., after school supplemental reading program). As necessary, regions organize additional trainings for teachers through continuous professional development (CPD) in teaching and assessing those standards effectively. Learners are assessed through standardized assessments - strongly aligned with standards. The policymakers and regional administrators use the feedback from the assessments for management and improvement purposes: to gauge their strengths and weaknesses; to identify schools that may need special help; and to be strategic in taking action and coordinating available resources to improve learner performance, e.g. through professional development, instructional materials, mentoring, and technical assistance. The framework also had the following key features:

- Coherence with other elements of education enterprise, which also illustrates the importance of standardized assessment of, for, and as learner learning
- Quality of standardized assessment with respect to reliability, validity, and fairness
- Engage learners in standardized assessments to improve their motivation and learning
- Engage teachers in assessment development and administration as a way to improve their professional practice and their capacity to support learners' learning and achievement
- Engage regional administrators in training of teachers on assessment administration and dissemination of results as a way to build their awareness and ownership.

SATS METHODOLOGY

The primary objectives of the SATs are to report learner performance growth from one year to the next and provide diagnostic information about what knowledge and skills (i.e., competencies/standards in the curriculum) learners have mastered and what they have not. Although the original intention of the SATs was to assess learner achievement at the end of each primary phase, lower primary (Grades 1-4) and upper primary (Grades 5-7), it was decided to administer the test to learners at Grades 5 and 7 (Wolfaardt, 2003) in alternate years. Grade 5 was chosen instead of Grade 4 because classroom instruction in English begins at Grade 4 in Namibia and learners at this level may not have adequate experience with English as a medium of instruction to perform on the SATs, which are written in English.

The first Grade 5 and 7 SATs were administered in November 2009 and 2010, respectively constituting the baselines. The first follow up tests for Grades 5 and 7 were administered in 2011 and 2012 and so on and so forth. English Second Language and Mathematics are tested in Grade 5. English Second Language, Mathematics, and Natural Science are tested in Grade 7. Additional subjects and grades may be added to these assessments in the future as the DNEA continues to develop and expand its capacity to implement these assessments.

The DNEA implements a census-based assessment system for the SATs, meaning that all Grade 5 and 7 learners in around 1100 primary and secondary schools in Namibia were tested in alternate years until 2013 and are currently being tested concurrently. A detailed competency based performance report is produced for each school for improving classroom instruction and monitoring the school's progress from year to year.

The following sections describe the details about how SATs are developed and implemented nationwide. The details can be classified primarily into three broad categories: (1) construction of reliable and valid test instruments, (2) development of the baseline reporting scale, (3) and maintenance of the scale for monitoring school performance longitudinally.

Construction of reliable and valid instruments

The test construction is one of the most critical components in a standardized assessment program. It provides information about the validity and integrity of the assessment program. It includes a number of steps: (1) identification of key competencies; (2) test specifications; (3) item development; (4) item banking; (5) pilot testing of the items; (6) test assembly; and (7) test administration. Each of these steps is described as follow.

Identification of key competencies

Since one of the objectives of SATs is to provide information about what learners know and be able to do in key competencies in the English Second Language, Mathematics and Natural Science curricula, therefore the first step in the SATs development process was to have the curricula reviewed carefully by an expert panel and identify the competencies that are absolutely necessary for learners to master at that grade level and can be tested through multiple choice items. The expert panel identified about 30 competencies in Grade 5 English Second Language, 88 in Grade 5 Mathematics, 28 in Grade 7 English Second Language, 63 in Grade 7 Mathematics, and 79 in Grade 7 Natural Science.

Test specifications

Test specification refers to a complete operational definition of test characteristics. For example, it must describe the type of the test format (multiple-choice), number of test forms, total number of items on each test form, cognitive classification of items, item scoring rules, time limit,

etc. Since one of the objectives of the SATs system is to provide information about what learners know and are able to do in key learning competencies in the curricular, DNEA utilizes a *multiple forms common-matrix sampled assessment design* (Petersen, Kolen, & Hoover, 1989) for the SATs, meaning that there is more than one test form for each subject and grade within the same administration year. However, each learner takes only one test form. By using the multiple forms assessment design, DNEA manages to cover about 80% of the key competencies in each subject area in the tests.

It is not easy to maintain the same level of difficulty between the forms (e.g., one form may get easy, one form may get difficult) when constructing the test forms, learners with same ability level get different scores depending on the easy or difficult form they take. In order to practice fair testing for learners, we report learners' scores on the same measurement scale, irrespective of the form they take and their difficulty levels. This is carried out through a procedure called test equating. A set of items is used as common items across the forms to bring the forms on the same reporting scale. In addition to the common items and unique core[9] items, there are a few items embedded per form called field test items. The field test items are the items that do not count toward learners' scores, but are used for pilot testing purposes and will potentially be used in future administration as core items. The visual representation of the SATs' design is as follows.

Figure 2: The visual representation of the SATs' Design

Test Form	Core Matrix Items			Common Items	Field Test Matrix		
Form A							
Form B							
Form C							

A few key features of the SATs' design include:

[9] Items are called core when they are counted toward learners' score. Common items can also be called core when they are counted toward learners' score.

- A total of three test forms (known as form A, B, and C) were administered for each subject.
- Each form consisted of 40-50 core items (including 12-15 common items) that are counted toward learners' scores; 5-10 items are embedded as field test items.
- The testing time for each Grade 5 and 7 subject was two hours.
- The test forms met the test blueprint, meaning the number of points per theme was maintained.
- The test forms met psychometric requirements, indicating that the test forms were of equivalent difficulty.

Item development

Item development refers to the activities in which a group of content experts, usually experienced teachers, get trained on item writing principles and procedures (Haladyna, Downing, & Rodriguez, 2002; Haladyna, 2004). And write items that are strongly aligned with competencies, have varied difficulty (e.g., easy, moderate, hard) and cognitive complexity (i.e., knowledge, comprehension, and application). Training of item writers is an important validity issue for test development process. To develop test items for SATs, DNEA organized a number of item development workshops for Grades 5 and 7 since 2008 and have developed over 9,000 items.

Item banking

Item Banking refers to the process of storing the items systematically and protecting them from exposures and theft. This is an essential component for a national standardized assessment system. Establishing a comprehensive item bank is also much more efficient than attempting to write new items every time when develop a new test.

Pilot testing of items

The items that are developed and banked must be pilot tested to check their psychometric properties before using them in an operational testing. The items with acceptable properties are considered for operational tests: items with difficulty level of 0.25-0.90 and discrimination level of 0.25 and above (APA, AERA, and NCME, 1999). DNEA pilot tested adequate number of Grade 5 and 7 items that are needed for next six years' administrations.

Test assembly

Test assembly refers to the rigorous process of constructing test forms with high technical quality. Test quality includes reliability, the consistency of measure (Cronbach, 1951; George & Mallery, 2003); validity, the degree to which accumulated evidence and theory support specific interpretations of test scores for the defined uses of the test results; accuracy of content and test keys. As stated earlier that DNEA has used a multiple test forms assessment design for SATs, a total of six test forms (2 subjects x 3 forms) for Grade 5 and nine test forms (3 subjects x 3 forms) for Grade 7 are assembled for an administration year. To ensure comparability among test forms for a given grade and subject, each form must meet the test blueprint. This means that the number of points per score reporting category as well as the number of items per themes is maintained. In addition to conforming to the content aspects of the blueprint, the test forms also conformed to the statistical requirements for a test form (e.g., equivalent test characteristic curves across forms).

Test administration

The SATs are administered using a standardized procedure. A systematic procedure is used to ensure random assignment of the test forms to the learners, i.e., if the first learner in the classroom is assigned form A, the second learner is assigned form B, the third learner form C, and the fourth learner is again assigned form A, so on and so forth. Before forms are assigned to learners, test administrators ensure that learners are seated in such a way so that the counting of learners and assignment of forms are made systematically. DNEA conducts a number of training workshops in each of the 14 regions for regional administrators, head teachers, and teachers on standardized test administration procedure. The regional administrators are later responsible for monitoring test administration in their respective schools.

Development of baseline reporting scale

The development of the baseline scale refers to activities that are related to statistical and psychometric analysis of the baseline (i.e., first operational administration) data, setting performance standards and determining cut scores for various performance level categories, and reporting the results. The baseline performance for each school or region is always referenced for comparison of their performances in other administration years. A reporting measurement scale is constructed in such a way that is fair, robust, and easily interpretable. Each of these steps is described as follows:

Statistical and psychometric data analysis

The baseline data are analyzed using both classical and item response theory, IRT (Lord & Novick, 1968; Hambleton & Swaminathan, 1984). The IRT is considered to be a more robust method of estimating learner ability and item parameters than classical test theory because it is not

population dependent. However, the classical test theory (CTT) is also used to cross validate the IRT-based results. Although there are various models (e.g., one, two, or three parameters) in IRT, a one-parameter model (Rasch, 1960/1980) is used for SATs.

For SATs, we have estimated item and test level statistics to report the quality of the instruments developed for Grade 5 and 7 assessments. For example, item difficulty (also called p-value), item discrimination (also called item-total correlation), IRT based item difficulty (also called b-value), differential item functioning (DIF) indicators, and test reliability (also called internal consistency of reliability), etc. In addition, learner ability statistic is also estimated using both classical (total score on the test) and IRT (referred to as theta,θ). Moreover, we also estimate item characteristic curve [ICC], item information curve [IFC], item standard error of measurement [SEM], test characteristic curve [TCC], test information curve [TIF]) based statistics.

Setting performance standards

In order to define learner scores in a more meaningful way with respect to content knowledge and skills (i.e., what learners with different score ranges, such as 0–14, 15–30, and so on, know and are able to do in key learning competencies specified in the curricula), we derive score ranges representing various categories (also called performance-level categories) through a psychometric procedure called standard setting (Perie, 2008; Beck, 2003). The first step of the standard-setting procedure is to decide the number of categories to be used for categorizing the learners, usually three or four categories in best practices; give appropriate labels to those categories; and describe the level of knowledge and skills the learners need to demonstrate on the assessment to be classified into various performance-level categories. The second step is to determine the cut scores corresponding to those performance-level categories. The Namibian education policymakers decided to classify learners into four performance level categories: *Below Basic*, *Basic*, *Above Basic*, and *Excellent*, defined as follows:

Below Basic: The learner demonstrates **insufficient** knowledge and skills across all themes in the syllabus.

Basic: The learner demonstrates **sufficient** knowledge and limited skills across all themes in the syllabus.

Above Basic: The learner demonstrates **proficient** knowledge and skills across all themes in the syllabus.

Excellent: The learner demonstrates **excellent** knowledge and advanced skills across all themes in the syllabus.

The cut scores for the performance-level categories are then determined on the reporting scale using a yes/no variation of the Angoff method for multiple cut scores (Plake, Ferdous, & Buckendahl, 2005). Two separate five days standard setting workshops for Grade 5 and 7 are conducted. A total of 17 teachers for Grade 5 English Second Language, 18 for Grade 5 Mathematics, 13 for Grade 7 English Second Language, 15 for Grade 7 Mathematics, and 15 for Grade 7 Natural Science representing all 14 geographical regions have attended the workshops. The first three days the participants have developed the detailed profiles of *below basic, basic, above basic*, and *excellent* learners, what learners in each of these categories know and be able to do in each key competency in the curricula; the last two days the participants set the cut scores for each of these categories.

Reporting school results

The purposes of the school reporting system are: (1) to inform the policy makers and other stakeholders (teachers, in particular) about what learners in each school know and be able to do in key learning competencies in the curricular so that teachers in the school can use that information in preparing their lesson plans and in classroom instruction; (2) to keep track of each school's progress from year to year.

The DNEA generates two reports for each school: one is called quantitative report, which provides information about how learners in a school perform (in percentage score) as compared to other schools in the region and the nation. It also provides information about what percentages of learners in the school, region, or the nation are classified into the four-performance level categories; these percentages of learners in four performance level categories and overall percent score are used to keep track of school progress from year to year.

The quantitative report is produced through an automated reporting system developed in SPSS. The other report is called qualitative report, which provides information about what learners in various performance categories know and be able to do (Figure 3) in key competencies in the curricula. For example, a below basic learner can only count in multiples of 2, 5, and 10 in the given competency, whereas a basic learner can count in multiples of 2, 5, 10, 20, and 100, which is more than what a below basic learner knows. Therefore, if we have a below basic learner in our class, then we know exactly what s/he can do in that competency and what additional things we have to teach to help her/him move to the next category (i.e., basic).

Figure 3: Examples of competency level descriptors

Competency	Below Basic	Basic	Above basic	Excellent
Count in multiples of 2, 3, 4, 5, 10, 20, 30, 50, and 100 (wn)	Count in multiples of 2, 5, and 10	Count in multiples of 2, 5, 10, 20, and 100	Count in multiples of 2, 3, 4, 5, 10, 20, 30, 50, and 100	Count in multiples of 2, 3, 4, 5,10 20, 30, 50, and 100

This information helps teachers to decide what targeted content and pedagogy supports s/he needs to provide in order improve learner performance. In order to have the teachers utilize the reports most for improved classroom purposes, it is necessary that teachers have complete understanding of the reports. DNEA provides thorough training to regional administrators (Subject Advisors and Inspectors) on interpretation and usage of the reports; regional administrators then conduct a number of training workshops on the same topic for teachers and head teachers in their respective regions, in the presence of DNEA's representatives.

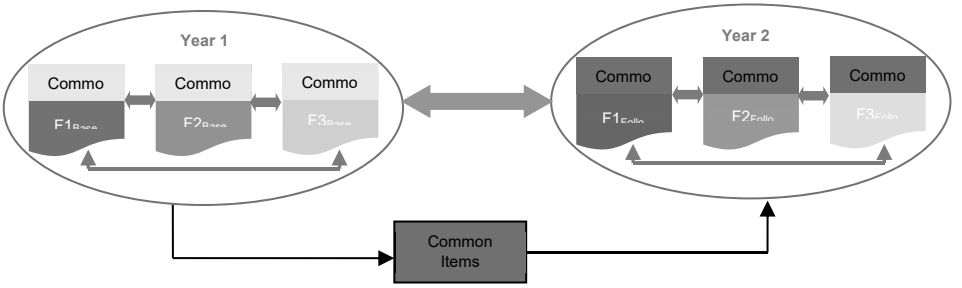
Maintenance of the scale for monitoring school performance

In order to monitor each school's progress longitudinally and accurately, it is important that difficulty level of tests in the baseline and the subsequent years must be made equivalent. Although constructing equivalent tests is not easy (literally impossible), they can be brought on the same measurement scale for calculating equivalence of scores between the baseline and follow-ups through one-to-one mapping. The procedure is called **equating** of test scores (Kolen & Brennan, 1995). Through test equating procedure, we answer to a question, if a learner is taking a Grade 5 test in 2011, what would have been his/her score on the baseline if he/she had taken the baseline test in 2009? Learners' scores in 2011 are reported after accounting for the difference in the difficulty level of the tests in 2009 and 2011.

Equating design and method

Given the objectives and design (i.e., three different test forms in each subject and year) of the SATs, a common-item nonequivalent group (internal) equating design and IRT-based mixed common item parameter approach was utilized. In other words, learners taking SATs in Year 1 and Year 2 are considered to be nonequivalent groups when equating is designed (Figure 4).

Figure 4: A Visual Representation of Common-Item Nonequivalent Groups Design



It is revealed from the visual representation of the SATs' equating procedure, (Left Panel – Year 1) there are three-test forms used in the baseline and each has a subset of common items (yellow) and a subset of unique items (green); the red lines represent the equating procedure implemented to bring all three baseline forms on the same measurement scale. Similarly, (Right Panel – Year 2) there are three test forms in the follow-up and each has a subset of common items (brown) that come from the baseline test forms (irrespective of common or unique items in the forms) and a subset of unique items (purple); the red lines showing the equating procedure implemented to bring all three follow-up forms on the same measurement scale. (Middle Panel) the blue line represents the equating procedure implemented to bring the baseline and follow-up test forms on the same measurement scale.

DNEA utilizes a fixed common item parameter (FCIP, Kim, 2006), a two-step calibration method. In this method, the parameters of its common items are fixed at the estimates obtained through the calibration of the reference test (i.e., baseline). As a result, the equated test score distribution is placed on the reference test scale.

RESULTS AND DISCUSSION

This section presents the national findings of the SATs on Grade 5 and 7 learners' performance in English Second Language, Mathematics and Science for the past 5 years. Before presenting learner performance on the SATs, it is important to review the reliability coefficients of internal consistency (Cronbach, 1951) for the reported test instruments used in latest test (2015).

Reliability coefficients

Internal consistency is usually measured with Cronbach's alpha, a statistic calculated to indicate how closely related a set of items (questions) are as a group. A high value of alpha (*α)* is often used as evidence that the items measure the same underlying construct. A commonly acceptable rule of thumb for describing internal consistency indicates that $0.7 \leq \alpha < 0.9$ is good for low-stakes testing (George & Mallery, 2003). For Grades 5 and 7 SATs for 2015, the reliability coefficients for test instruments were estimated at 0.74 – 0.89 (Figure 5).

Figure 5: Reliability Coefficients for Grade 5 & 7, 2015 instruments

Forms	English 2nd Language	Mathematics	English 2nd Language	Mathematics	Natural Science
Form A	0.85	0.84	0.88	0.83	0.83
Form B	0.85	0.83	0.89	0.81	0.81
Form C	0.86	0.84	0.89	0.81	0.74
Average	**0.85**	**0.84**	**0.89**	**0.82**	**0.79**

The reliability as shown in Figure 5 can be interpreted as, for example **Grade 7 English 2nd Language Form B**, $\alpha = 0.89$ means that if a

learner who takes a test that has a reliability coefficient of 0.89 will obtain a similar score on a test of equal difficulty 89 out of 100 times. In other words, given a learner who took Form B on the Grade 7 English 2nd Language test and got a score of 30 out of 50, if such a learner takes 100 similar but different tests (with equivalent difficulty), then such learner will get about 30 out of 50 in 89 of the 100 tests. Therefore, the learners' true ability was estimated in Grades 5 and 7 English 2nd Language, Mathematics and Natural Science through the 2015 SATs in the same manner it could have been done with 100 similar tests.

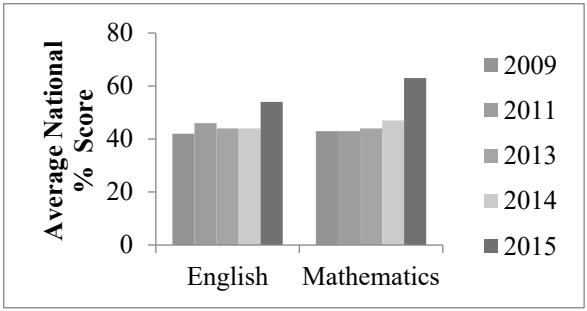
National performance

Overall Performance

The Grade 5 learners of 2015 obtained an average score of 54% (21.6 out of 40) in English 2nd Language, which shows a drastic improvement when compared to 2014 cohort of learners. In Mathematics, learners on average scored 63% (25.2 out of 40) in 2015 which shows a much bigger improvement of 16% when compared to 2014 (Figure 6). The bigger improvement could be attributed to the teacher training interventions undertaken by Ministry through the Directorate Programme and Quality Assurance as a reaction to the previous performance in SATs.

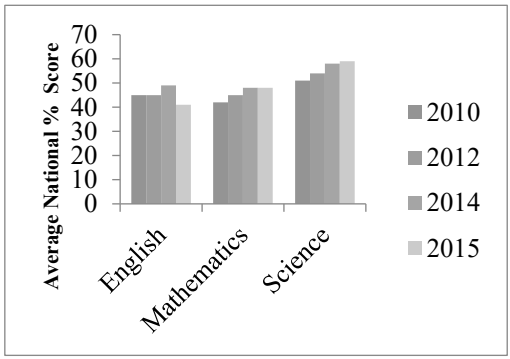
Figure 6: Grade 5 Average National Percentage Scores

Subject	2009 (baseline)	2011	2013	2014	2015
English 2nd Language	42	46	44	44	54
Mathematics	43	43	44	47	63



Furthermore, the Grade 7 learners of 2015 have shown slight improvement in Natural Science, same performance as that of 2014 in Mathematics and a significant slump in English 2nd Language. Learners obtained an average score of 41% (20.5 out of 50) in English 2nd Language, which shows a decline 8% when compared to 49% (24.5 out of 50) in 2014. In Mathematics, learners obtained an average score of 48% (24 out of 50) in 2015 which is the same as the performance in 2014 while in Natural Science the performance improved slightly to 59% (29.5 out of 50) in 2015 when compared to 58% (29 out of 50) in 2014. It is worth noting that learners have been showing continuous improvement in Natural Science since the baseline tests in 2010 (Figure 7).

Figure 7: Grade 7 Average National Percentage Scores



Subject	2010 (baseline)	2012	2014	2015
English 2nd Language	45	45	49	41
Mathematics	42	45	48	48
Natural Science	51	54	58	59

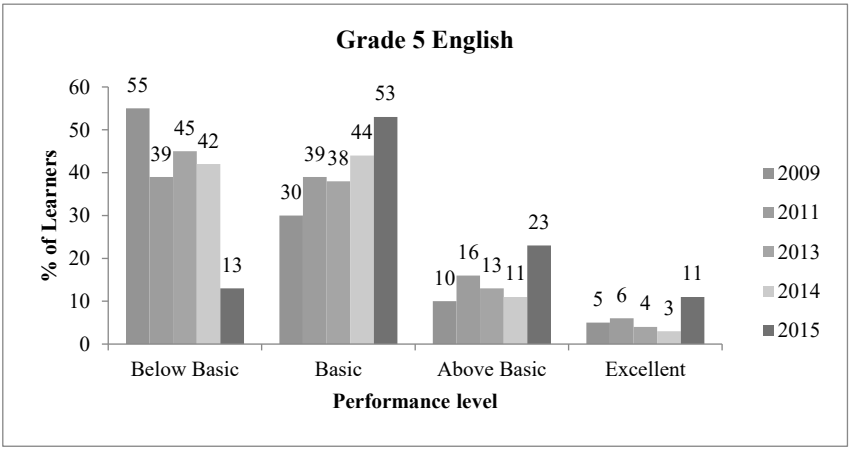
Performance Level Categories

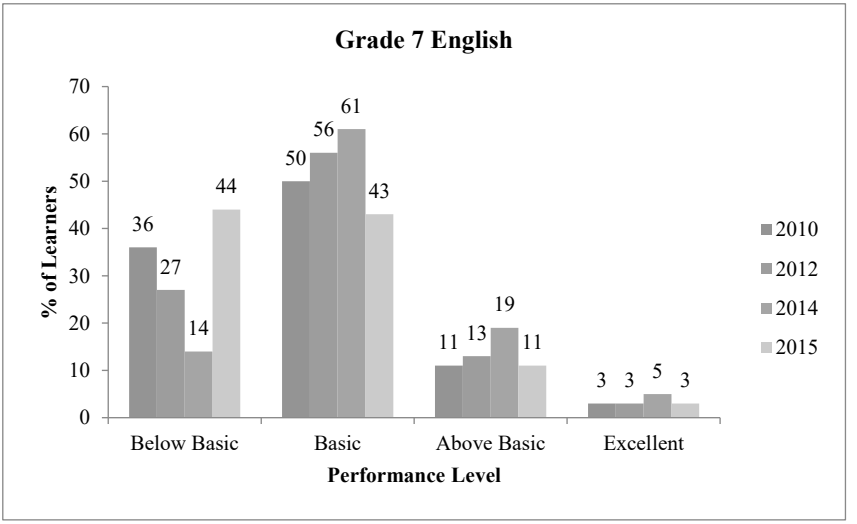
One of the major objectives of Standardised Achievement Tests is to keep track of the schools' performance from one administration to another. In this regard, learners' performance improvement from year to year is monitored through the classification of learners into four performance level categories, namely: *below basic achievement, basic achievement, above basic achievement and excellent achievement.* Ideally, the percentage of learners in the lower categories (*below basic and basic achievement)* should be continuously decreasing while those in the higher categories (*above basic and excellent achievement)* should increase to indicate improvement; the reverse indicates an undesirable trend.

In 2015, 66% of the Grade 5 learners were classified under *below basic* and *basic achievements* categories nationally in English compared to an average of 83% for the past four years. Although 66% is still high, it shows nonetheless that there was a significant decrease of in the percentage of learners falling in these two lowest performance level categories. This is indeed a positive development since learners need to move to advanced categories. Similarly, the percentage of learners classified under above basic and excellent categories has increased significantly (34%) compared to an average of the previous years.

For Grade 7, 87% of learners who took the English test were classified into *below basic* and *basic* achievement categories compared to an average of 81.3 % for the previous year (Figure 8). This is a worrying development as it was expected that the number of learners falling into these lower categories decreases with the subsequent cohorts of learners.

Figure 8: Classification of learners in Performance Level Categories in



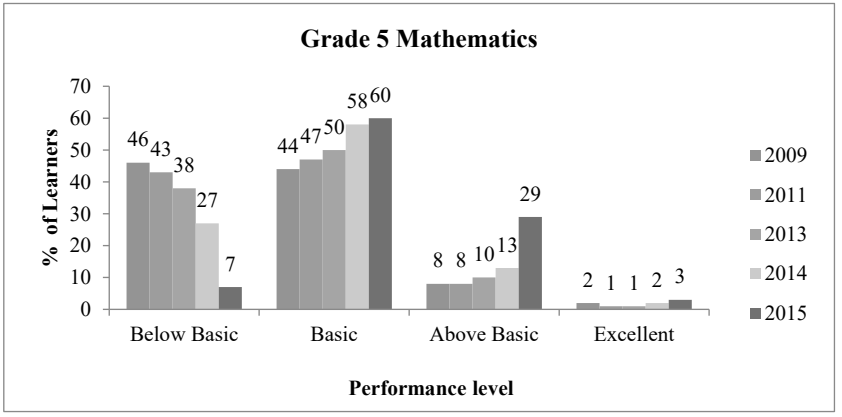


English

For Grade 5 Mathematics, the concentration of learners in the lower categories (*below basic and basic*) has decreased to 67% nationally in 2015 when compared to an average of 88.3% for the previous years. This indicates a significant improvement in the performance of the 2015 cohort as the percentage of learners in the higher performance level categories (*above basic and excellent*) has increased to 32% which is much higher than ever recorded in the previous years (Figure 9).

In Grade 7, a similar trend in performance as in English (Figure 8) is observed. In 2015, 84% of the learners were classified in the lower achievement categories nationally compared to an average of 82% recorded in in the previous years (Figure 9). The increase in the number of learners in the *below basic* and *basic* categories is undesirable as it represents the undoing of the previous year's successes.

Figure 9: Percentage of Learners classified in Performance Level



Categories

The percentage of learners classified into the *below basic* and *basic* categories for Grade 7 Natural Science continues to show improvement since the baseline in 2010. In 2015, only 49% of the learners were classified into the two lower categories compared to an average of 68.3% in for the previous years (Figure 10). This indicates that the number of learners falling in lower performance level categories is continuously decreasing as expected. The percentage of learners classified in the two higher categories has increased to 51% compared to an average of 31.7 % for the previous years. Natural Science continues to demonstrate the Ministry of Education, Arts and Culture's desire to have fewer learners in lower performance level categories and more learners in the higher achievement categories overtime.

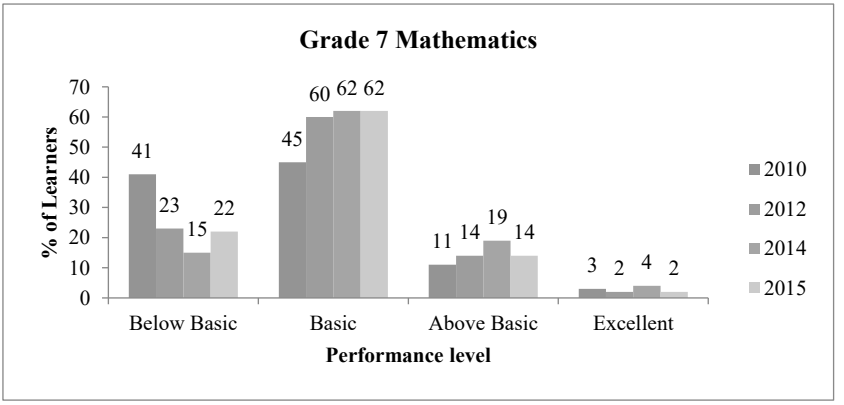
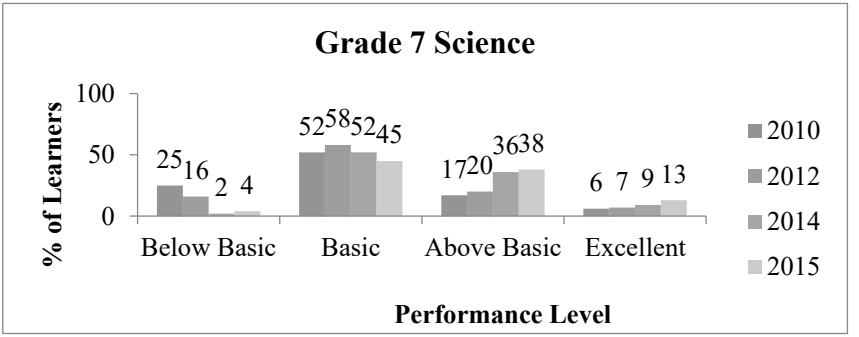


Figure 10: Percentage of Learners classified in Performance Level



Categories

Teachers' perceptions of standardized achievement tests

A perception survey was conducted amongst primary school teachers in Namibia (N = 130), with years of teaching experience within two categories namely, 0 to 3 years and more than three years. Majority (66.2%) of the respondents have teaching experience of more than 3 years. Hence, it can be assumed that they have been exposed to the SATs' information within the last three to four years. This is the period in which the SATs have been implemented in primary school phase country wide. Figure 11 below shows how teachers responded to nine of the indicators on the questionnaire. The questionnaire comprised of ten indicators, with three levels to choose from, namely, *Agree*, *Not sure* or *Disagree* and the tenth indicator required respondents to choose between *Regularly*, *Rarely* or *Not at all*.

Figure 11: Teachers' responses to the questionnaire

Indicators	Agree	%	Not sure	%	Disagree	%	Total	% Total
SATs provide quality feedback for improving teaching and learning	120	92.3	6	4.6	4	3.1	130	100
SATs results help teachers identify competencies learners find difficult to learn.	123	94.6	5	3.8	2	1.5	130	100
SATs help monitor the performance of learners from year to year	115	88.5	12	9.2	3	2.3	130	100
SATs school report is clear and understandable	112	86.2	18	13.8	0	0	130	100
SATs school report contains sufficient information needed by the teachers	104	80.0	25	19.2	1	0.8	130	100
All the competencies tested as shown in the SATs reports are in the syllabus	115	88.5	10	7.7	5	3.8	130	100
SATs school report is useful for the planning of teaching and learning	120	92.3	10	7.7	0	0	130	100
SATs has improved the quality of teaching	78	60.0	47	36.2	5	3.8	130	100
SATs are relevant in Namibia and should continue	118	90.8	11	8.5	1	0.8	130	100

As evident in Figure 11, teachers' responses revealed a largely positive perception about SATs, with an average of 85.9% agreeing with the indicators. Also important to note from the responses was the overwhelming majority of the respondents agreeing with notion that SATs as a national assessment is relevant to the Namibian education system and hence the need for it to continue providing diagnostic information about learners' ability.

The tenth indicator required respondents to indicate how often they used the SATs reports and the responses are summed up in Figure 12 below.

Figure 12: Responses on how often respondents used SATs reports

	Frequency	Percent
Regularly	82	63.1
Rarely	41	31.5
Not at all	7	5.4
Total	130	100.0

Although most of the respondents indicated that they use the SATs report regularly, there was a significant number of respondents who indicated that they rarely used the SATs report for varying reasons. These reasons ranged from lack of incentives to compel teachers to use the SATs report as it is low-stakes in nature to lack of support from school management in popularizing the SATs in schools. In some instances, SATs reports do not reach schools early enough due to logistical issues in the regions, despite the early release by the DNEA.

CONCLUSION

The Namibian SATs were developed based on best practices that not only provide reliable and valid diagnostic information about what learners know and are able to do in key learning competencies in the curricular but also assist the Ministry of Education to keep track of each school's progress from one year to the next. When developing the SATs, particular attention was paid to: (1) SATs as system vs. an isolated activity; and (2) expedited reporting of the assessment results

back to school within a reasonable time frame. Each of these aspects is discussed in the following.

SATs as system vs. isolated activity: In many developing countries, national assessment is viewed as an isolated routine activity; it lacks coherence and congruence with other educational initiatives for improving learner learning outcomes. In many occasions, although the national assessment data is used for decision making process, trustworthiness of the data is very limited; either data being manufactured during collection or wrong analysis being used for producing results. Therefore, SATs were designed as an integrated and robust assessment system that is not only coherence with other educational initiatives by ministry or other developing partners, but also has involved all stakeholders (e.g., University, other Education Directorates, Regional Administrators, Schools, etc.) within the education enterprise (Abdullah & Nyambe, 2013).

Expedited reporting of the assessment results: Reporting assessment results back to schools or to the system has always been the issue in most countries. It takes about 2-4 years in producing national assessment results in many countries; effectiveness of the results diminishes as people seems losing their interest and some of the assumptions made during assessment design may not be even valid any longer. Hence, an automated reporting system in SPSS by which DNEA can use to produce school reports within two months of test administration. For example, if the test is administered in taken in November, 2015 (end of school year) schools receive assessment reports by February, 2016 just a few weeks after schools reopened for new school calendar year so that teachers can still utilize the reports for improved classroom instruction and lesson planning.

No matter how soon we produce reports and what psychometric procedures we use, nothing will help improving learner learning outcomes until teachers, head teachers, and the system understand and use the reports, and make necessary intervention plans for

improving teachers' classroom performance. Assessment is just a tool that only provides reliable and valid information for making policy decisions.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Beck, M. (2003). *Standard setting: If it is science, it's sociology and linguistics, not psychometrics*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16 (3): 297–334.

Ferdous, A. A., & Nyambe, C.M. (2013). Design and Development of Successful National Assessment Program: A Namibia Model. Journal of Educational Assessment in Africa, 8 (pp).

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges*. Washington, DC: Partnership of Educational Revitalization in the Americas.

George, D. & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference 11.0 update* (4th ed.). Boston, MA: Allyn and Bacon.

Haladyna, T. (2004). *Developing multiple-choice test items* (3rd Ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A review of multiple-choice item-writing guidelines for classroom

assessment. *Applied Measurement in Education, 15(3),* 309-334.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Hammond, D. & Wentworth, L. (2010). *Benchmarking Learning Systems: Student Performance Assessment in International Context*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and Practices*. New York: Springer-Verlag.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

Marope, M. T. (2005). *Namibia human capital and knowledge development for economic growth with equity*. Africa Region Human Development, Working paper series No. 84. The World Bank.

National Institute for Educational Development (2008). *The national curriculum for basic education*. Ministry of Education.

Office of the President. (2004). *Namibia Vision 2030: Policy framework for long-term national development*. Windhoek, Government of the Republic of Namibia.

Parie, M. (2008). *A guide to understanding and developing performance-level descriptors*. Educational Measurement: Issues and Practices, 27(4), 15-29.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: Macmillan.

Plake, B. S., Ferdous, A. A., & Buckendahl, C. W., (2005). *Setting multiple performance standards using the Yes/No method: An alternative item mapping method*. Paper presented to the meeting of the National Council on Measurement in Education, Montreal. Canada.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Ravela, P. (2005). A formative approach to national assessments: The Case of Uruguay. *Prospects* 35(1): 21-43.

Wolfaardt, D. (2003). *Symposium proceeding: The influence of English in the Namibian examination context. Namibian Ministry of Basic Education, Sport, and Culture.*