IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388,p-ISSN: 2320–737X Volume 6, Issue 4 Ver. I (Jul. - Aug. 2016), PP 65-75
www.iosrjournals.org

Vertical Scaling in Standards-Based Educational Assessment and Accountability in Educational Systems

Dr Nsikak-Abasi Udofia[1], Mary Patrick Uko[2]
[1,2]Department of Educational Foundations, Guidance and Counselling,
University of Uyo, Uyo, Nigeria

Abstract: *The purpose of this paper is to inform state policy-makers and assessment and accountability specialist about vertical scaling methods and the potential advantages of vertical scaling in standard-based educational accountability in educational system. The possible application of vertical scale to support growth model is carried out. Final Test scores of SS1 and SS2 students for four consecutive years from 3 secondary schools was use to realize this. A statistical analysis of moving average was used to interpret growth in the student performance over time. The result revealed that growth can be tracked down over years across class levels, with the implementation of vertical scaling.*

Keywords: *Measurement; Vertical Scaling; Academic Performance; Achievement Testing*

I. Introduction

The Nigerian Educational Research and Development Council (NERDC) has the mandate to develop curricular to meet the targets of the reform in the context of National Economic Empowerment and Development Strategies (NEEDS) and the Millennium Development Goals (MSG). Following the Federal Government reforms in education and the need to attain the Millennium Development Goals (MDGs) and the critical targets of the National Economic Empowerment and Development Strategies (NEEDs), which can be summarized as value-reorientation, poverty eradication, job creation, wealth generation and using education to empower the people, it has become imperative that the existing curricula for Senior Secondary School should be reviewed and re-aligned to fit the reform programme. The National Council on Education (NCE) at its meeting in Ibadan in December 2005 directed the Nigerian Educational Research and Development Council (NERDC) to carry out this assignment. Between January 2007 and March 2008, the Nigerian Educational Research and Development Council convened a meeting of experts and also organized several workshops to produce the senior secondary school curriculum, which would ensure continuity and flow of themes, topics and experiences from Junior Secondary one to three and Senior Secondary one to three levels.

The implementation of common educational assessment measures (such as the National Common Entrance Examination, the West African School Certificate Examination, General Certificate in Education, the Joint Admission and Matriculation Board Examination and of course, Internal Joint Admission and Matriculation Board Examination) in the Nigerian educational system are essential in the development of the world finest education system. Such policy is believed to motivate the children, lift some students to world class standards, help increase the national productivity and contribute to the restoration of our global competitiveness.

An assessment system that assesses student performance at different class or class level based on public adopted standard of what is to be taught is a standard - Based Assessment. A standard- Based assessment is designed to hold schools publicly accountable for each students meeting those high standards. Often, standards-Based Assessment Systems have different levels of achievement that defines performance categories. Standards are the centerpiece of MDGs therefore a method should be employed to set this standard that will cut across the different class level assessment, for proper accountability and measures of growth of individual or group of students over time. A standard-based test addresses standard in two ways first, the test questions pertain to a particular set of content standards that is statement of objectives defining the domain of knowledge and skills to be learned and assessed at a particular class and in a particular subject. Secondly, result are reported in context of performance standard which is relative to various threshold scores, which creates test-score ranges that corresponds to different categories or levels of performance.

Student's performance relative to established thresholds or "cut-scores" is the focus of the reporting of result in standard-based program. Nom-reference testing addresses common denominator content and skills; they are not designed to comprehensively cover a set of content-standard that may be used by state or local school Authority. In other words, nom-reference Test do not measure how well schools are teaching or students are learning the material defined by relevant content standard. Vertical scaling can be conceptualized as measurement process that models latent variables estimate derived from set of test forms of increasing difficulty and place the ability estimate in appropriate relation so that comparison may be made for examinees taking forms of different difficulty. The scores on a vertical scale, when properly constructed represents unit on a single, equal-interval scale applied across all class levels.

Modern educational assessment programs are used for a variety of purposes: to improve student learning of content standards through improved instruction based on the assessment results; to complement curriculum or teaching methods; to inform teachers/students of their progress; to inform the public about school performance; to be used as a guide in decision making about students, teachers, or schools; and to provide various data comparisons [1].

The information that serves these purposes is derived from individual tests that make up an assessment program. The purpose of any particular assessment, however, is more specific-that is, its purpose is to give users an accurate description of what students know and are able to do. The most important characteristic of any assessment procedure is its impact on validity. Standards for Educational and Psychological Measurement define validity as “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests” [2].

The creation of a set of tailored instruments that measure educational achievement for groups of systematically different abilities is the goal of vertical scaling. This paper describes vertical scaling methods and helps examine the question of whether the resulting changed measures can measure changes in individuals over years with enough accuracy to support the relatively high stakes that accountability testing necessarily entails. It explains what vertical scales are, how they are constructed, and what their advantages are. Specifically, this study is designed to:

1. To investigate whether the curriculum content standard is vertically articulated and if the content standard permits meaningful interpretation of a vertical scale.
2. To determine the classes to be included in the vertical scale and to what extent the overlapping should be present between class levels?
3. To determine the specifications that should describe growth in vertical scaling?
4. To explore the effect of vertical scaling on the maintenance of standard in schools assessment program.
5. To investigate the effects of vertical scaling on content systematic progression across classes.
6. To explore ways of interpreting vertical scale scores with regards to students performance level, when the scores are not ordinal across classes.

The creation of a set of tailored instruments that measure educational achievement for groups of systematically different abilities is the goal of vertical scaling. This paper describes vertical scaling methods and helps examine the question of whether the resulting changed measures can measure changes in individuals over years with enough accuracy to support the relatively high stakes that accountability testing necessarily entails. It explains what vertical scales are, how they are constructed, and what their advantages are.

Vertical scaling is the term used for the process of linking assessments to describe student growth over time. Although the methods can be complex, the goal is quite simple: to create a framework and metric for reporting the educational development of individuals and groups. The challenge of vertical scaling of assessments has existed since the first use of standards-based assessments to measure individual and group progress [3]. Today, vertical scaling is needed for assessments of growth toward college and career readiness standards and for adaptive testing. In these applications, comparative information about results from assessments of different levels of difficulty is needed to build a vertical scale. When the assessments are aligned to appropriate cross-class content standards, valid use of test scores as indicators of growth based on standards can he achieved. Points on a vertical score scale are a kind of cognitive map to the future. They provide a basis for describing progress, setting goals, and ultimately determining whether students are on track for college and career readiness.

Vertical scale places the scores of different tests onto a common metric so that, in principle, comparisons can be made between scores that span through different class levels. One of the main motivations to the development of these vertical scales is to support direct. Inferences about student growth. Because each students has a different background and capabilities, teachers, administrators and infact educators should use all appropriate management tools, techniques and practices which will cause the state’s educational program to be more effective. Assessment test should serve as the measure for the state wide school accountability system and growth modeling.

Vertical scales result from the application of a number of psychometric procedures, and understanding vertical scales, their usefulness and limitations, requires an understanding of several of these key building blocks. In particular, the fundamental components that the review for this vertical scaling discussion include: Measurement Model,Scaling, Equating, Linking, Standard setting, Appropriate contexts for vertical scaling, Characteristics of well-constructed vertical scales, Advantages of vertical scale and some recent development on Vertical Scaling Technology.

1.1. Appropriate Contexts for Vertical Scales

Although the statistical and psychometric scaling and linking procedures required to produce vertical scales might be employed in any of a wide variety of contexts, meaningful interpretations about growth and progress will only be supported when the test forms and the student populations involved have certain characteristics. Reference[3] described the general requirements in this way: "When differences in population proficiency at adjacent levels are modest in comparison to differences between examinees within levels, and when the expectations or standards against which examinees are to be measured overlap extensively, then linking the adjacent test levels to a common scale will make sense and help to provide meaningful information." They note that these conditions are generally well-satisfied in achievement test batteries measuring proficiency over a range of adjacent class levels in broad domains such as reading and mathematics.

A significantly more refined examination of the appropriate contexts for conducting vertical scaling is the "vertical alignment" work sponsored by CCSSO's Technical Issues in Large-Scale Assessment (TILSA) State Collaborative on Assessment and Student Standards (SCASS). In particular, reference [4] describes a systematic approach to assessing and creating the necessary conditions for meaningful vertical scales. It is simply noted here that some sets of content standards are clearly more amenable to vertical scaling than others. In North Carolina, for example, the K-5 Language Arts content standards are based on common goals "in order to provide continuity of language study and increasing language skill development" [5].

By contrast, North Carolina's Social Studies content standards indicate that the state's history and geography will be the focus of class4, whereas United States history is the focus of class5. Given these characterizations, one would expect a vertical scale to provide meaningful information for language arts but not social studies in North Carolina. As states add testing at a high school class to their classes 3-8 testing programs, it is particularly appropriate to examine the degree to which the high school content standards align with those of class8 before making a commitment to place the high school test on a 3-8 vertical scale.

1.2. Advantages of Vertical Scales

Vertical scaling may bring several compelling features to achievement tests.

1. Vertical scales facilitate the estimation and tracking of growth over time, as repeated measures (i.e., comparable scale scores) on individual students using different, age- and class appropriate test forms becomes possible. This should help determine how much growth has occurred over time and in different regions of the proficiency range.
2. Second, it would appear that vertically-scaled achievement tests allow comparisons of one class level to another and one cohort of students to another at any point in time.
3. Vertical scaling of test forms also enables important comparisons regarding test items.
4. Vertical scaling can lead to more efficient field testing of new content, as items targeted for one class might be found to be of more appropriate difficulty for an adjacent class, assuming that the targeted content standard is present in both classes.
5. Final form selection for a target class can then identify appropriate items from a larger pool, when each item in the pool has parameters on a common scale.
6. In addition, as noted above, standard setting judgments can be made more developmentally appropriate.
7. The standards may be made more precise in theory, since a richer set of items (from adjacent levels of the test) may be ordered and the scale may thus be more finely segmented as the density of items increases.

1.3. Characteristics of Well-Constructed Vertical Scales

In order to construct vertical scales that will support meaningful interpretations, several elements are required. A state must have a set of vertically-aligned content standards with considerable class-to-class overlap and a systematic, intentional increase in difficulty. [6] A robust vertical scaling design specifying the psychometric procedures and data collection approaches to be used, including sufficient numbers of common items across levels or sufficient numbers of students taking multiple forms, is needed. It is highly desirable that the data collected to create the vertical scales be gathered during an operational administration or under conditions closely approximating the operational conditions, and that statewide data or large, statistically representative samples of students be involved in the vertical scale data collection.

When vertical scales have been well constructed for use in large-scale educational testing programs, one would expect to see a number of technical characteristics. Since the forms are intended to progress in difficulty, one should see evidence that this has been achieved. [4]

Test characteristic curves, for example, should show evidence of increasing difficulty across classes. For sufficiently large and diverse samples of students, scale score means 18 would be expected to increase with class level, and the pattern of increase would be expected to be somewhat regular and not erratic. Erratic or non-increasing patterns of mean scale score growth or large differences in scale score standard deviations from one class to the next would warrant special scrutiny. When scrutinizing such results, it is important to consider as fully as possible the context of the test (e.g., Are the stakes lower in some classes than other? Are there significant changes in the curriculum at certain classes?), as well as all of the psychometric and statistical issues.

States should expect that all information pertaining to their vertical scales would be well documented in an appropriate technical report. Statistics reported should also include correlation of adjacent-level test scores (under the common examinee design) and/or correlation of item difficulties across test levels in the common item design. High degrees of correlation suggest that the examinees and/or items would be ordered the same way on adjacent test levels, which may be taken as a degree of validation that vertical scaling is appropriate. The specific methods employed, the data collection design, appropriate descriptive statistics regarding test items and groups of examinees, and the resulting scale score patterns should be thoroughly documented.

1.4. Recent Developments in Vertical Scaling Technology

Vertical scaling has seen a resurgence of interest in recent years, perhaps much of it attributable to the change in federal education policy in the United States that resulted in many states adopting new testing programs for classes 3-8.

For example, reference [7] examined approaches to modeling cross-class growth trajectories in a hierarchical modeling framework. The approach may support the identification of more regular underlying growth patterns than those that may be inferred by fitting simpler IRT models and examining the resulting class-by-class trends.

Reference [8] also examined the application of multidimensional IRT models to vertical scaling problems, as have reference [9]. It is widely recognized that the assumption of unidimensionality underlying standard IRT measurement models is an over-simplification of reality, and these recent investigations into multidimensionality in vertical scaling data support this observation. Nonetheless, it is not yet clear whether or how more complex multidimensional models might bring greater validity to the essentially unidimensional classification decisions required by standards based accountability testing. This remains an area rich with research possibilities. Reference [10] added a chapter on vertical scaling to their measurement text, and they include a discussion of several ways to evaluate the appropriateness and validity of vertical scaling results. Reference [11] applied the Kolen and Brennan criteria when examining vertical scaling results under a variety of model estimation and forms linking strategies.

1.5. Research Questions

The following Research Questions were raised to guide the study.

1. Is the curriculum standard or content standard vertically articulated and to what extent does the content standard permit meaningful interpretation of a vertical scale?
2. What class levels should be included in the vertical scale, and how much overlapping should be present between class levels?
3. What are the specifications that should describe growth in vertical scaling?
4. What is the effect of vertical scaling on the maintenance of standard in school assessment program?
5. What is the effect of vertical scaling on content systematic progression across classes?
6. What is the effect of vertical scaling treatment on the student performance level across the class levels?

II. Research Methodology

2.1. Design

The research design adopted for this study is Longitudinal Survey Design. The design permits the measurement of difference or change variable from one period to another which therefore makes it convenient for the researcher to adopt the design.

2.2. Area and Population

The study area is in Akwa Ibom State, Nigeria. The population of this study is made up of all Senior Secondary One and Two (SS1 & SS2) students offering English Language, Biology and Mathematics in the 23 Secondary Schools in Abak Local Government Area of Akwa Ibom State comprising a total of seven thousand, two hundred (7,200) students in secondary schools in Abak Local Government Area.

2.3. Sampling

Simple random sampling of schools and class levels within schools was employed to acquire samples of approximately 779 students who were administered off- class level items in each year of the study per class level and content area.

2.4. Instrumentation

The instrument used in this study was the final term or promotion examination score of the student in Mathematics, Biology and English Language. The test score revealed strength and weaknesses of the student Mathematics, English and Biology abilities based on detail analysis of the specific skills involve in their successful performance and a thorough study of the growth pattern made by students in a particular year was carried out.

On-class level English Language, Biology and Mathematics operational test items were administered to all classes SS1 to SS2 students during regular test administration in each year of the study. The off- class level linking items selected from the adjacent class operational assessments were administered to samples of the same students during a separate administration approximately two weeks after the operational test administration. Approximately ten items from the level below and ten items from the level above operational test were administered to SS1 students. Fifteen items from SS1 test and SS3 test were administered to SS2 students.

Linking items were selected to conform to the learning standards assessed in class in which they were administered. The on-class level assessments contained multiple-choice (MC) and constructed response items (CR). The off-class level linking item sets comprised of MC items only selected from the WAEC, NECO, JAMB and JSSCE past questions. These off-class level items were used for linking adjacent classes but did not contribute to the test scores.

2.5. Data Preparation

Vertical scales for the SS1 and SS2 English language, Biology and Mathematics assessment were developed using item response they model known as The Rasch Partial-Credit model (RPCM). The RPCM which is an extension if the Rasch one-parameter item Response model was use to place test items and measures of students proficiency on the same scale across assessment. The RPCM was so preferred because it maintains a one-to-one relationship between Rasch-based performance estimate (Q), scale scores, and raw scores, meaning each raw score is associated with a unique scale score.

III. Result

3.1. The Vertical Scale Score Analysis

The result of the final analysis of growth over 4years is presented in table 3 and 4 followed by discussion of findings.

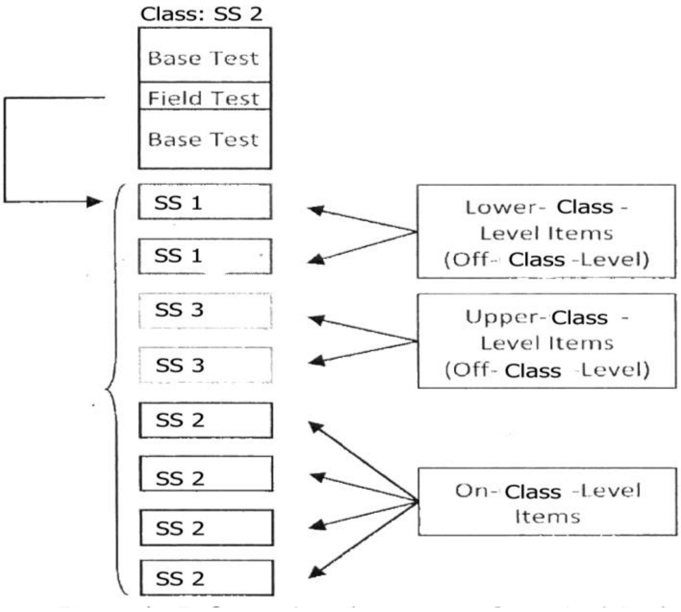
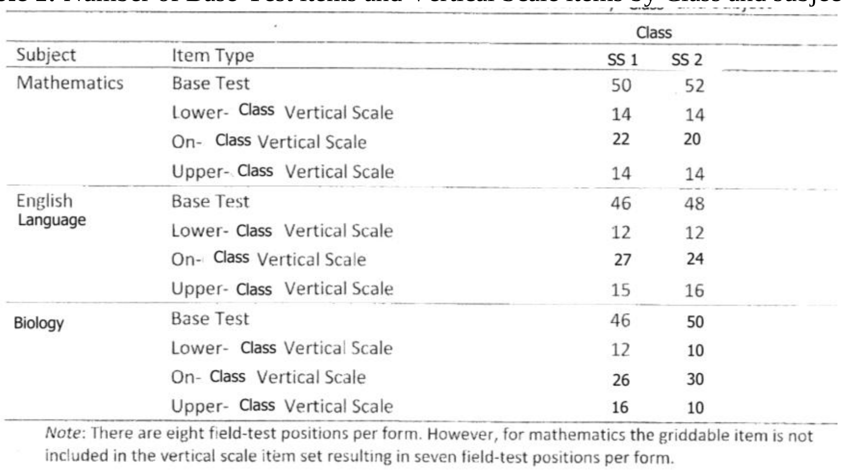


Figure 1: Example referencing the types of vertical scale items

English, Biology and Mathematics components of State Assessment Test based on the WAEC, NECO and AKSSSPE Standards were examined. Commonly referred to as Akwa Ibom State Secondary School Promotion Exams (AKSSSPE) Assessment Test and Promotion Exams score for SS I on English, Biology and Mathematics are the criterion-referenced components of the statewide assessment program.

Table 1: Number of Base-Test items and Vertical Scale items by Class and subject



		Class	
Subject	Item Type	SS 1	SS 2
Mathematics	Base Test	50	52
	Lower- Class Vertical Scale	14	14
	On- Class Vertical Scale	22	20
	Upper- Class Vertical Scale	14	14
English Language	Base Test	46	48
	Lower- Class Vertical Scale	12	12
	On- Class Vertical Scale	27	24
	Upper- Class Vertical Scale	15	16
Biology	Base Test	46	50
	Lower- Class Vertical Scale	12	10
	On- Class Vertical Scale	26	30
	Upper- Class Vertical Scale	16	10

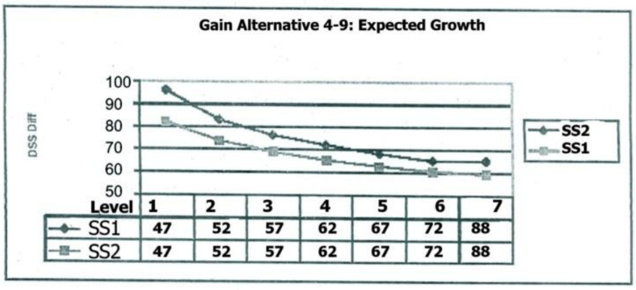
Note: There are eight field-test positions per form. However, for mathematics the griddable item is not included in the vertical scale item set resulting in seven field-test positions per form.

The resulting Developmental Scale Scores (DSS) or Vertical Scaling Score (VSS) was used for analysis.

Figure 2 shows the relationship between these scores and the corresponding Achievement Levels by content area. This level were grouped under the five achievement or proficiency levels as follows:

Level 1	**Level 2**	**Level 3**	**Level4**	**Level 5**	**level 6**
0-39	40-44	45-49	50-54	55-59	60-64
Level 7	**Level 8**	**Level 9**			
65-69	70-74	75-100			

Levels 1 to 3	-	Below Basic Academic Performance
Levels 4 to 6	-	Basic Academic Performance
Levels 5	-	Satisfactory Academic Performance
Levels 8	-	Proficient Academic Performance
Levels 9	-	Advance Academic Performance



Level	1	2	3	4	5	6	7
SS1	47	52	57	62	67	72	88
SS2	47	52	57	62	67	72	88

Figure 2: State Assessment Test (SAT) and WAEC Achievement Levels for the Developmental Scale
Source: *WAEC Grading*

To achieve this, a moving average was calculated on the mean scores difference obtained from AKSSSPE and SS I Promotion Exams scores for each year. This was used to compare with the vertical scaled score or standard score of the achievement levels to find out range of growth of students in each year.
The result of the mean difference and comparison using moving average is presented on table 2 to 3

3.2. English

Sample sizes and means for class level aggregate data for the Comprehensive Assessment Test State Standards (AKSSSPE) English Vertical Scale Scores (VSS) are shownin Table 2.

Table 2: Mean Aggregate English VSS for Performance of Students making Gains by Achievement Level 4 Criteria

Class	Year	Number of Students	Mean
SS2	2011	569	71.8
SS1	2010	569	43.5

SS2 SS1	2012 2011	477 477	93.6 30.5
SS2 SS1	2013 2012	557 557	102.1 128.3
SS2 SS1	2014 2013	574 574	105.0 28.3

The mean difference and comparisons to expected growth for class level aggregate data for the SAT English VSS are shown in Table 3.

Table 3: Mean difference and comparisons to expected growth for class level aggregate data for the SAT English VSS

Year & Class	No. of Student	Mean Difference	VSS Expected	Met Expectation
SS1 – SS2 2010-2011	569	28.3	62	No
SS1 – SS2 2011-2012	477	6.31	62	Yes
SS1-SS2 2012-2013	557	-26.2	62	No
SS1-SS2 2013-2014	574	76.7	62	Yes

3.3. Biology

Sample sizes, means, and standard deviations for class level aggregate data for the SAT Biology VSS are shown in Table 4

Table 4: Mean Aggregate Biology VSS for performance of Students making Gains by Achievement Level Criteria (2011 to 2014).

Class of Students	Year	Number	Mean
SS2 SS1	2011 2010	569 569	195.0 96.3
SS2 SS1	2012 2011	477 477	78.9 33.4
SS2 SS1	2013 2012	557 557	164.8 64.5
SS2 SS1	2014 2013	574 574	158.6 53.2

The Mean difference and comparisons to expected growth for class level aggregate data for the SAT Biology VSS are shown in Table 5.

Table 5: Mean difference and comparisons to expected growth for class level aggregate data for the SAT Biology VSS

Year &Class	No.ofStudent	MeanDifference	VSSExpected	MetExpectation
SS1 – SS2 2010-2011	569	98.7	62	Yes
SS1 – SS2 2011-2012	477	45.5	62	No
SS1-SS2 2012-2013	557	100.3	62	Yes
SS1-SS2 2013-2014	574	105.4	62	Yes

3.4. Mathematics

Sample sizes, means, and standard deviations for class level aggregate data for the SAT Mathematics VSS are shown in Table 6

Table 6: Mean Aggregate Mathematics VSS for Performance of Students by Achievement Level Criteria from 2010 to 2014.

Class	Year	Number of Students	Mean
SS2 SS1	2011 2010	569 569	136.7 43.8
SS2 SS1	2012 2011	477 477	98.6 108.3
SS2	2013	557	143.3

SS1	2012	557	55.9
SS2 SS1	2014 2013	574 574	158.7 73.8

The mean difference and comparisons to expected growth for class level aggregate data for the State Assessment Test (SAT) Mathematics Vertical Scale Score (VSS) are shown in Table 7.

Table 7: mean difference and comparisons to expected growth for class level aggregate data for the State Assessment Test (SAT) Mathematics Vertical Scale Score (VSS)

Year &Class	No. ofStudent	MeanDifference	VSSExpected	MetExpectation
SS1 – SS2 2010-2011	569	92.9	62	Yes
SS1 – SS2 2011-2012	477	-9.7	62	Yes
SS1-SS2 2012-2013	557	87.4	62	Yes
SS1-SS2 2013-2014	574	84.9	62	Yes

The study was combining data across multiple years for improvement- based approach. Rolling average was used in comparing the average scores of 2011, 2012, 2013 with the average score of 2012, 2013 and 2014. Percentage of proficient is given as a mean score diff x Expected Growth.

Table 8: English

Year	2011	2012	2013	2014
Mean score difference	28.3	63.1	-26.2	76.7
% Proficient	17.5%	39.1%	16.2%	47.6%

Using 3 years rolling average, the percentage proficient is compared for 2011-2012, 2013 and 2012, 2013, 2014.
The average of 2011, 2012 and 2013 = 24.3% proficient
The average of 2012, 2013 and 2014 = 34.3%
There is a 10% gain.

Table 9: Biology

Year	2011	2012	2013	2014
Mean score difference	98.7	45.5	100.3	105.4
% Proficient	61.2%	28.2%	62.2%	65.3%

Using 3 years rolling average, the percentage proficient is compared for 2011-2012, 2013 and 2012, 2013, 2014.
The average of 2011, 2012 and 2013 = 50.5% proficient
The average of 2012, 2013 and 2014 = 51.9.3%
There is a 1.4% gain.

Table 10: Mathematics

Year	2011	2012	2013	2014
Mean score difference	92.9	-9.7	87.4	84.9
% Proficient	57.2%	6.0%	54.2%	52.7%

Using 3 years rolling average, the percentage proficient is compared for 2011-2012, 2013 and 2012, 2013, 2014.
The average of 2011, 2012 and 2013 = 39.1% proficient
The average of 2012, 2013 and 2014 = 37.6%
There is a 1.5% gain.

IV. Analysis Approach

Components of both the School Accountability System, as envisioned by the MDG Plan for Education, and the Akwa Ibom State Free and Compulsory Education Act of 2007 were used to approach the question of differential growth for various groups of students within the Achievement Levels of Basic and Below Basic.

Within the School Accountability System, making annual learning gains on AKSSSPE account for three categories that are used in calculating school performance level.

1. When students improve their AKSSSPE Achievement Levels (1-2, 2-3, 3-4,4-5); or (6, 7, 8 or 9)
2. When students maintain a relatively high Achievement Level (3, 4 or 5); or
3. When students demonstrate more than one year's growth within Levels 1 or 2, as measured by an increase in their AKSSSPE Developmental Scale Scores from one year to the next.

The focus of the analysis was on the students eligibility for demonstrating gains via Vertical Scaling (VS). The reason being that these students have demonstrated two consecutive years of Basic or Below Basic performance. Students that demonstrated gains via VS level have shown a substantive improvement in their performance. For example, a student that has improved from Achievement Level 3 to Achievement Level 4, has gone from a classification of Basic to Proficient Students that demonstrated gains VS level have shown consistently high performance as Achievement Level 4 or higher is classified as Proficient. To make Annual Yearly Progress (AYP), the percentage of students earning a score of Proficient or above in Biology, English and mathematics has to meet or exceed the annual objectives for the given year. For the purpose of this study the student has to meet or exceed the cut score that is used as basis for analysis. Though not all students in the analyses are non-Proficient, tracking the composition and progress of these groups could lead to an understanding that may facilitate policy and instruction.

V. Discussion

Vertical Articulation is “the interrelationship and continuity of contents, curriculum, instruction and evaluation within programs which focus on the progress of the student in learning both to comprehend and communicate. Since the content standard and was vertically articulated, there was evidence of continuity because the vertical scale was able to measure and interpret growth of the student in each of the years and across the class levels. The vertical scale align content standard and objectives related from one class to the next, knowledge or skill is extended to wider range of content, deeper understanding (cognitive processes) for the same content is gained, and new content and skills are acquired.

From the result of the vertical scale, there is sufficient overlapping items between groups of student to form the links on the scale scores. The overlapping from below and above and at least 20% of the Items was included in the test which is enough overlapping. Any class can be used as base class since the overlapping covers both below that class level and above that class levels for the vertical scale to be useful in interpreting the result. This has to be so for the gap to be closed both from the bottom and from the top for a meaningful interpretation of the score with the vertical scale.

From the result we can see a gain of 10%, 1.4% and 1.5% respectively, for each of the subject in each of the years compared. This shows that there is a significant effect of the vertical scale on the students and the measurement of growth over time. Vertical scales are used for high stakes purposes such as building students growth models and it is important that the scale have strong technical underpinnings. The process of developing the vertical scale shows that vertical scales have the set of items ministered to multiple class groups in this case SS 1 and SS 2 have items that are representative of the skill and knowledge and content area.

The gain of 1.5% for Mathematic 10% for English and 1.4% for Biology that is seen from the analysis can be used to judge the Educational standard of the schools. Therefore vertical scale helps to maintain the standard of the schools assessment programs. The result can also be used to ascertain future student achieve met, judge whether a student is on track to reach proficiency or some other achievement outcome referred to as growth-to-standard, and also determine whether student are on track to reach the designated target which is the standard set for the test program.

The achievement levels describes the extent to which students have mastered the intended content for the class levels and are ready for subsequent element of the curriculum. This mastery is a progressive process from class to class and content to content. The vertical scale score of the students informs whether the student have mastered the content in a particular class which will now qualifies the student to move to another level of content mastery. Developing vertical scale for growth means defining content standards that describe continuous learning. It also means measuring progress towards those standards with valid and reliable measures that are true to the goals and objectives implied by the standards. (Wise & Alt. 2005).

The mean were computed for each subject in each year and the SS2 as the base class as can be seen in table – based on vertical scale constructed for each content area. The mean increases as the class increases, suggesting that student growth from one class to the next is observed in all the content area. The mean difference between the consecutive class SS 1 and SS 2 were computed to analyze the growth pattern over class. The mean difference as presented in Table – indicate a decelerating growth pattern as the class increases. Therefore, there is an effect of the vertical scale on student performance level across the class levels.

Comparisons of mean scores in measuring one year’s growth State Assessment Test (SAT) in English, Biology and Mathematics as well as across SS1 and SS2 within each content area. Lastly, differential growth within each class for both content areas discussed.

To further specify what constitutes one year’s growth as defined in the State School Accountability System, it is necessary to revisit the developmental scale or vertical scale specifically as it applies to Gain Alternative 4. The definition is based on the numerical cut-scores for the AKSSSPE Achievement Levels that have been approved by the State Board of Education. The following steps were applied to the cut scores, separately, for each subject and each class-level.

The increase in the VSS necessary to maintain in the same relative standing within Achievement Levels from one class to the next was calculated for each of the seven cut scores that separate between the Achievement Levels into five. The median value of these four differences was determined to best represent the entire student population. Mean gain expectations were then calculated for each. This was adopted because it best captures the theoretical expectation of greater gains in the early classes due to student maturation. Graphs of these curves and the resulting expected growth on the development scale is shown in Figure 1 for SS1 and SS2.

To be denoted as making gains under Gain Alternative 4, students must demonstrate more than the expected growth on the developmental scale. Therefore, they must score at least one developmental scale score point more than the values listed above. These criteria were applied to mean differences for the performing students.

VI. Conclusion

The purpose of measurement is to provide information that can be used to improve instruction and learning. Assessment of any kind has value to the extent that, it results in better decisions for students. The most valid assessment of achievement for a particular school is one that most closely defines that school's education standard and goals for teaching and learning. Vertical scaling offers many compelling features to an assessment system. Item response theory is a powerful framework for building tests and understanding their measurement properties, and IRT-based vertical scales promise greater comparability of scores, greater efficiency in test construction and field testing, and better standard setting judgments.

This power of IRT is derived in large part from some relatively strong assumptions, and the validity of interpretations based on the vertical scales depends upon evidence available to support assumptions. For these reasons, it makes sense to examine these assumptions critically in light of the available data.

Most fundamentally, creation of valid vertical scales requires a set of content standards that reflect a degree of continuity over class levels. Assessment programs that do not employ vertical scales eliminate opportunities for interpretation on the cross-class comparisons that vertical scales invite.

Building a vertical scale is not merely a matter of psychometric procedures but instead requires careful design work at all stages of test development, from creating test blueprints to setting performance standards. These additional complexities add some cost and complexity to test development, although some of this cost may be offset by more efficient field testing and item utilization that vertical scaling analyses enable.

It is noted that neither vertical scales nor item response theory are required for measuring growth using tests. Relatively straightforward pre-test/post-test designs using alternate test forms can and have been employed to measure growth in simple and interpretable ways. Such approaches may have limited application under AKSSSPE, however, where one summative test is of interest and where comparability must be maintained as items on the test are changed each year. Nonetheless, the absence of a vertical scale in AKSSSPE assessment systems implies some significant limitations. Most importantly, perhaps, the developmental appropriateness of a progression of class-level performance standards may not be directly observed and assessed.

When their use is appropriate and their construction is sound, vertical scales can significantly enrich the interpretations of test scores and growth trajectories. They provide a systematic way to examine the developmental characteristics and appropriateness of systems of state performance standards across class spans.

VII. Recommendations

In the light of the findings of this study the following recommendations are hereby made:

(i) Teachers should be free to construct and/or administer vertical scale tests on students so that the growth of individual students or group of students can be meaningfully tracked over time.

(ii) Provision should be made for availability and use of vertical scale testing during planning of academic programs.

(iii) In-serving training should be conducted regularly to acquaint teachers on the need and use of vertical scale testing in schools assessment program.

(iv) Vertical scaling plays an important role in the school accountability system because it allows scores to be compared from one year to the next. Therefore, the technical quality of vertical scaling methods used and the documentation of the vertical scaling processes employed should be properly documented. They directory relevant to any accountability system that reflect annual growth and ongoing improvement.

(v) Effort should be made to procure the software that is used for calibration in order to have a well-constructed scale.

(vi) Educators should employ vertical scaling method which defines longitudinal score scale for measuring growth in achievement, and produce reliable sores that satisfy the demand of the user and meet professional test standards.

(vii) The State Ministry of education should determine a general cut-score that will serves as a criteria for measuring growth, proficiency and standard setting for each subject.

References

[1]. Redfield, D. (2001). Critical issue in large-scale assessment: A resource guide. Washington, DC: council of chief state school officers.

[2]. Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C., (1999). Uncommon measures: Equivalence and linkage among educational tests. Washington, DC: National Academy Press.

[3]. Patz, R. J., & Hanson, B. (2002). Psychometric issues in vertical scaling. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

[4]. Wise, L., & Alt, M. (2005). Assessing vertical alignment. Washington, DC: Council of Chief State School Officers.

[5]. North Carolina Standard Course of Study, available at http://www.ncpublicschools. org/curriculum/languagearts

[6]. Smith, R. L, & Yen, W. M. (2006). Models for evaluating class-to-class growth. In R. W. Lissitz (Ed.), Longitudinal and value added modeling of student performance (pp. 82-94). Maple Grove, MN: JAM Press.

[7]. Patz, R. J., & Yao, L. (2006). Vertical scaling: Statistical models for measuring growth and achievement. In S. Sinharay& C. Rao (Eds.), Handbook of statistics, 26: Psychometrics. Amsterdam: North Holland.

[8]. Patz, R. J., & Yao, L. (in press). Methods and models for verticalscaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), Linking and aligning scores and scales. New York: Springer-Verlag.

[9]. Reckase, M. D., & Martineau, J. (2004). The vertical scaling of science achievement tests. Paper commissioned by the Committee on Test Design for K-12 Science Achievement. Washington, DC: National Research Council.

[10]. Kolen, M., & Brennan, R. (2004). Test equating, scaling, and linking: Methods and practices (2nd ed). New York: Springer-Verlag.

[11]. Karkee, T., Lewis, D. M., Hoskens, M., & Yao, L. (2003). Separate versus concurrent calibration methods in vertical scaling. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.