Assessing the Potentials of Compurized Adaptive Testing to Enhance Mathematics and Science Student’s Achievement in Secondary Schools

Mary Patrick Uko, Ph.D
Department of Educational Foundation, School of Education, College of Education, Afaha Nsit, Akwa Ibom State, Nigeria

I.O. Eluwa, PhD
Department of Science Education, College of Education, Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria

Patrick J. Uko, PhD
Department of Science Education, Faculty of Education, Akwa Ibom State University, Ikot Akpaden, Akwa Ibom State, Nigeria

Abstract:
This study focuses on Computerized Adaptive Testing (CAT), designed to provide Mathematics and Basic Science teachers in secondary schools with an innovative method of assessment, that can enhance student’s achievement. A descriptive survey design was adopted for the study. Three research questions and two null hypotheses were formulated to guide the study. Using purposive sampling technique, a sample size of four hundred (400) Junior secondary three (JSS3) students and four private schools with computer laboratories in Akwa Ibom State were selected for the study. Each group had 100 members with mixed Mathematical and Basic Science competencies of high, medium, and low. The instrument for data collection was a Mathematics and Basic Science Computerized Adaptive Test (MABSCAT) developed by the researchers, tested, and modified after trial testing. Descriptive statistics, one-way ANOVA was used to analyse the data. Results indicated that: test lengths, testing times, and Mathematical and Basic Science competencies were significantly different at p=0.05 among four groups of stopping criteria ($\text{SEE} \leq 0.20$, $\text{SEE} \leq 0.30$, $\text{SEE} \leq 0.40$, and $\text{SEE}_{m-1} - \text{SEE}_{m-1} \leq 0.005$); test lengths and testing times were significantly different at p=0.05 among the three groups of Mathematical and Basic Science competencies (low, moderately high, and high); and there were 72.25%, 16.75%, and 8% of the students having a moderately high, low, and high Mathematics and Basic Science achievement respectively. It was recommended among others that, Government should establish item banks, CAT should be introduced in assessment of learning for the actual abilities of the students to be tested with sufficient accuracy.

Keywords: Computer Adaptive Testing, Mathematics, Basic Science, Competency

Introduction
A new system-wide educational change has been introduced in education, requiring amongst other things, that technologies for education be

Suggested Citation
utilized at all levels of education. Assessment is one of the most important components of the educational system. It is an integral part of the learning process and its major purpose is to improve learning. Assessment of students' achievement is a key factor in schooling and is an issue of major concern in education, as test scores are index of academic achievement (Evans, Uko & Ekim 2022; Joshua, 2018). The results of assessment can predict whether an education product is successful or not. The reliability of the outcome using assessment tools is important to students and teachers (Joshua, 2018; Udofia and Uko, 2016). There are many kinds of educational assessment tool including observation, interview, questionnaire, test and so on. The most popular one currently used by teachers is the achievement test. Most teachers try to increase efficiency in measuring achievement, but many, if not most, lack expertise in measurement.

Assessment process in Nigeria is marred by three major problems which are teachers, the testing process and the students. For instance, most teachers lack knowledge of good test items construction and they often encounter difficulties in getting help. They cannot fall back on the test item banks for help, as they are not readily available for the teachers use. Also, they neither know how to calibrate test items nor realize the significance of constructing and calibrating same. With regard to testing problems, appropriate test items and item banks are not available in many subjects for teachers to use. The test items which are used in schools are teacher made and most often, are of poor quality (Iwuoha-Njoku, 2022). Testing arrangements in Nigerian Secondary school are often inappropriate, and test items sometimes do not indicate the actual abilities of the students with sufficient accuracy, as well as the fact that the Students do not often prepare themselves for testing.

In traditional testing, all students in a class have to do the same test at the same time, irrespective of individual differences in students' abilities. Using the same test usually causes some problems, as some items are too easy, or too hard for individuals whose abilities are quite low, or high. Arising from this, students sometimes guess answers test items; spend too much time on some items to their own detriment; they are bored, and are careless when writing test. Such problems cause errors in measurement (Famoroti, 2023; Uko, 2021). The Nigerian educational system is not satisfactorily successful, partly due to the testing process which emphasises the knowledge and memory domain, and the challenge of developing a measuring and evaluating tools, which correspond to the development of skills and abilities of learners. These variables are actually lacking (Uko, 2021). For the educational development to be successful, it is necessary to improve the assessment tools (Eyenaka, Uko & Eyenaka, 2016).

The shift to digital delivery of educational tests has ignited interest in developing new approaches to collecting evidence of students' learning with efforts geared towards Computer Adaptive Testing where questions are presented to candidates based on their ability level (Eluwa and Uko, 2020). The benefits of CAT are ability level items, shorter tests, reduction in test anxiety, test efficiency and higher measurement precision. Considering the importance of student's assessment in learning, teachers need to develop an effective assessment instruments in order to achieve specific learning objectives. Effective assessment begins with identifying the right learning objectives for the understanding of the key scientific ideas and to be able to utilize them through scientific practices. Upon identifying the learning objectives, the next step is to monitor student's progress in specific objectives. The last step is to assess how much of the objectives have been achieved. Based on these steps, a conclusion can be drawn up, that, assessment means, an activity of collecting information on students to assist teachers in decision making (Istiyono, 2018). In this twenty-first century, the mode of assessing students or examinees for ability estimation has shifted toward the use of CBT (Bandele, 2019; Sheu, 2019).

Many researchers who have carried out studies on CBT operations, application and usage have similarly described Computer-based test as a
modern method of administering test to examinees through the use of computers and as a viable alternative to traditional methods of Paper and Pencil Test (Bennett, 2015; Khoshshima, Hosseini & Hashemi, 2017; Okocha, Eyiolorunshe & Owolabi, 2017). The fact that assessment for the estimation of students’ cognitive ability at any level of education cannot be trivialised and left on a spot, necessitated continuous research on how CBT can be better explored for assessment use, more so that almost all examiners are shifting to it. The Computer Based Test (CBT) as a means of assessment has many types that can be used to assess test takers; the major decision on which one to use lies with the examiner who conducts the test (Ejim, 2017). This study narrows down the type to Computer Adaptive Test.

Computerized Adaptive Testing (CAT) is a type of CBT that adapts the ability of the examinees to the difficulty of the items. CAT involves modifying the assessment to take account of the test taker’s ability. In most cases, this involves choosing questions for candidates based on the score they got on previous questions in order to maximize the precision in measuring their level. As examinees do well, the test questions get harder, if they do less well, the questions get easier. At some point the examinee’s level settles out, and a test outcome is decided (Magis & Barrada, 2017). CAT performs iterative and adaptive administration by selecting and administering items one by one to the test taker according to the principle that the selection of the next item is conditional upon the response of the respondent to the previously administered item and the current estimate of ability (Magis & Barrada, 2017). Estimating students’ ability with CAT has its significant benefits to include: the time taken for an assessment to reach a judgment about a candidate’s level of ability can be significantly shortened, perhaps by as much as 50%, candidates tend to be better motivated as their time in the assessment is not spent answering questions that are way too hard or much too easy, the test can potentially measure a very broad range of skills, reducing the test length to half of what it used to be and still maintain a substantial level of precision (Han, 2018). This will benefit both the teachers and the students because the students will not be wasting time attempting the 'too easy items' or 'too difficult items' while the test administrator will benefit from the reduced cost of the examination (Seo & Choi, 2018).

Another advantage ascribed to CAT is its item selection algorithm’s capacity to reduce exposure level of some of the items to examinee because, all the examinees will receive different questions unlike what operates in the fixed linear testing where the same set of questions will be administered to all examinees (Famoroti, 2023). Other advantages of CAT include, reducing human errors mostly on the concept of test reliability, curbing examination malpractices because if two or more examinees are sitting very close to one another, they will not be administered the same question, since CAT administers questions based on the estimated ability of individual examinee in the examination centre (Famoroti, 2023). Mathematics and science are very important for all levels of education in Nigeria. They are offered as compulsory subjects in the basic education year. In the secondary education levels, mathematics is one of the core subject and so it is important for all students. The importance of mathematics and science can be seen in terms of daily living, in civic life, and in working. Moreover, mathematics is a necessary basic skill for studying many other subjects at school and in life. Many students have problems that cause low mathematics and science achievement. These problems arise from the curriculum, teaching, learning, and assessing (Patarapichyatham, Locke & Lewis, 2021; Uko & Uko, 2019). Assessing in mathematics and science is often a problem in secondary schools, for example, most tests are only used once, test items are often of poor quality, the items are not matched to the abilities of the students because the items are analysed by using Classical Test Theory (True Score Theory). One significant problem that causes low Mathematics and science achievement is students often have negative attitudes to Mathematics and Science. Most researchers, teachers and educators try to develop the new innovations to solve the
problems (Patarapichyatham, Locke & Lewis, 2021). The researchers are interested in using technology in mathematics and science testing, by creating computer programs for Mathematics and science Adaptive Testing, and see how it can enhanced students’ achievement in Mathematics and Basic Science in Secondary schools in Akwa Ibom State.

**Theoretical Framework**

Analysis of assessment instrument is necessary in order to obtain a good testing. The analysis of test includes the analysis of the characteristics of measurement instrument used and the analysis of test participants’ capabilities. This test analysis is based on two theories which are, Classical Test Theory (CTT) and Item Responses Theory (IRT). This study was premised on the Modern test theory which is also known as item response theory (IRT). The 3-Parameter Logistic Model (3-PLM) of Item Response Theory (IRT) for dichotomously scored responses was used. The theory establishes the examinees' interaction level with the items in the test, based on the probability of correct response to an item. CAT is premised on Item Response Theory (IRT) which explains examinees’ responses to test items using a mathematical function modeled using either one (difficulty-\(b\)), two (discrimination-\(a\) of an item after item difficulty parameter has been computed) or three (guessing-\(c\) in addition to \(b\) and \(a\)) parameter to explains the level of interaction of the examinees with test items based on the probability of correct response (Oladele, Ayanwale and Owolabi, 2020). The 3-PLM IRT model adopted for this study considered the estimates of difficulty (b-parameter), discrimination (a-parameter), and pseudo guessing (c-parameter). Item Response Theory is increasingly employed in construct development due to its many advantages over classical test theory approaches.

In all, Item Response Theory is very useful in the construct development phase of testing. It now includes a vast array of models that postulate qualitatively different types of underlying constructs. Comparative fit indices for different Item Response Theory models can provide interpretations about the constructs that are measured. For example, inconsistent findings about the number and nature of constructs involved in specific tests result, in part, from applying methods that are inappropriate for item-level data. Applying multidimensional Item Response Theory models to item level data results in more valid findings. Furthermore, it is often suspected that some test items are population-specific; that is, performance may differ qualitatively over different groups of persons. Sometimes the populations are intrinsic to the measure, such as employing different strategies to solve the items. Other times the populations differ in background, such as defined by gender, racial-ethnic background, native language, or clinical status (handicaps, disabilities). Item Response Theory models are available not only to assess these differences, but their application can provide solutions.

One aspect of this research examines the effect of experience with Computer-Based Tests and how it may impact students’ achievement. Therefore, Theory of Constructivism is also considered. Bruner’s Theory of Constructivism includes student readiness and scaffolding. Information must be introduced to students at an appropriate age and developed over time. Therefore, Bruner felt that teachers should use a spiral curriculum, in which students are introduced to content and skills and then revisit content to better develop their understanding (Schunk, 2016). Vygotsky’s Zone of Proximal Development expands on this concept, whereby students can learn new content but may need guidance from adults or peers to accomplish a task. “The experiences one brings to a learning situation can greatly influence the outcome” (Schunk, 2016). These theories indicate that students may need practice with and guidance from teachers and peers before taking Computer Adaptive Tests.

The purpose of this study is to compare the achievement scores of students that took computer adaptive test, to find out if it can enhance their achievement in Mathematics and Basic Science. Therefore, one must consider the constructivist theory, by which people develop their knowledge and understanding through interactions with persons and situations.
Constructivism also proposes that one’s learning is influenced by one’s own environment (Schunk, 2016). When considering the implications of constructivism, it is important for educators to allow students to interact with computer tools and computer-based assessments in order to develop a deep understanding of the expectations and format.

Self-regulated learning is a vital element of student development. This means, being involved in one’s learning and performance on a multi-dimensional level, including behaviorally, cognitively, metacognitively, and motivationally (Schunk, 2016; Uko, 2021). Self-regulated learning (SRL) is multi-faceted and includes self-monitoring and self-reinforcement. The model of Self-Regulated Learning is explained by the “Cyclical Phases Model” developed by Zimmerman and Moylan in 2009, this illustrates the thinking that is needed to complete adaptive tests and to grow over the course of the academic year (Panadero, 2017). According to Zimmerman and Moylan (2009), the model depicts that, self-regulation includes not only strategy and time management, but also self-consequences and metacognitive monitoring. After the performance, students should exhibit self-judgement and self-assessment, which should lead to forethought for future performances. This can include goal setting and planning for future assessments.

In many curriculum-based tests, the forethought process may be less valuable because tests on the same topic or chapter, in schools are not going to take place, as the teachers most often move on to a new chapter and do not test old materials. However, with computer-adaptive tests, this forethought and goal-setting can be very important to the students’ achievement, as the content may change but the strategies used by the students might improve over time and contribute to their growth and achievement. Azevedo (2011), opined that, “Students who are effective at self-regulating their learning will continue to capitalize on the opportunities of computer-based learning environments (CBLE), while those who lack this ability will find themselves at a serious disadvantage. Educators would do well to consider preliminary and formative assessments of their students’ SRL skills, knowledge, and motivation while using CBLEs and then design scaffolding interventions accordingly”. Without self-regulated learning skills, students’ achievement on assessments may be hindered, especially when using a new format of testing, such as CAT. Without the experience of computer-based assessments, student achievement may be hindered. With the shift in assessments, it is important to see how the new trend and experience with a program impacts growth, achievement scores, attitudes, and motivation of those taking the tests.

**Statement of the Problem**

The primary goal of educational reforms is to provide students with 21st century skills required to respond to global challenges. It therefore implies that many new knowledge are invented. Students and educators should be made to be conversant with deep knowledge of the 21st century skills. This is to allow the achievement of learning objectives and correct utilization of science and technology. It is an established fact in the literature that CBT, whether linear or adaptive, is a modern means of estimating the ability of examinees in this twenty-first century, especially the latter one. It is also a fact in the literature that CAT, a type of CBT, estimates the ability of the examinees more precisely than LCBT and PPT. Many researchers have compared the ability estimates under the traditional method of testing known as the paper and pencil test and the ability estimate under the linear computer-based test; some even investigated the perceptions of the examinees about the linear computer-based test. However, despite the fact that CBT, especially the aspect of computerized adaptive testing, is more precise in ability estimation and also receiving more attention in the educational field, it can still be said that in Nigeria as a whole and in Akwa Ibom State in particular, much research work has not been done to exclusively assess the potential of CAT in enhancing junior secondary school students achievement in Mathematics and Basic Science. Therefore, this study is interested in assessing the potentials of computerized adaptive testing in enhancing Mathematics and
Basic Science students’ achievement in secondary schools in Akwa Ibom State.

**Purpose of the Study**
The aims of this study were to:

1. Investigate the changes that exist if any, between test length and testing times in relation to different stopping criteria in Computerized Adaptive Testing;
2. Compare the test length, testing time, and mathematics competency for different stopping criteria;
3. Compare the test length and testing time among differences in mathematics achievements of the examinees.

**Research Questions**

1. What changes exist between test length and testing times in relation to different stopping criteria in Computerized Adaptive Testing?
2. What changes exist between test length and testing times in relation to differences in Mathematics and Basic Science competency of the examinees?
3. What are the changes in measured Mathematics and Basic Science competencies using Computerized Adaptive Testing when different stopping criteria are applied?

**Hypotheses**

**HO1.** There is no significant mean differences in test length and testing times, among stopping criteria and Mathematics and Basic Science competencies of students.

**HO2.** There is no significant mean difference between Mathematics and Basic science competencies and different groups of stopping criteria.

**Methodology**

**Design of the Study**
A descriptive survey design was adopted for the study. This design is considered appropriate because a group of people or items are being studied by collecting and analyzing data from only a few people or items considered to be representative of the entire population (Nworgu, 2015). Therefore, the design served better in finding out the potentials of CAT in enhancing students’ achievement in Mathematics and Basic Science in private secondary schools as a representative of the entire population of students in Akwa Ibom State.

**The Area of the Study**
The area of the study was Akwa Ibom State in the South-South Zone, Nigeria which has thirty one local government areas. Akwa Ibom States has three Senatorial Districts which are; Akwa Ibom North East, Akwa Ibom North West and Akwa Ibom South. There are 243 public secondary schools in the state. The choice of junior secondary schools in Akwa Ibom State for this study is that the researchers have observed that Akwa Ibom State have not introduced CBT into the system yet, talkless of CAT and this may be as a result of some constraint in developing items and creating item banks. Again the JSS level is the foundation classes and if a good foundation is laid there then it will help them in the senior levels. Therefore this study will in no small way be very beneficial to the State.

**Population and Sample**
The population for this study consisted of all Junior secondary School Students, but focus was on (JSS3) in private secondary schools in Akwa Ibom State, Nigeria who have completed the Junior Secondary Mathematics and Basic Science’ syllabus and are equiped with functional computer laboratory. The choice of private schools was preferred because of the clause of completion of syllabus and well equiped computer laboratory which may not likely be achievable in public schools. This study employed a purposive sampling technique to identify and select schools who have completed Junior Secondary School Mathematics and Basic Science’ syllabus as at the time of the study. The purposive sampling technique was also used to select four schools with an enabling environment where computer facilities can be used among the schools selected for the study.
Two schools from the Akwa Ibom North West, one each from Akwa Ibom North East and Akwa Ibom South were selected on the basis of computer facilities. Intact classes of J. S. S. 3 students in the selected schools were used for the study. A sample size of four hundred (400) Junior secondary three (JSS3) students were selected for the study. The gender-ability mix of students from each school were randomly assigned to four subgroups. Each group contained 100 members whose different mathematical competencies were mixed: 30 students in high ability group, 40 in medium, and 30 in low competency level. Four stopping criteria techniques (SEE ≤ 0.20, SEE ≤ 0.30, SEE ≤ 0.40 and SEE_{m} - SEE_{m-1} ≤ 0.005); were used for simple random selection of the students for each group. Stopping criteria (SC) refers to the values specified for stopping the testing of each examinee. The standard error of estimation (SEE) of an examinee’s ability was used.

Instrument for Data Collection

Instrument for data collection for the study was Mathematics and Basic Science Computerized Adaptive Test (MABSCAT). The structure of the test was categorised according to the objectives and the time used. The researchers developed the multiple-choice Mathematics tests on equations, Basic Science test on living and non-living things based on the objectives, contents, and time required in learning and teaching mathematics (equations), Basic Science test (living and non-living things) from the curriculum and the scheme of work prescribed for the JSS3 students. The initial Mathematics and Basic Science test items for the item bank consisted of 300 items developed according to the sub-objectives. The 300 items were divided into six papers with 50 items each, and each paper contained 40 different items and 10 common items. All the items in the six tests were arranged from easy to hard. The detailed test items included item number, stem and choices, difficulty, standard error (SE), and key answer (1 for choice a, 2 for choice b, 3 for choice c, and 4 for choice d) for each item. After the pilot test, the responses of the examinees were analysed with 3-Parameter Logistic Models of the multidimensional item response theory (MIRT) package in R-software environment with a view to establishing and ascertaining all the necessary psychometrics properties of the instrument. The 180 items that did not satisfy the difficulty, discriminating and guessing parameters criteria in the initial pool of items were removed. The 120 items with difficulty indices between -3 and +3; and discriminating indices between 0.5 and above were finally selected for the CAT. The reliability of the items was established by using Kuder Richardson 20 Formula (KR 20). The items fitting the model were stored in the item bank to be used for the computer adaptive testing.

Method of Data Collection

The process of Computerized Adaptive Testing administration started from getting the results of a sample of the second terminal examination scores on Mathematics and Basic Science from the four sampled schools in order to classify the students into three different groups of Mathematical and Basic Science competency, high, medium, and low. A total of 785 students, based on the scores gained from the exams, were obtained. The researchers then selected 400 students and divided them into four subgroups having the same features of Mathematical and Basic Science competency levels by using a stratified random sampling technique. Each group contained 100 members whose different Mathematical and Basic Science competencies were mixed: 30 students in high ability group, 40 in medium, and 30 in low level. Four stopping criteria techniques were used for simple random selection of the students for each group. Each student’s information such as name and surname, student code, class, and school was installed into the database of the Computerized Adaptive Testing program with passwords by the researchers. Then the students were given their passwords before they enter the examination room. The researchers set the computers up for Computerized Adaptive Testing by installing the Computerized Adaptive Testing program into the computers on the basis of four stopping criteria. After this process, each of the students were allowed to enter the room and sit in front of the computers in accordance with the specific computer for him or her. After
that, the researchers explained Computerized Adaptive Testing and how to do the test, and demonstrated answering the test through the projector. They were allowed to ask questions about the testing. The students were then told to begin the test by following the instructions mentioned on Computerized Adaptive Testing from the beginning until the last process. After the completion of the test, the result of each of the test takers was automatically recorded into the computer for further statistical analysis.

**Method of Data Analysis**

The SPSS computer program was used to analyse data from the 400 JSS3 students. The frequencies and percentages as well as One Way ANOVA were used as the indicators to examine Mathematics and Basic Science competencies of the students. This was done by examining differences in test length and testing times among the different groups relating to stopping criteria and also to examine any differences among the different groups for stopping criteria for Mathematics Basic Science competencies. Where the F statistics were significantly different, the Sheffe Multiple Range test was used to determine which groups the true differences lie.

**Result**

The result of the analysis of Mathematics and Basic science competencies of JSS3 students is set out in Table 7. It presents frequencies and percentages of Mathematics and Basic science competencies of the students in the three groups (low, moderately high, and high).

**RQ1.** What changes exist between test length and testing times in relation to different stopping criteria in Computerized Adaptive Testing?

<table>
<thead>
<tr>
<th>Achievement</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>67</td>
<td>16.75</td>
<td>16.75</td>
</tr>
<tr>
<td>Moderately high</td>
<td>285</td>
<td>71.25</td>
<td>88.00</td>
</tr>
<tr>
<td>High</td>
<td>48</td>
<td>12.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As can be seen from Table 1, the results showed that there are 67 (16.75%) of students that were regarded as having a low mathematics and Basic science achievement (mathematics and Basic science measures were from -1.02 to 0.00 logits). From 0.00 to 1.00 logits, there were 285 (71.25%) students that were regarded as having a moderately high Mathematics and Basic science achievement. From +1.00 to +3.00 logits, there were 48 (12.00%) students that were regarded as having a high Mathematics and Basic science achievement.

**RQ2.** What changes exist between test length and testing times in relation to differences in Mathematics and Basic Science competency of the examinees?

From Table 1 above, there were 71.25%, 16.75%, and 12.00% of the JSS3 students having a moderately high, low, and high Mathematics and Basic Science achievement respectively.

The four groups of stopping criteria were $SEE \leq 0.20$ (1), $SEE \leq 0.30$ (2), $SEE \leq 0.40$ (3), and $SEE - SEE_{m-1} \leq 0.005$ (4). Test lengths were significantly different among four groups of stopping criteria. The mean highest test length of 8.14 items and the mean lowest test length of 3.14 items were in group 3 stopping criteria and group 2 stopping criteria. Each group was significantly different from the others. Testing times were significantly different among the four groups of stopping criteria. The mean highest testing time of 5.74 minutes and the mean lowest testing time of 2.38 minute were in group 3 stopping criteria and group 2 stopping criteria. Each group was also significantly different at from the others.
RQ3. What are the changes in measured Mathematics and Basic Science competencies using Computerized Adaptive Testing when different stopping criteria are applied?

The results of the analysis of the test of different Mathematics and Basic Science competencies among four groups for stopping criteria of Mathematics and Basic Science Computerized Adaptive Testing are shown in Tables 3 and 4. Table 3 shows the F values to examine the difference in Mathematics and Basic Science competencies among the different groups for stopping criteria, while the Sheffe Multiple Range test results for the differences is shown in Table 4.

### Table 3 ANOVA of Mathematics and Basic Science Competencies for the Different Groups by Stopping Criteria

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3.41</td>
<td>3.00</td>
<td>1.14</td>
<td>5.09</td>
<td>.00*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>88.46</td>
<td>396</td>
<td>.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>91.87</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** p means significance based on the F value. * p < 0.05.

As can be seen from Table 3, the F test shows that the difference in the means of the students for the four groups by stopping criteria, $\text{SEE} \leq 0.20 \ (1)$, $\text{SEE} \leq 0.30 \ (2)$, $\text{SEE} \leq 0.40 \ (3)$, and $\text{SEE}_{m} - \text{SEE}_{m-1} \leq 0.005 \ (4)$ were significantly different at the 0.05 significance level in regards to mathematics and Basic Science competency ($F = 5.09, df= 3, 396, p = 0.00$). That is, there were significant differences in the mean mathematics and Basic Science competency levels of students in the four groups by stopping criteria. To determine between which groups mathematics and Basic Science competencies are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 4.

### Table 4. Sheffe Multiple Range Test of Differences in Mathematics and Basic Science Competency by Stopping Criteria

<table>
<thead>
<tr>
<th>Stopping Criteria</th>
<th>(2) Mean</th>
<th>(3) Mean</th>
<th>(1) Mean</th>
<th>(4) Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SEE} \leq 0.30 \ (2)$</td>
<td>0.38</td>
<td>0.16</td>
<td>0.19*</td>
<td>0.25*</td>
</tr>
<tr>
<td>$\text{SEE} \leq 0.40 \ (3)$</td>
<td>0.55</td>
<td>0.03</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>$\text{SEE} \leq 0.20 \ (1)$</td>
<td>0.58</td>
<td></td>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>$\text{SEE}<em>{m} - \text{SEE}</em>{m-1} \leq 0.005 \ (4)$</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** *The mean difference was significant at the 0.05 level.

As can be seen from Table 4, the results showed that mean Mathematics and Basic Science competency for the four groups by stopping criteria was 0.58 logits for the first criteria, 0.38 logits for the second, 0.55 logits for the third, and 0.66 logits for the fourth. There were two main points of difference in Mathematics and Basic Science competency by stopping criteria. The second group was significantly different from group 1 and group 4 at p=0.05. Mathematics and Basic Science competencies were significantly different among the four groups of stopping criteria. The mean highest Mathematics and Basic Science competency of 0.66 logits and the mean lowest Mathematics and Basic Science competency of 0.38 logits were in group 4 stopping criteria and group 2 stopping criteria. Students mathematics and Basic Science
competency in group 2 stopping criteria was significantly different from group 1 stopping criteria and group 4 stopping criteria as can be seen above.

**Hypotheses**

The results of the analysis of the test relating to different test lengths and testing times among four groups for stopping criteria [see description of stopping criteria on page 9] and three groups of mathematics and Basic science competencies of the students with the Mathematics and Basic science Computerized Adaptive Testing are set out in Tables 5 to 12. Tables 5, 7, 9, and 11 show the F values to examine the difference in test length and testing times among the different groups for stopping criteria and mathematics competencies, while the Sheffe Multiple Range test results for the differences are set out in Tables 6, 8, 10, and 12.

**HO1.** There is no significant mean difference between Mathematics and Basic science competencies and different groups of stopping criteria.

### Table 5. ANOVA Analysis of Test Length for the Different Groups by Stopping Criteria

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>1560.31</td>
<td>3.00</td>
<td>520.10</td>
<td>191.30</td>
<td><strong>.00</strong>*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>1076.63</td>
<td>296.00</td>
<td>2.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2636.94</td>
<td>299.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** p means significance based on the F value. *p < 0.05.

The result of the F test in Table 5, shows that the difference in the means of the students for the four groups by stopping criteria, $\text{SEE} \leq 0.20$ (1), $\text{SEE} \leq 0.30$ (2), $\text{SEE} \leq 0.40$ (3), and $\text{SEE}_m - \text{SEE}_{m-1} \leq 0.005$ (4), were significantly different at the 5 per cent significance level in regards to test length ($F = 191.30$, $df = 3, 396$, $p = 0.00$). That is, there were significant differences in the mean test length levels of students in the four groups by stopping criteria. To determine which groups test lengths are significantly different, the Sheffe Multiple Range test was performed. The results are shown below.

### Table 6. Sheffe Multiple Range Test of Differences in Test Length by Stopping Criteria

<table>
<thead>
<tr>
<th>Stopping Criteria</th>
<th>(2)</th>
<th>(1)</th>
<th>(4)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SEE} \leq 0.30$ (2)</td>
<td>3.24</td>
<td>4.44</td>
<td>6.82</td>
<td>8.34</td>
</tr>
<tr>
<td>$\text{SEE} \leq 0.20$ (1)</td>
<td>4.44</td>
<td>1.20*</td>
<td>3.69*</td>
<td>5.00*</td>
</tr>
<tr>
<td>$\text{SEE}<em>m - \text{SEE}</em>{m-1} \leq 0.005$ (4)</td>
<td>6.82</td>
<td>2.49*</td>
<td>3.20*</td>
<td>1.31*</td>
</tr>
<tr>
<td>$\text{SEE} \leq 0.40$ (3)</td>
<td>8.34</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** *The mean difference was significant at the 0.05 level.

The result in Table 6 showed that, mean test length (number of items) for the four groups by stopping criteria was 4.44 for the first criteria, 3.24 for the second, 8.34 for the third, and 6.82 for the fourth. There were six main points of difference in test length by stopping criteria. The third group was significantly different from groups 2, 1, and 4 at $p=0.05$; the fourth group was significantly different from groups 2 and 1 at $p=0.05$; and the first group was significantly different from group 2 at $p=0.05$.

### Table 7. ANOVA Analysis of Testing Times for the Different Groups by Stopping Criteria
The F test result in Table 7, shows that the difference in the means of the students for the four groups by stopping criteria, $SEE \leq 0.20$ (1), $SEE \leq 0.30$ (2), $SEE \leq 0.40$ (3), and $SEE_m - SEE_{m-1} \leq 0.005$ (4) were significantly different at the 0.05 significance level, in regards to testing time ($F = 53.85$, $df = 3$, 396, $p = 0.00$). This means that, there were significant differences in the mean testing time levels of students in the four groups by stopping criteria. To determine between which groups, testing times are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 5.

### Table 8. Sheffe Multiple Range test of Differences in Testing Time by Stopping Criteria

<table>
<thead>
<tr>
<th>Stopping Criteria</th>
<th>Mean (2)</th>
<th>Mean (1)</th>
<th>Mean (4)</th>
<th>Mean (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SEE \leq 0.30$</td>
<td>2.45</td>
<td>0.95*</td>
<td>2.88*</td>
<td>3.36*</td>
</tr>
<tr>
<td>$SEE \leq 0.20$</td>
<td>3.58</td>
<td>1.93*</td>
<td>1.93*</td>
<td>2.41*</td>
</tr>
<tr>
<td>$SEE_m - SEE_{m-1} \leq 0.005$</td>
<td>5.46</td>
<td></td>
<td></td>
<td>0.48</td>
</tr>
<tr>
<td>$SEE \leq 0.40$</td>
<td>5.87</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The mean difference was significant at the 0.05 level.*

As can be seen from Table 8, the results showed that mean testing times for the four groups by stopping criteria was 3.58 minutes for the first criteria, 2.45 minutes for the second, 5.87 minutes for the third, and 5.46 minutes for the fourth. There were five main points of difference in testing times by stopping criteria. The third group and the fourth group were significantly different from groups 2 and 1 at $p=0.05$; and the first group was significantly different from group 2 at $p=0.05$. The third group with the stopping criteria of $SEE \leq 0.40$ was not significantly different from group 4.

**HO2.** There is no significant mean differences in test length and testing times, among stopping criteria and mathematics and Basic Science competencies of students.

### Table 9. ANOVA Analysis of Test Length for the Different Groups of Mathematics and Basic Science Competency

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>207.04</td>
<td>2</td>
<td>103.52</td>
<td>16.91</td>
<td>.00*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>2,429.89</td>
<td>397</td>
<td>6.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2,636.94</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The mean difference was significant at the 0.05 level.*

As can be seen from Table 9, the F test shows that the difference in the means of the students in the different groups of mathematics and Basic science competency, high, moderately high, and low, were significantly different at $p=0.05$ (at the 5 per cent significance level) in regards to test length (number of items) ($F = 16.91$, $df = 2$, 397, $p = 0.00$). This means that, there were significant
differences in the mean test length levels of students in the three groups of the mathematics and Basic science competency. To determine which groups, test lengths are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 5.6.

Table 10. Sheffe Multiple Range Test Differences in Test length by Mathematics and Basic Science Competencies

<table>
<thead>
<tr>
<th>Mathematics and Basic science competencies</th>
<th>(2)</th>
<th>(3)</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.22</td>
<td>6.33</td>
<td>7.09</td>
</tr>
<tr>
<td>Moderately high (2)</td>
<td>5.22</td>
<td>1.10</td>
<td>1.86*</td>
</tr>
<tr>
<td>High (3)</td>
<td>6.33</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Low (1)</td>
<td>7.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *The mean difference was significant at the 0.05 level.

As can be seen from the results Table 10, the mean test length (number of items) for the three groups of mathematics and Basic Science competency was 7.09 for the first (low), 5.22 for the second (moderately high), and 6.33 for the third (high). The test length of the students in the first group with the low mathematics competency was significantly different from that in group2 (moderately high) at the 5 per cent significance level.

Table 11. ANOVA of Testing Times for the Different Groups by Mathematics and Basic Science Competency

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>52.40</td>
<td>2</td>
<td>26.20</td>
<td>4.07</td>
<td>.02*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>2,556.27</td>
<td>397</td>
<td>6.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2,608.67</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: p means significance based on the F value. * p < 0.05.

The F test result in Table 11 shows that, the difference in the means of the students in the different groups of high, moderately high, and low mathematics and Basic Science competency, were significantly different at 0.05 significance level, in regards to testing times (F = 4.07, df = 2, 397, p = 0.02). This hows that, there was significant differences in the mean testing time levels of students in the three groups of the mathematics and Basic Science competency. To determine which groups, testing times are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 12.

Table 12. Sheffe Multiple Range test of Differences in Testing Time by Mathematics and Basic Science Competencies

<table>
<thead>
<tr>
<th>Mathematics and Basic science competencies</th>
<th>(2)</th>
<th>(3)</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.97</td>
<td>4.60</td>
<td>4.96</td>
</tr>
<tr>
<td>Moderately high (2)</td>
<td>3.97</td>
<td>0.53</td>
<td>0.94*</td>
</tr>
<tr>
<td>High (3)</td>
<td>4.60</td>
<td></td>
<td>0.41</td>
</tr>
<tr>
<td>Low (1)</td>
<td>4.96</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *The mean difference was significant at the 0.05 level.
The results in Table 12 above showed that, mean testing time for the three groups of mathematics competency was 4.96 minutes for the first (low), 3.97 minutes for the second (moderately high), and 4.60 minutes for the third (high). There was one main point of difference in testing time by mathematics and Basic Science competencies. The first group with the low mathematics and Basic Science competency was significantly different from group 2 (moderately high) at the 5 per cent significance level.

Discussion of Findings

Computerized Adaptive Testing software can be used to examine differences in Mathematics and Basic Science competency of students. The findings indicated that the Mathematics and Basic Science competencies were significantly different among four groups of stopping criteria. The mean highest Mathematics and Basic Science competency and the mean lowest Mathematics and Basic Science competency were in group 4 and group 2 respectively. Also, there were changes in test length and testing times related to differences in stopping criteria in Computerized Adaptive Testing. Test lengths and testing times were significantly different among four groups of stopping criteria. The mean highest test length and testing times and the mean lowest test length and testing times were in group 3 and group 2 respectively. Each group was significantly different in both test length and testing times from the others.

Again there were changes in test length and testing times related to difference in Mathematics and Basic Science competency of the participants. Test lengths and testing times were significantly different among the three groups of Mathematics and Basic Science competencies. The mean highest test length and testing times and the mean lowest test length and testing times were in group 1 (low Mathematics and Basic Science competency) and group 2 (moderately high Mathematics and Basic Science competency). There was only one significantly different test length and also testing times between students in group 1 (low Mathematics and Basic Science competency) and group 2 (moderately high Mathematics and Basic Science competency). For the Mathematics and Basic Science ability, there were 72.25%, 16.75%, and 8% of the JSS3 students having a moderately high, low, and high Mathematics and Basic Science achievement respectively. Measured Mathematics and Basic Science competencies were significantly different among four groups of stopping criteria. The mean highest and the mean lowest Mathematics and Basic Science competencies were in group 4 stopping criteria and group 2 stopping criteria.

This suggests that Computerized Adaptive Testing might help students improve their learning achievement in Mathematics and Basic Science. This conclusion is backed up with the findings of Matilda and Helen (2019); Humes, 2021; Alashwal (2020), which revealed that computerized adaptive testing is likely to be accurate in assessing individual student’s ability in any testing situation. The teachers can use it with individual students or groups without worrying about cheating in the examinations. Computerized Adaptive Testing could help prevent examinees from getting bored with having too many test items. Also, through Computerized Adaptive Testing, each examinee does different test items and different number of items. This depends upon an individual's ability. In addition, data gained from the test can be used for many purposes, such as, to follow up an individual's learning progress, to diagnose deficiencies in each student, and to assess students' achievement. Student's weaknesses in any subject matter can consequently be remedied. Computerized Adaptive Testing is an efficient and authentic assessment of student's learning. In relation to school network, it would be useful for schools in all the senatorial district to have access and to network with the school that has the item banks through Computerized Adaptive Testing. The school network could either develop a bank containing tests of
different subject areas or different banks for different subject areas. This can be done by establishing one school as the item bank, equipped with a central computer, while other member schools in the network can access the bank through the networking computers in their schools. This can save time and school resources in preparing tests and conducting examinations whenever they need. Regarding the development of the test items, teachers in every school network could cooperate to construct, try out, analyse, and select qualified items to store in the item bank. If this process is continuously done, the item bank will become large with thousands of well-calibrated items by difficulty equated on the same scale. Moreover, Computerized Adaptive Testing is a new approach for learning assessment and evaluation which is likely to be the future of assessment. There is a large monetary cost to implement this, but it would be well worth it, and probably necessary in the future.

Conclusion

Computerized adaptive testing is likely to be accurate in assessing individual student’s ability in any testing situation. The teachers can use it with individual students or groups without worrying about cheating in the examinations. Computerized Adaptive Testing could help prevent examinees from getting bored with having too many test items. Also, through Computerized Adaptive Testing, each examinee does different test items and different number of items. This depends upon an individual’s ability. In addition, data gained from the test can be used for many purposes, such as, to follow up an individual’s learning progress, to diagnose deficiencies in each student, and to assess students' achievement. Student's weaknesses in any subject matter can consequently be remedied. Computerized Adaptive Testing is an efficient and authentic assessment of student's learning. In relation to school network, it would be useful for schools in all the senatorial district to have access and to network with the school that has the item banks through Computerized Adaptive Testing. The school network could either develop a bank containing tests of different subject areas or different banks for different subject areas. This can be done by establishing one school as the item bank, equipped with a central computer, while other member schools in the network can access the bank through the networking computers in their schools. This can save time and school resources in preparing tests and conducting examinations whenever they need. Regarding the development of the test items, teachers in every school network could cooperate to construct, try out, analyse, and select qualified items to store in the item bank. If this process is continuously done, the item bank will become large with thousands of well-calibrated items by difficulty equated on the same scale. Moreover, Computerized Adaptive Testing is a new approach for learning assessment and evaluation which is likely to be the future of assessment. There is a large monetary cost to implement this, but it would be well worth it, and probably necessary in the future.

Recommendation

1. It is recommended that teachers should develop Computerized Adaptive Test with well-calibrated items by difficulty equated on the same scale and item banks for use in secondary schools in Akwa Ibom in different subjects. This can save time and school resources in preparing tests and conducting examinations whenever they need.

2. In relation to school networking, it is recommended that schools in all the senatorial district should have access and to network with the school that has the Computerized Adaptive Test item banks. The school network could either develop a bank containing tests of different subject areas or different banks for different subject areas. This can be done by establishing one school as the item bank, equipped with a central computer, while other member schools in the network can access the bank through the networking computers in their schools.
3. The Ministry of Education, through the state secondary Education Board should launch the pooling of resources between different schools. Commissioner of Education should run in-service courses on CAT and item banks development, with items appropriate to many school subjects. Moreover, Computerized Adaptive Testing is a new approach for learning assessment and evaluation which is likely to be the future of assessment.

4. To attract examinees' attention, concentration and willingness, a test in a form of multimedia, i.e., with moving graphics, pictures and sound should be constructed. This will make the test items more challenging.

Further Research

For the future research, there are several recommendations:

1. A study that develops CAT in the form of multimedia such as, moving graphics, pictures and sound that will attract examinees' attention, concentration and willingness should be conducted. This will make the test items more challenging.

2. Presently, the Internet is becoming more significant in all subject areas for different purposes like, in presentations of lectures, in teaching and learning, in conducting research, in transferring information, and in measurement and evaluation. In the area of measurement and evaluation, further research on conducting CAT on the internet is suggested.

3. Generally, as observed in the literature, most of the CATs are developed on cognitive traits. Very few studies have been developed on affective traits, like, diligence, discipline, honesty, attitudes and Such traits were measured using rating scale questionnaire. There are only a few qualified instruments for measuring the affective trait, therefore, more studies on CAT for the affective traits employing the Partial Credit Model of Rasch in analyzing the questionnaire items are worth conducting.

4. This study was conducted in Mathematics and Basic Science, a replication of this study should be done on other subjects.

References


