From Computer Adaptive Testing to Automated Scoring in a Senior Secondary School Physics Essay Test in Osun State, Nigeria

Dr. J. Gbenga Adewale[1] and Olubunmi A. Etuk-Iren[2]

[1]International Centre for Educational Evaluation
Institute of Education
University of Ibadan
Ibadan

[2]School of Science
Federal College of Education (Technical)
Akoka, Yaba, Lagos

Abstract

Many teachers see teaching as enjoyable but scoring as a task they would like to avoid. This implies that many teachers would prefer automated scoring to manual. Computer-based testing supports the use of multiple-choice, drag-and-drop and fill-in-the-blank tests in objective type of testing as well as essay test items. With a well-designed protocol, this form of assessment provides instant feedback to students which allows them monitor their academic progress. It should be noted that scoring objective tests such as supply (completion and fill-in) and select (yes/no, matching and multiple choice item) is not challenging. Many examination bodies can score with little or no support from the manufacturers of such scoring software. By contrast, essay tests, which can be categorized into restricted and extended response tests, do not have the characteristic of a "one cap fits all" approach**.** Each question has to have its own software. Software designed for a particular examination cannot be used for another examination even in the same subject. A battery of physics tests was prepared using an appropriate text editor and uploaded to the internet. Students were expected to answer the essay questions and submit their responses online. Software (based on moodle platform) was prepared to score each student's responses by extracting predetermined features from their responses. The number of the predetermined features extracted from the students' responses corresponded to their scores depending on the predetermined marks for each of the features. In order to validate the automated scoring, a manual marking was done using a prepared marking guide. The two sets of scores were correlated and a positive and high correlation was established between the automated and manual scorings. Therefore, we advocate automated scoring.

Keywords: Computer Adaptive Testing, Automated Scoring, Moodle, Learning Management System, Physics Essay Test

Introduction

Teaching, which is the transfer of information from someone with adequate knowledge to the learner, is done with ease. In fact, everybody can teach if the materials to be taught are available, although it is another issue for the teaching to be done professionally. The focus of this paper is not on teaching but on scoring using information and communication technology. Before we score, it is important to survey the test scripts to be scored.

Students' response in the test scripts is a function on the type of tests given. This implies that there are different types of tests; however, achievement test is our focus in this paper. Achievement tests in a particular subject, according to Obemeata (2000); Ayodele, Adegbile and Adewale (2001); and Amoo and Adewale. (2007) are a series of questions given (using a criterion-referenced test- the curriculum) to learners in order to determine their mastery level in the subject.

Achievement tests are usually designed as a terminal evaluation to determine the status of an individual on completion of a course of study or training. These tests are characterized in terms of their content validity (Adewale, 2006). There is the need for value judgement and rational decision on what the tests should be measuring. This judgement and decision are informed by the objectives of the course of study or training. These objectives which are stated in terms of desired behaviour provide clues for the achievement test construction. An achievement test should be measuring fully the status of the individual in all the hierarchical levels of understanding as proposed in the Bloom's taxonomy of educational objectives. The test should measure:

i. Remembering previously learned materials (knowledge)
ii. Grasping the meaning of learned materials (comprehension)
iii. Making use of learned materials in new and concrete situation (application)
iv. Breaking learned materials into its component parts so as to understand its organizational structure (analysis)
v. Putting parts together to come out with new whole form (synthesis) and
vi. Judging and assessing the value of learned materials for a given purpose (evaluation).

Test experts have classified achievement tests using different parameters. Whereas some classify tests on the basis of the behaviour that is being measured, others classify considering the types of items contained in the test, the purpose of tests, etc Okpala, Onocha and Oyedeji (1993). However, achievement tests may be classified on the basis of the objective-type and essay-type.

Usually, there are two types of objective test: supply and select. In supply type, candidates are expected to fill-in the gap like short answers and completion. However, in select type, candidates are expected to pick an option (s)he thinks is correct out of many options. Examples are the true/false items, matching items and multiple choice items. In this type of test, the candidate is given some statements to which (s)he should respond. The statements have to be marked as either "true" or "false", "yes" or "no", "agree" or "disagree". For example, Liberia is in Africa (True or False). This type of question is greatly influenced by guessing. In the fill-in type (short answer or completion), candidates' knowledge on dates, names, items, figures or facts is tasked e.g. the name of the author of the story "The man Died is _____. In the matching type, the candidates are presented with two columns, each consisting of a list of names, facts, places, etc which they are expected to match based on the instruction given. Lastly, the multiple choice type is used to measure simple learning outcomes and learning tasks involving the entire six educational

objectives in cognitive domain. In this type of test, the candidates are required to choose one correct response to a problem. Objective test are difficult to construct but easy to score.

The second type of test is the essay type**.** The essay test has been a very popular type of achievement test. Its usage started earlier than 2300 B.C. in ancient China. The candidate is expected to generate ideas, organize ideas, express idea and integrate ideas as (s)he solves a problem. It is described as an attempt at answering questions in the form of continuous but connected writing in which the candidate is free to express himself in his own way. The two main forms of essay test are the extended response and the restricted response. In extended type, candidates are expected to organize their answers and express them in their own words in a continuous and connected form. A good example of this *"Discuss in details how examination malpractice, sexual harassment and cultism affect quality of education in Nigeria".* In the restricted or short-answer essay, the candidate is given a number of topics or questions and (s)he is asked to write briefly on them. It limits both the content and the type of candidate's response. For example, mention any four factors affecting learning or list three layers of soil. Here, learners' opportunity to demonstrate depth of knowledge is extremely limited.

We have discussed the two main types of tests. Either objective or subjective, they can exist in either paper pencil based testing (PBT) or computer based testing (CBT). In paper pencil testing, candidates use their pencils on their Optical Marks Recognition (OMR) sheet. Whereas on the computer based testing which refers to a situation in which some aspects of computer technology is deployed in the assessment process, for example, interactive tests completed on a computer. Some of the merits of computer based testing according to Desforges, (1989) are listed:

i. It reduces the risk involved in travelling from candidates' location to the examination centre when computer based testing is in operation because computer based testing allows the candidate to take the examination close to home.
ii. Since there is less travel, less amount of time and money is expended on taking an examination.
iii. Examination questions and visual images are easily viewable from the computer screen, and are generally of higher quality when presented on a computer. The advantage is that images are often easier to read on the computer screen than on the printed copy.
iv. The test items could be generated from the pool of test or item banking 15 minutes before the start of an examination, hence eliminate problems associated in test security.
v. Impersonation is also removed because each candidate bio data (biometrics) is completed during registration for the examination and before the candidates are allowed into the examination, verification is done e.g. thump printing could be used as the candidate password before the question portal is opened.

vi. Candidates' selection of responses to test items is done with easy. The candidate is expected to click an icon that represents the option s/he thinks is the key to a particular test item instead of spending seconds shading the answer in the paper pencil based testing.

vii. Unlike paper and pencil test which publishes candidates' results at a later time, computer based testing marks immediately and publishes the results indicating candidates' characteristics generated during registration and the examination scores.

viii. Computer based testing is far more consistent in its scoring, and it is rare for errors to occur. When errors do occur in Computer Assisted Testing, trouble spots are obvious, easier to identify, and easier to correct.

Disadvantages of computer based testing are:

i. Students without or limited knowledge of computer application may not be able to attempt questions from a computer based test.

ii. One cannot go back and change any of answers once entered and submitted.

iii. There is no way of "jumping ahead" on the computer test unlike the paper – pencil test. One must take the questions in the sequence the software presents them

iv. This testing is limited to schools with a large number of desktops or laptops. Alternatively, if a school does not have large number of computers, students in such a school should be able to go to school with their computers. In a situation where this is not possible, the use of computer in the computer based testing will be difficult

Marking the objective test is one of the easiest things as it was pointed out earlier that one of the advantages of the objective test is that it is very easy to score most especially if adequate instructions are given on the question paper. For each item, only one possible answer is provided. If it is truly a very good objective test, a scoring key is available indicating the answer for each item. The most plausible way of obtaining total score for each student is to simply count the number of correct answers. In any form of objective test, there is always a main stem and options from where to select the correct one. Also, there is usually only one correct answer. As the name indicates, the marking of objective tests lays itself to objectivity. Inter-rater scores are constant and the correct options are predetermined, thus marking objective tests present little or no problem as regards the objectivity and reliability of the marking. Objective tests can be marked by another person who is not in the field of the subject or discipline. The two basic methods of marking objective tests according to Yoloye (2005) are:

(1) Hand marking or self-marking and

(2) Machine marking.

In hand marking or self-marking, the examiner marks the answer of the examinees manually, either on the question booklet or on the separate answer sheet provided. There is marking and counting of item by items the testee's correct responses. In a situation where separate answer sheets are used, a stencil i.e. a regular answer sheet containing

holes punched at the position of the correct option to the items can be used. The examiner places the stencil on the examinees answer sheet and counts the corresponding correct responses, (i.e. the number of shaded options appearing in the holes).

The use of stencil makes marking very fast, but disastrous errors could be made if the stencil is not well placed on the examinee's answer sheets. In order to ensure that objective tests are objectively marked (i.e error free) by using stencil the following procedures can be taken:

(i) Punch the position where the candidate's identification number and the first and last item number should appear.

(ii) Examine the candidate's answer sheet to ensure that not more than one response has been shaded for each item.

(iii) Fix the stencil to the candidate answer sheet edge to edge and use clip to hold the stencil and the answer sheet in place so that the papers will not shift.

(iv) In some cases, instead of just counting the shaded points that appear in the stencil, identification marks in form of ticks could be used before counting the correct responses. This method provides a feedback to the candidate on the correct options.

The second method is the machine marking, as a result of increase in the number of subjects as well as the number of examinee's, machine scoring or marking has been introduced. This is to reduce the burden of the examiners on hand marking. Thus, the exercise of marking objective tests becomes faster and errors are also reduced as observed in some public examinations like the junior and senior secondary school certificate examinations and the University Matriculation Certificate Examination. Scoring machines are available in various configurations and levels of sophistication in terms of capacity and complexity.

Scoring of objective tests is not a challenge as this can be done with some instruction previously coded into the computer, on the contrary, scoring of essay questions using computer is somewhat challenging. Marking essay questions in students' script is one of the challenges faced by teachers. Automated scoring according to Bennett (2011) refers to a large collection of grading approaches that differ dramatically depending upon the constructed-response task being posed and the expected answer. The assessment task for which automated scoring can be used is patterned after a scoring "model." The model extracts features from the student response and uses those features to generate a score. Although, there are many software in the market that assert to the possibilities of marking essay questions, some of these are the blackboard platform, ActionScript platform and moodle platform (the authors are familiar with these), however, it is important to note that, regardless of all the claims of these software's vendors, a machine does not read, understand, or grade a student's essay in the same way a human rater would. Most of the software are designed to look out for some prescribed features in the students responses in which a machine recognizer must first process the response and generate hypotheses for each word. If automated scoring is adopted, there is the possibility of the teacher

concentrating his/her teaching on such features that promote short answers. Assessments demanding for short responses are associated with questions in the realm of factual information, ability to recall facts, names, dates, places, etc. For some of these short answer responses, automated scores could agree with human scores provided there is a high inter-rater reliability. Using automated scoring for short answer questions, student's score are likely to be enriched.

On the contrary, students are not likely to express themselves. Questions that demand for higher order cognitions – like application, analysis, synthesis and evaluation in the Bloom's Taxonomy of educational objectives will not be tested. It may be difficult for computer to mark such questions that bother on describing how a student spent his last holiday. We can get as many responses as the number of students who respond to such question. Apart from such question that demand from students skills in description or narrative essay, testing objectives with such active words like demonstrate, discover, modify, operate, produce, relate, draw conclusions, determine evidence, break down, discriminate, illustrate, point out, predict, construct, create, design, generate, appraise, compare, conclude, contrast, criticize, justify, interpret, etc. may also be difficult.

Whether we use the manual or machine marking, there are some procedures we need to follow; some of them are:

i. Criteria for scoring essay items - the following three criteria are applied to scoring content, organization and process criteria.

ii. Standardizing the making scheme - the chief examiners are expected to prepare a marking scheme. This document which is usually general shows the proportion of marks to be allocated to various aspects of the answer. After the examination, the chief examiner marks a sample of the scripts. In the light of candidates' interpretations of questions, the chief examiner in consultation with the team leaders may amend the marking scheme. This is followed by a standardization meeting where the marking scheme is discussed in great detail and a high degree of agreement is reached. The examiners the mark scripts allocated to them. In order to ensure that the agreed marking scheme is applied as uniformly as possible, each examiner is expected to send several samples of his marked scripts to his Team Leader, who will compare his marking with the marking scheme.

iii. Marking procedures - marking of essays may be analytic or impression. In analytic marking, decision is made on what we are trying to assess in the essay and allocation of marks to the various aspects of the essay is carefully considered. For instance, are we interested in grammar, spelling and handwriting? In impression marking, the examiner reads a question rapidly, forms a general impression, and then records a rating.

iv. Reliability of essay marks - the reliability of essay marks which is a measure of the consistency of the examiner could be influenced by several factors. For example, the consistency of the marking, the variability of the examiners, the

suitability of the topic-whether the essay is measuring the same ability each time it is administered. Reliability of essay marks could be improved by the use of team marking, if only one scorer, score every essay question twice. If the essay question consists of several items, mark the answers to the first item on all papers, then go on to mark the second item, and score each paper without looking to see the name of the candidate. This is called scoring blindly.

We have discussed how impossible it is to score essay questions using computers, notwithstanding, an attempt was made is scoring four essay questions which involved definitions and stating of some concepts

Hypotheses

i. There is no significant deference between scores obtained from automated marking and the ones obtained from manual marking.

ii. There is no significant relationship between scores obtained from automated marking and the ones obtained from manual marking.

RESEARCH METHODOLOGY

This study adopted a survey research in descriptive setting. The population for the study consisted of all Senior Secondary School II students offering physics in the State of Osun. The study adopted a multi-stage sampling technique. All the 30 Local Government Areas in the State of Osun were stratified into three groups; rural, semi-urban and urban. Two strata were randomly selected and one Local Government Area was randomly selected from each stratum. One school was then selected from each cluster making 6 schools in total.

The main instrument used for this study was the Physics Achievement Test (PAT). The instrument was constructed by the researchers to measure acquisition of knowledge in four selected topics in SS2 physics. Ambitiously, we started with 20 essay questions. Test re-test was used to determine the stability of the instrument using 73 samples. The samples used were equivalent but not the same samples to be used for the study. The samples were made to provide answers to the questions and after two weeks interval, they responded to the same questions again. The test-retest reliability coefficient was gotten to be 0.74. The content validity was established using the table of specification drawn from the selected topics in the physics syllabus. The items were developed to reflect the four cognitive domains of Blooms taxonomy of educational objectives; which are Knowledge, Comprehension, Application and Analysis. The table of specification for the test items is shown in Table 1

Table 1: Table of Specification showing distribution of items

Topic/Instructional Objectives	**Knowledge** (50%)	**Comprehension** (25%)	**Analysis** (25%)	**Total**
Production and propagation of waves (45%)	1,2,4,8,17	6	3,5,9	9
Types of waves (20%)	10	7,12	11	4

Properties of waves (5%)		13		1
Light waves (30%)	14,15,16,20	19	18	6
Total	10	5	5	20

In addition, the content validity of the instrument was also determined using Lawshe Content Validity Index of 0.6 when administered to 20 Physics secondary school teachers.

Physics Achievement Test Marking Guide

The physics achievement test marking scheme was constructed by the researchers. It was constructed in line with what was included in the teachers' instructional guide to ensure that it conforms with what students were taught. It was given to two practicing teachers in the field of physics to ensure that it is appropriate for the level of students involved. The reliability of the instrument was ensured by making sure that only one scorer was used. This was done in absence of inter-scorer reliability.

Data collection procedure

The sample cut across all the geographical locations (rural, semi-urban and urban) in the State of Osun. This implies that the use of internet was limited only to urban cities and some semi-urban cities. It was difficult to have students using the link to get to the questions, so the questions were given to them. They used paper and pencil for uniformity sake. Their responses were typed out and the researchers fed their responses in the space provided for answers in the moodle learning environment system. The scores were obtained and recorded. Since they had responded to the questions in paper and pencil mode, their responses were marked. Their scores were also obtained and recorded.

Data Analysis

The statistical tool used in this study to determine the relationship between the scores obtained in the automated scoring and manual scoring were subject to Pearson Product Moment Correlation (r).

Moodle Platform Implementation

Moodle (acronym for Modular Object-Oriented Dynamic Learning Environment) is a free software e-learning platform, platform, also known as a Learning Management System (LMS), or Virtual Learning Environment (VLE). As of June 2013 it had a user base of 83,008 registered and verified sites, serving 70,696,570 users in 7.5+ million courses with 1.2+ million teachers. Moodle has several features considered typical of an e-learning platform. Some typical features of Moodle that relevant to this paper are:

- Assignment submission
- Online quiz
- Grading

These are the steps used in hosting the questions in the moodle platform:

1. We selected the question category
2. We gave our question a descriptive name.
3. We created the question text. If you're using the HTML Editor, you can format the question just like a word processing document.
4. We set the 'default question grade' (i.e. the maximum number of marks for this question).
5. We added general feedback. This is text that appears to the student after he/she has answered the question.
6. We decided not to use case-sensitive because case sensitivity can be tricky where capitalization is important. For example, can one accept *Ban Ki-moon* as well as *ban ki-moon* as an answer?
7. We filled in the answers we accepted. We gave common misspellings that could not be resolved a partial credit with this option.
8. We added grade for each answer.
9. We created feedback for any and all answers. This appearred if the student enters that answer.
10. We clicked Save Changes to add the question to the category.

Results and Discussion

The two hypotheses were tested at 0.05 significant level.

Hypothesis One

There is no significant deference between scores obtained from automated marking and the ones obtained from manual marking.

Table 2 presents the mean scores of students whose scripts were automated marked and manually marked. It also shows the degree of differences between the two sources of variations.

Table 2: Mean Scores of Automated and Manual Scorings and students t-value

Source of variation	N	Means score	S.D.	t-value	Sig. 7
Automated scoring	100	7.8450	2.65208	1.875	0.0890
Manual scoring	100	8.4450	2.69691		

The total score for the test is 11 points. Students' whose scripts were marked manually scored more than those scored with machine. The reasons could be connected to the fact that in manual marking, arrangement of how answers are presented is not an important issue. For example, the question that demanded for properties of waves is used as an illustration. There are four properties of waves namely:

A. Reflection
B. Refraction
C. Diffraction

D. Interference **2marks**

Correct mentioning of a property of waves earns half a mark. In manual marking, it is possible to mark these properties irrespective of the ways they are arranged. For example, interference could come first or refraction coming last. In as much as these properties are listed, full marks will be awarded. On the other hand, the arrangement is important in auto scoring. In order to overcome this problem, we provided the four properties four times and instructed the computer to select any of the four properties in number one, this was repeated for the remaining three. So we have had 16 options. Moreover, some spelling error could be overlooked in manual making, but this is not so in automated marking. We made effort to provide as many variances of some words after going through the students' scripts, for example, reflection could be spelt refelction, eflection, flection, reelection, relection, refletion, etc. In order words, we go a 10! (10 factorial i.e. 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1= 3,628,800 possible spellings of reflection. We did the same for the rest properties of wave. We took care of arrangements and spellings but we did not take care of punctuation marks like commas and full stops during demonstrations. These were incorporated later. Another source of variation between the automated and manual scorings could be due to other features we were not aware of. The student t-test shows a value of 1.875 which is not significant at $P<0.05$. The difference obtained between the automated scoring and manual scoring does not warrant the rejection of the hypothesis. The implication is that the scores obtained the students is not influenced by the mode of scoring.

Hypothesis 2

There is no significant relationship between scores obtained from automated marking and the ones obtained from manual marking.

Table 3 shows the relationship (correlation) between the scores of students in the two modes of scorings.

Table 3 Relationship between Automated and Manual Scorings

Source of variation	N	r	Sig. r
Automated scoring	100	0.853	0.000
Manual scoring	100		

Table 3 shows a positive and relatively high correlation between the automated scoring and manual scoring. The hypothesis is therefore rejected. This implies that the two modes of scoring could be used interchangeably depending on the ease of use. As at now, the automated scoring is somewhat difficult in Nigeria for the following reasons:

i. Electricity challenges
ii. Internet presence in rural schools
iii. Knowledge of the use of software by the examiners.
iv. The rigour involve in getting all the variances of arrangements of ideas or concept and spelling, and use of punctuation marks.

These are non-issue with the manual marking. On the other hand, if automated scoring is adopted, it is possible to reduce the cost of examination, as no paper would be needed for printing. Cost of transporting the examination materials also will be eliminated. Finally, the examiner also could sit at a centre close to his/her home without having to travel a long distance (cost and risks of travelling will be reduced or eliminated).

Conclusion and Recommendations

The scores obtained by the students in physics are not significantly influenced by the mode of scoring. Again, there is a relatively high correlation between the automated scoring and manual scoring. This implies that either of the automated scoring or manual could be used to score students' scripts. Since an of the two modes of scoring could be used in scoring students' scripts, it is recommended that automated scoring should be encourage in schools even though, the rigour involved in the automated scoring could serve as a source of discouragement to the examiners, the advantage outweighs the disadvantage. This method of scoring could be used if four difficulties identified in the discussion are taken care of.

References

Adewale, G. (2006): Item Analysis of Life Skills Achievement Test for Nigerian Non – Formal Education Learners: Implications for MDGs. Adult Education in Nigeria Vol. 13, 77-92.

Amoo S. A. and Adewale J. G. (2007). Junior Secondary School Students' Competency in ICT: An Assessment of Effectiveness of SchoolNet Nigeria Project in Southwest, Nigeria. Journal of Computer Literacy Vol. 7. No 1, 32–56.

Ayodele, S. O., Adegbile, J. A. and Adewale, J. G. (2003 and 2009): Evaluation Studies. Ibadan. Powerhouse Press and Publishers; ISBN 978 – 35794 – 6 – 0.

Bennett, R. E. (2011). Automated Scoring *of* Constructed-Response Literacy and Mathematics Items. https://www.google.com.ng/search?noj=1&q=bennetts+and+automated+scoring&oq=bennetts+and+automated+scoring&gs_l=serp.12...230.6445.0.8692.14.11.0.0.0.0.0.0..0.0....0...1c.1.26.serp..14.0.0.0cYrqDKp6_A

Desforges, Charles (1989). *Testing and Assessment*. London: Cassell Education Limited.

Obemeata, J. O. (2000). Principles of Test Construction. Sterling Horden, Ibadan.

Okpala, P.N., Onocha, C.O. and Oyedeji 1993. *Measurement and evaluation in education.* Ibadan: Stirling-Horden Publishers (Nig.) Ltd.

Yoloye, T. W. (2005) Marking and scoring of Items In Yoloye, T. W. and J. G. Adewale (eds) ***Training Manual*** on Measurement and Evaluation for the International Association for Educational Assessment (IAEA)Workshop University of Ibadan, Nigeria.