ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

Development of a Software Package for Computing Psychometric Properties of Multiple Choice Tests Using Norm-Referenced Approach.

Kalu Eke Osonwa, Ngozi N. Agu, Virgy E. Ejiofor, Uduma E. Osonwa

Faculty of Education, Nnamdi Azikiwe University

DOI: https://doi.org/10.51244/IJRSI.2024.1103015

Received: 02 March 2024; Revised: 14 March 2024; Accepted: 19 March 2024; Published: 05 April 2024

ABSTRACT

This work is on the development of a software package for computing the psychometric properties of multiple choice tests using norm-referenced approach. Focus was on item indices which are difficulty, discrimination and distracter indices. There are no easily affordable statistical packages in the Nigerian market for the computation of these indices. Most studies involving them have always been done manually due non-availability of affordable statistical software. Hence, the researcher was motivated to embark on this investigation. Three research questions were formulated to guide this study. Analysis of the existing system was done on BILOG-MG statistics. This study deployed the following methodologies: mathematical, structured systems analysis and design methodology (SSADM), the object-oriented analysis and design methodology (OOADM) and prototyping. The design for the study was instrumentation. The population for the study was 225 students which consisted of all the JS3 students of 2016/2017 session in Port Harcourt, Rivers State. The researcher made use of two instruments for the study. The first was the Rivers State Basic Education Certificate Examination (BECE), 2017 multiple choice questions in Mathematics. The second was a group of system software: HTML, PHP, JavaScript and MySQL. The responses of the students were used to test the effectiveness and efficiency of the software package developed in this study. Data analysis was done first with manual computation and second with the software package that was developed. The manual computation took the researcher 180 hours while the electronic computation with the developed software took only six hours to complete. The findings from the results of the study showed that the developed software was accurate in computing the difficulty, discrimination and distactor indices of the test items and it has a very high rate of accuracy. The study recommended that the developed software should be used to compute the difficulty, distractor and discrimination indices of multiple choice tests.

Keywords: software package, psychometric properties, norm-referenced tests, difficulty, discrimination and distracter indices, BILOG-MG Statistics, SSADM, OOADM, Prototyping, HTML, PHP, JavaScript and MySQL.

1. Dr Kalu Eke Osonwa, Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.
2. Prof Ngozi N. Agu, Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.
3. Prof Virgy Ejiofor, Department of Computer Science, Faculty of Sciences, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.
4. Prof Uduma Eke Osonwa

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

INTRODUCTION

Measurement is an integral and inalienable aspect of any formal educational environment or research studies. It determines the extent to which the learning objectives have been achieved. Psychometricians and psychologists are constantly working on ways to improve the accuracy of the instruments used in measuring the various domains of man which simply consist of the cognitive, affective and psychomotor. These measuring tools used in ascertaining the presence, absence as well as the degree of the existence of traits in any of these domains of man are simply referred to as tests.

To Orluwene (2014, p.4), "a test can be regarded as an instrument used to determine the relative presence or absence of the trait measured for." A test is therefore a measuring instrument, device or tool developed, adopted or adapted by investigators for data collection concerning a trait in the cognitive, affective or psychomotor domain. Tests could be classified on several bases and one of them is the method based on types of students' responses. This method consists of the objective tests and the essay tests.

The objective test is a type of test which yields the same score when different independent examiners mark the same script with the same mark guide. This type of test could take different forms which include multiple-choice test items, short-answer, alternate response, arrangement type and matching type. Kpolovie (2014) explained that a multiple choice test is composed of a stem, a key, and distractors. The key and the distractors are collectively known as the options while the stem states the problem or question. The correct response is known as the key and the incorrect responses are called the distractors. The essay tests, on the other hand, are tests in which the students are required to express their views or ideas in writing in an attempt to provide the answers to questions. They are two types of essay tests namely restricted-response type and unrestricted-response type essays. In restricted-response essay type, students are limited in the number of lines, words or pages they are to write while in the unrestricted-response type; students are given the freedom to write all they know concerning the concept or construct of interest. The interest of this study is not on the essay tests but on the multiple-choice test which is a form of the objective test.

A test undergoes several stages in its development process. It is not developed in a haphazard manner whether it is a teacher-made or a standardized. It requires great skills by trained personnel especially when the test has to be standardized. A standardized test is a measurement tool in education that has systematically undergone the various stages of an instrument development and is also accompanied with a manual that gives instruction on its purpose, properties, administration and scoring procedure. A teacher-made test does not pass through all these rigourous steps while being developed. It is developed by a teacher who may not be an expert in measurement and evaluation. However, efforts are made by the developer to ensure that such tests are considered valid and reliable for the purposes they were meant. By implication, both the teacher-made tests and the standardised tests are expected to be accurate, efficient and effective as measurement instruments.

The accuracy, efficiency or effectiveness of any test instrument is simply explained by its psychometric properties. The psychometric properties refer to the quantifiable attributes that relate to the statistical strength or weakness of a test or measurement instrument (Alvior, 2013 & The Free Dictionary, 2015). The psychometric properties of a test are qualities that are contained in a test instrument which explain its usefulness as a tool for data collection. In essence, the psychometric properties of test instruments are a function of their reliability, validity and item indices.

The reliability of a test instrument is the consistency with which it measures what it was designed to measure. Kpolovie (2014) defined reliability as the degree or magnitude of consistency with which a test produces consistent scores. Nworgu (2015, p.197) similarly stated that, "the reliability of an instrument refers to the degree of consistency with which the instrument measures whatever it measures." There are

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

various measures of reliability estimates that a test could be subjected to. These include estimates of temporal stability, internal consistency, equivalence and scorer/rater variability. Each of these methods is either used to establish the homogeneity of the items that make up the test or to ascertain the stability of the entire instrument.

Another vital property of test instruments is validity. This refers to the extent to which an instrument measures up its purpose (Nworgu, 2015). It is the degree or extent to which a test measures what it purports to measure. In other words, the validity of an instrument refers to the ability of an instrument to cover its content area as well as the underlying construct it was designed for. The various types of validities as detailed by Shuttle worth (2009) are: external validity, internal validity, test validity, criterion validity, content validity, construct validity and face validity. While reliability and validity are qualities of the entire test, there may be need to look into the qualities of the items that make up the entire test. These qualities are revealed by its item indices.

The item indices give detailed properties of each of the items that make up the entire test instrument. The item indices include the difficulty index, the discrimination index and the distractor index (Fidelis, 2014; Kpolovie, 2014). The item difficulty index is a measure of the proportion of examinees who answered the item correctly; for this reason it is frequently called the p-value which is the proportion of examinees who got the item right (Professional Testing, 2015). The Professional Testing further noted that the p-value might be more appropriately called the item easiness index, rather than the item difficulty. Item difficulty ranges from 0.0 to 1.0. An index of 0.0 indicates that the item is very difficult such that none of the examinees got it correct. An index of 0.5 shows average difficulty; while 1.0 implies that the item is very easy such that all the examinees got it correct.

The discrimination index is a measure of the ability of a test item to differentiate between students with high ability level and those with low ability level. It indicates the extent to which an item can discriminate between the dull students and the bright ones. Orluwene (2014) stated two methods of establishing the discrimination index of an item; these are the extreme group method and the point bi-serial correlation method.

In the extreme group method, the examinees are divided into three groups using specific percentages to get the upper norming and the lower norming groups. The middle group is discarded. Alternatively, a cut-off mark could be used to divide the testees into the upper criterion (high ability) group and the lower criterion (low ability) group. The number of examinees in the lower group getting the item correct is subtracted from the number of examinees in the upper group getting the item correct. After this, the difference is divided by the average of the number of examinees in the upper group and the lower group put together. The second method of establishing the discrimination index is the Point Bi-serial Correlation method. Kpolovie (2010) and Agu (2014) also confirmed that the Point Bi-serial Correlation is a statistic for establishing the discrimination index of a test item. Statistics Solutions (2015) explained that Point-Biserial Correlation analysis, like all other correlation statistics, measures the strength of association or co-occurrence between two variables. It discriminates between examinees that fall into the two groups – high ability and low ability. When the discrimination index of an item has been established, a third aspect of the item indices that needs to be established is the distractor index.

The distractor index of an option attempts to find out whether a particular option appeals more to the low ability students than the high ability ones. In the same vein, Carlenton State University (2016) explained that all of the incorrect options, or distractors, should actually be distracting also that each distractor should be selected by a greater proportion of the lower scorers than of the top group. These statements reveal that the distractor index is more analytical than the difficulty and the discrimination indices because it deals with each of the options that make up an item except the key which is the correct answer.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

The validity, the reliability and the item indices jointly determine the quality of an instrument. They do not operate independently for it is possible for an instrument to be reliable but not valid. For instance, a test can consistently measure a wrong construct but produces almost the same reliability coefficient. This implies that the instrument is reliable but invalid. Similarly, a test may meet both the validity and the reliability properties, but defaults in the item parameters. The calculation of these psychometric properties of tests is very cumbersome to execute manually. This necessitates the use of a software package for their computations.

A computer software package is defined by Jones, Roach, Setter and Esling (2008) as a computer program that is sold together with instructions on how to use it. It is an invisible aspect of a computer which contains programs and applications. A package can contain several related programs. Mitchell (2014) explained that computer software mainly refers to the set instructions which have been grouped into programs that make the computer to function in the desired ways the user intends it to. There are many software packages in circulation in the world today. Some of them have been commercialised and made available to public users while a good number are specialised, customised or tailored to satisfy the need of a particular organisation or corporate body. The available commercialised software packages which are designed for statistical computations include but not limited to: the Statistical Package for Social Sciences (SPSS), Statistical Analysis System (SAS), Minitab, Liserel, Ststistica, Bilog-MG and Microsoft Excel.

It is unfortunate that most of these commercialised statistical packages listed, except the Bilog-MG, do not make provision for the computation of item indices despite their importance in the instrumentation of research studies and the construction of valid tests for effective teaching and learning process. The Bilog-MG can only be used to conduct item analysis on Item Response Theory and not based on Classical Test Theory (CTT).

The Classical Test Theory is a theory of test development which originated from physical measurement and the study of errors in measurement and is primarily aimed at assigning higher scores to students or testees who posses more knowledge/skills/trait than those who posses less of it. It is also regarded as the true score theory. Classical test theory holds the assumption that each observed score of a testee is a combination of the testee's true score and some error components which invalidate the essence of the test. Item analysis using this test theory cannot be done with the Bilog-MG.

Another major disadvantage of Bilog-MG is that it is very expensive which makes it difficult for an average user to acquire. Third, the Bilog-MG does not accept raw data in the form of responses such as a, b, c, d or e which are used in the multiple choice type of tests. These responses are vital in item analysis. Fourth, it does not provide step-by-step results on how the item indices were computed. Fifth, it is not a web-based application, hence it does not make provision for multiple users in the case of cloud computing. Sixth, it is complex and requires specialised training before an average user can run an analysis with it. Sometimes, the user may be required to have knowledge of coding which is a specialised skill. Considering these shortfalls, there is need to develope a software package that will close this gap. It is against this background that the researcher conceived the idea to embark on a study to develop a software package for computing the psychometric properties of test instruments with specific reference to difficulty, discrimination and distractor indices.

Research Questions

The following research questions guided the conduct of the study.

1. How accurate is the developed software in computing the difficulty indices of test items?
2. How accurate is the developed software in computing the discrimination indices of test items?

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

3. How accurate is the developed software in computing the distractor indices of test items?

METHOD

The study adopted an instrumentation design because this software package is an electronic instrument. The study was conducted using 225 JSS3 secondary school students. The researcher made use of two instruments for the study. The first was in hard copy while the second was a set of system software. The first was the Rivers State Basic Education Certificate Examination (BECE, 2017) multiple choice questions in Mathematics. This examination which consisted of 60 items was used because it had 5 response options from A to E. Responses from tests with multiple options A-D could also be analysed using the software that was developed. The second instrument that was used for the study was a group of system software which consisted of the following technologies: HTML (Hypertext Mark-up Language), PHP, JavaScript and MySQL.

The validity and the reliability of the first instrument (which was Rivers State Basic Education Certificate Examination (BECE) question paper in Mathematics) were not necessary in the study because the intention of the researcher was not to establish its psychometric properties. The second set of research instrument which consisted of HTML, PHP, JavaScript and MySQL has proved valid and reliable in writing programs or applications in this information age (21st Century). Applications written with them cut across various fields of human endeavour: economics, politics, technology, medicine, education, agriculture, etc. To say the least, this instrument can be used to develop software in any problem area.

The researcher administered the Mathematics BCCE for 2017 to the students. Their scripts were collected and scored for the manual item analysis. Secondly, the students' responses for each question fed into the software that was developed for the computerized calculation.

This study deployed the following methodologies: Mathematical/Statistical; the Structured Systems Analysis and Design Methodology (SSADM); the object-oriented method; and Prototyping.

RESULTS

Research Question 1

How accurate is the developed software in computing the difficulty indices of test items?

Table 1: Results of the Difficulty Indices of Items Computed Using the Developed Software (IASP) and those Obtained from Manual Calculations

	SOFT.	MAN.		SOFT.	MAN.		SOFT.	MAN.		SOFT.	MAN.
QST.	**Diff**	**Diff**	**QST.**	**Diff**	**Diff**	**QST.**	**Diff**	**Diff**	**QST.**	**Diff**	**Diff**
1	0.9	0.9	**16**	1	1	**31**	0.9	0.9	**46**	0.4	0.4
2	0.4	0.4	**17**	0.5	0.5	**32**	0.5	0.5	**47**	0.9	0.9
3	0.4	0.4	**18**	0.3	0.3	**33**	0.4	0.4	**48**	0.5	0.5
4	0.3	0.3	**19**	0.5	0.5	**34**	0.5	0.5	**49**	1	1
5	0.4	0.4	**20**	0.4	0.4	**35**	0.5	0.5	**50**	0.5	0.5
6	0.4	0.4	**21**	0.6	0.6	**36**	0.5	0.5	**51**	0.4	0.4
7	1	1	**22**	0.4	0.4	**37**	0.9	0.9	**52**	0.5	0.5
8	0.3	0.3	**23**	0.7	0.7	**38**	0.6	0.6	**53**	0.5	0.5

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

9	0.4	0.4	**24**	0.4	0.4	**39**	0.9	0.9	**54**	0.4	0.4
10	0.3	0.3	**25**	0.4	0.4	**40**	0.5	0.5	**55**	0.5	0.5
11	1	1	**26**	0.4	0.4	**41**	0.4	0.4	**56**	0.5	0.5
12	0.5	0.5	**27**	0.4	0.4	**42**	0.4	0.4	**57**	0.5	0.5
13	0.4	0.4	**28**	0.3	0.3	**43**	0.9	0.9	**58**	0.3	0.3
14	0.3	0.3	**29**	0.4	0.4	**44**	0.4	0.4	**59**	0.5	0.5
15	0.9	0.9	**30**	0.3	0.3	**45**	0.4	0.4	**60**	0.6	0.6

*SOFT. = indices obtained from the developed software

*MAN. = indices obtained by manual calculation

*QST. = question (item) numbers

*Diff. = difficulty indices

Research Question 2

How accurate is the developed software in computing the discrimination indices of test items?

Table 2: Results of the Discrimination Indices of Items Computed Using the Developed Software (IASP) and those Obtained from Manual Calculations

	SOFT.	MAN.		SOFT.	MAN.		SOFT.	MAN.		SOFT.	MAN.
QST.	Disc.	Disc.	QST.	Disc.	Disc.	QST.	Disc.	Disc.	QST.	Disc.	Disc.
1	0.2	0.2	**16**	0	0	**31**	0	0	**46**	0.3	0.3
2	0.2	0.2	**17**	0.5	0.5	**32**	0.6	0.6	**47**	0	0
3	0.4	0.4	**18**	0.5	0.5	**33**	0.4	0.4	**48**	0.5	0.5
4	0.5	0.5	**19**	0.6	0.6	**34**	0.5	0.5	**49**	0	0
5	0.4	0.4	**20**	0.4	0.4	**35**	0.4	0.4	**50**	0.4	0.4
6	0	0	**21**	0.6	0.6	**36**	0.6	0.6	**51**	0.2	0.2
7	0	0	**22**	0.5	0.5	**37**	0.1	0.1	**52**	0.5	0.5
8	0.4	0.4	**23**	0.4	0.4	**38**	0.7	0.7	**53**	0.8	0.8
9	0.3	0.3	**24**	0.3	0.3	**39**	0	0	**54**	0.5	0.5
10	0.4	0.4	**25**	0.4	0.4	**40**	0.6	0.6	**55**	0.3	0.3
11	0	0	**26**	0.5	0.5	**41**	0.2	0.2	**56**	0.6	0.6
12	0.7	0.7	**27**	0.7	0.7	**42**	0.6	0.6	**57**	0.4	0.4
13	0.4	0.4	**28**	0.4	0.4	**43**	0.1	0.1	**58**	0.5	0.5
14	0.2	0.2	**29**	0.3	0.3	**44**	0.4	0.4	**59**	0.4	0.4
15	0.1	0.1	**30**	0.5	0.5	**45**	0.3	0.3	**60**	0.2	0.2

*SOFT. = indices obtained from the developed software

*MAN. = indices obtained by manual calculation

*QST. = question (item) numbers

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

*Disc. = discrimination indices

Research Question 3

How accurate is the developed software in computing the distractor indices of test items?

Table 3: The Results of the Distractor Indices of Items Computed Using the Developed Software (IASP) and those Obtained from Manual Calculations

QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.
1	A	0.1	0.1	6	A	0	0	11	A	0	0	16	A	0	0
	B	0.1	0.1		B	0	0		B	0	0		B	0	0
	C	0	0		C	0	0		C	0	0		C	0	0
	D	0	0		D	0	0		D	0	0		D	0	0
	E	-0.2	-0.2		E	0	0		E	0	0		E	0	0
2	A	-0.2	-0.2	7	A	0	0	12	A	0.1	0.1	17	A	0.1	0.1
	B	0	0		B	0	0		B	0.1	0.1		B	-0.5	-0.5
	C	0	0		C	0	0		C	-0.7	-0.7		C	0.2	0.2
	D	0.1	0.1		D	0	0		D	0.2	0.2		D	0	0
	E	0	0		E	0	0		E	0.2	0.2		E	0.1	0.1
3	A	0.1	0.1	8	A	0.1	0.1	13	A	0	0	18	A	0.1	0.1
	B	0.1	0.1		B	0.1	0.1		B	0.1	0.1		B	-0.5	-0.5
	C	-0.4	-0.4		C	-0.4	-0.4		C	-0.4	-0.4		C	0.3	0.3
	D	0.1	0.1		D	0.1	0.1		D	0.2	0.2		D	0	0
	E	0.2	0.2		E	0.1	0.1		E	0	0		E	0	0
4	A	0.1	0.1	9	A	-0.3	-0.3	14	A	0	0	19	A	0.1	0.1
	B	-0.5	-0.5		B	0.1	0.1		B	0.1	0.1		B	-0.6	-0.6
	C	0.3	0.3		C	0.1	0.1		C	-0.2	-0.2		C	0.2	0.2
	D	0	0		D	0	0		D	0.2	0.2		D	0.1	0.1
	E	0	0		E	0.1	0.1		E	0	0		E	0.1	0.1
5	A	0.1	0.1	10	A	0	0	15	A	0	0	20	A	0.1	0.1
	B	0.1	0.1		B	0.1	0.1		B	-0.1	-0.1		B	-0.4	-0.4
	C	0.2	0.2		C	-0.4	-0.4		C	0	0		C	0.2	0.2
	D	-0.4	-0.4		D	0.1	0.1		D	0	0		D	0.1	0.1
	E	-0.1	-0.1		E	0.2	0.2		E	0	0		E	0	0

QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.
21	A	0.2	0.2	26	A	0.3	0.3	31	A	0	0	36	A	0.1	0.1
	B	0.2	0.2		B	-0.5	-0.5		B	0	0		B	0.2	0.2
	C	0.2	0.2		C	0.1	0.1		C	0	0		C	0.2	0.2

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

	D	-0.6	-0.6		D	0.1	0.1		D	0	0		D	-0.6	-0.6
	E	0.1	0.1		E	0.1	0.1		E	0	0		E	0	0
	A	0.2	0.2		A	0.2	0.2		A	0.1	0.1		A	0	0
	B	0.1	0.1		B	-0.7	-0.7		B	0.1	0.1		B	0	0
	C	-0.5	-0.5	27	C	0.2	0.2	32	C	-0.6	-0.6	37	C	-0.1	-0.1
	D	0	0		D	0	0		D	0.1	0.1		D	0	0
	E	0.2	0.2		E	0.1	0.1		E	0.2	0.2		E	0	0
	A	0	0		A	0.1	0.1		A	0	0		A	0.2	0.2
	B	-0.4	-0.4		B	0	0		B	0.1	0.1		B	0.2	0.2
23	C	0.3	0.3	28	C	-0.4	-0.4	33	C	-0.4	-0.4	38	C	-0.7	-0.7
	D	0	0		D	0.1	0.1		D	0.1	0.1		D	0.1	0.1
	E	0.1	0.1		E	0.2	0.2		E	0.1	0.1		E	0.2	0.2
	A	0.1	0.1		A	0.1	0.1		A	0.1	0.1		A	0	0
	B	0.1	0.1		B	0	0		B	0	0		B	0	0
24	C	0.2	0.2	29	C	0.2	0.2	34	C	-0.5	-0.5	39	C	0	0
	D	-0.3	-0.3		D	-0.3	-0.3		D	0.1	0.1		D	0	0
	E	-0.1	-0.1		E	0	0		E	0.2	0.2		E	0	0
	A	0.1	0.1		A	0.1	0.1		A	0.1	0.1		A	0.3	0.3
	B	0.1	0.1		B	-0.5	-0.5		B	0.1	0.1		B	-0.6	-0.6
25	C	-0.4	-0.4	30	C	0.3	0.3	35	C	0.1	0.1	40	C	0.2	0.2
	D	0.1	0.1		D	0	0		D	-0.4	-0.4		D	0.1	0.1
	E	0.2	0.2		E	0	0		E	0	0		E	0	0

QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.	QST.	OPT.	SOFT. Dist.	MAN. Dist.
	A	-0.2	-0.2		A	-0.3	-0.3		A	0	0		A	0.2	0.2
	B	0	0		B	0	0		B	0.1	0.1		B	-0.6	-0.6
41	C	0	0	46	C	0	0	51	C	0.1	0.1	56	C	0.2	0.2
	D	0	0		D	0	0		D	0	0		D	0.1	0.1
	E	0.1	0.1		E	0.2	0.2		E	-0.2	-0.2		E	0.2	0.2
	A	0.1	0.1		A	0	0		A	0.1	0.1		A	0.1	0.1
	B	-0.6	-0.6		B	0	0		B	0.1	0.1		B	0.1	0.1
42	C	0.2	0.2	47	C	0	0	52	C	-0.5	-0.5	57	C	0.1	0.1
	D	0.1	0.1		D	0	0		D	0.1	0.1		D	-0.4	-0.4
	E	0.1	0.1		E	0	0		E	0.2	0.2		E	0.1	0.1
	A	0	0		A	0.2	0.2		A	0.2	0.2		A	0.1	0.1
	B	0	0		B	-0.5	-0.5		B	0.2	0.2		B	-0.5	-0.5
43	C	0	0	48	C	0.2	0.2	53	C	-0.8	-0.8	58	C	0.3	0.3
	D	0	0		D	0	0		D	0.2	0.2		D	0	0
	E	-0.1	-0.1		E	0.1	0.1		E	0.2	0.2		E	0	0
44	A	0.1	0.1	49	A	0	0	54	A	0	0	59	A	0.1	0.1

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

	B	0.1	0.1		B	0	0		B	-0.5	-0.5		B	-0.4	-0.4
	C	-0.4	-0.4		C	0	0		C	0.3	0.3		C	0.2	0.2
	D	0.1	0.1		D	0	0		D	0	0		D	0.1	0.1
	E	0.1	0.1		E	0	0		E	0.1	0.1		E	0	0
45	A	0.1	0.1	50	A	-0.4	-0.4	55	A	-0.1	-0.1	60	A	-0.2	-0.2
	B	0.1	0.1		B	0.2	0.2		B	0.1	0.1		B	0	0
	C	0.1	0.1		C	0.1	0.1		C	0.1	0.1		C	0	0
	D	-0.3	-0.3		D	0.1	0.1		D	0.1	0.1		D	0	0
	E	0	0		E	0	0		E	-0.3	-0.3		E	0.2	0.2

*SOFT. = indices obtained from the developed software

*MAN. = indices obtained by manual calculation

*QST. = question (item) numbers

*Dist. = distractor indices

Summary of Results

The analyses revealed that:

1. The developed software is 100 percent accurate in computing the difficulty indices of the test items. It has a very high rate of accuracy.
2. The developed software is 100 percent accurate in computing the discrimination indices of the test items. It has a very high rate of accuracy.
3. The developed software is 100 percent accurate in computing the distractor indices of the test items. It has a very high rate of accuracy.

DISCUSSION OF FINDINGS

Accuracy of the Developed Software in Computing Difficulty Indices

The result showed that the item difficulty indices computed with the developed software were exactly the same as those calculated manually. This indicates 100% accuracy, which indicates a very high level of accuracy of the software in computing the difficulty indices of test items (Table 1). This result is expected and not surprising because the instrument used to establish the accuracy of the developed software in this study is an objective test. An objective test yields the same result whether scored and computed manually or electronically. Secondly, the packages (PHP, MYSQL and HTML) used to develop the software package in this study (IASP) have been confirmed to be highly efficient in software programming. Since the algorithms were correctly coded and the raw data obtained from the field were accurately fed into the system, therefore it is expected to produce accurate indices.

The finding of the present study is in agreement with that of Osuo-Genseleke (2016) who conducted a study on deploying data mining algorithm for rule discovering and decision making. The result of his study revealed that accuracy improved from 94.6% in the existing software to 96.5% in the developed software.

A finding which the present study is discordant with is that of Ugwu (2016) who developed an application of data mining for the prediction of election results. The findings revealed that there was an average level of accuracy in the prediction of the election results using the developed software. This divergent result from

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

the present study may be attributed to the fact that the data used in the prediction of election results are highly subjective while the data used in computing the difficulty indices are highly objective. Also, errors could have such been introduced while feeding data into the existing software, but s errors were completely eliminated in typing data into the present software.

Accuracy of the Developed Software in Computing Discrimination Indices

This result implies 100% accuracy which indicates a very high level of accuracy of the software in computing the difficulty indices of test items (Table 2). This result is expected and not surprising because the instrument used to establish the accuracy of the developed software in this study is an objective test. An objective test yields the same result whether scored and computed manually or electronically. Secondly, the packages (PHP, MYSQL and HTML) used to develop the software package of in this study (IASP) have been confirmed to be highly efficient in software programming. Since the algorithms were correctly coded and the raw data obtained from the field were accurately fed into the system, therefore it is expected to produce accurate indices.

The finding of the present study is in agreement with that of Babatubo (2016) who developed an automated online manager and result generator. The system had high accuracy in generating results and in the performance of auto grading of students based on the questions answered correctly.

However, findings which the present study is discordant with is that of Ugwu (2016) who developed an application of data mining for the prediction of election results and the findings revealed that there was an average level of accuracy in prediction of the election results using the developed software. The divergent results from the present study may be attributed to the fact that the data used in the prediction of election results are highly subjective while the data used in computing the difficulty indices are highly objective. Also, errors could have been introduced while feeding data into the existing software, but errors were completely eliminated in typing data into the present software.

Accuracy of the Developed Software in Computing Distractor Indices

This result implies 100% accuracy, which indicates a very high level of accuracy of the software in computing the distractor indices of test items (Table 6). This result is expected and not surprising because the instrument used to establish the accuracy of the developed software in this study is an objective test. An objective test yields the same result whether scored and computed manually or electronically. Secondly, the packages (PHP, MYSQL and HTML) used to develop the software package of in this study (IASP) have been confirmed to be highly efficient in software programming. Since the algorithms were correctly coded and the raw data obtained from the field were accurately fed into the system, therefore it is expected to produce accurate indices.

The finding of the present study is in agreement with that of Salako (2016) who developed an e-commerce sales forcasting model using Bayesian Network. The project was designed to efficiently predict future values, having a significant and positive impact on sales and operations as well as overall financial health of e-commerce business. There was fairly high rate of prediction of sales which was based on probability of demand rate in a particular month.

However, findings which the present study is discordant with is that of Ugwu (2016) who developed an application of data mining for the prediction of election results and the findings revealed that there was an average level of accuracy in prediction of the election results using the developed software. The divergent results from the present study may be attributed to the fact that the data used in the prediction of election results are highly subjective while the data used in computing the difficulty indices are highly objective. Also, errors could have been introduced while feeding data into the existing software, but errors were

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

completely eliminated in typing data into the present software.

CONCLUSIONS

From the results obtained in this study, conclusion is drawn that the developed software gives very high level of accuracy in computing the item indices of test instruments at one decimal place. These item indices are the difficulty index, the discrimination index and the distractor index. Any error that may be seen in the results generated by the system may arise from the user and not from the developed software.

RECOMMENDATIONS

With regard to the results of this study, the researcher recommends that:

1. Since the developed software is very accurate and would likely be affordable, researchers, evaluators and educational administrators should use it in computing the difficulty indices of their test instruments.
2. In the same vein, since the developed software has a very high level of accuracy and also affordable, researchers, evaluators and educational administrators should use it in computing the discrimination indices of their test instruments.
3. Lastly, the manual computations of the distractor indices of the options of test items should be replaced with electronic computation using the developed software.

REFERENCES

1. Agu, N. (2014). Basic statistics for education and behavioural sciences. Awka: J'Goshen.
2. Alvior, M. G. (2013). What are the psychometric properties of a research instrument. Retrieved April 10, 2018 from http//simplyeducateme.
3. Babatubo, I. O. (2016). Automated online examination manager and result generator. Unpublished Masters Thesis, Department of Computer Science, University of Port Harcourt.
4. Carlenton State University (2016). Item analysis. Retrieved June 24, 2016 from https://carleton.ca/edc/wp-content/uploads/Item-Analysis.pdf.
5. Fidelis, I (2014). Comprehensive guide to test construction and administration. Omoku: Chifas Nigeria.
6. Jones, D., Roach, P., Setter, J. & Esling, J. (2008). Cambridge advanced learner's dictionary. Cambridgeshire: Cambridge University Press.
7. Kpolovie, P. J. (2010). Advanced research methods. Owerri: Springfield Publishers Ltd.
8. Kpolovie, P. J. (2014). Test measurement and evaluation in education. (2nd Ed.) Owerri: Springfield Publishers Ltd.
9. Mitchell, A. S. (2014). Online with computers, book 5. Rasmed Publications Limited. India.
10. Nworgu, B. G. (2015). Educational research: Basic issues and methodology. (3rd ed.). Nsukka: University Trust Publishers.
11. Orluwene, G. (2014). Introduction to test theory and development process. Port Harcourt: Chris-Ron Integrated Services.
12. Osu-Genseleke, M. (2016). Deploying data mining algorithm for rule ddiscovery and decision making. Unpublished Masters Thesis, Department of Computer Science, University of Port Harcourt.
13. *Professional Testing (2015).* Item analysis index. Retrieved June 24, 2016 from http://mededuunit.blogspot.com.ng/2015/07/the-difficulty-index-and-discrimination. html.
14. Salako, A.O. (2016). E-commerce sales forecasting model using Bayesian network. Unpublished Masters Thesis, Department of Computer Science, University of Port Harcourt.
15. Shuttleworth, M. (2009). Types of validity. Retrieved March 7, 2016 from

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue III March 2024

https://explorable.com/types-of-validity.

16. Statistics Solutions (2015). Advancement through clarity. Retrieved July 3, 2016 from http://www.statisticssolutions.com/point-biserial-correlation/.
17. The Free Dictionary (2015). Psychometric properties. Retrieved December 29, 2015 from http://medicaldictionary. thefreedictionary.com/psycho metric +properties.
18. Ugwu, F. C. (2016). Application of data mining for prediction of election results. Unpublished Masters Thesis, Department of Computer Science, University of Port Harcourt.