PSYCHOLOGY AND EDUCATION (2022) 59(1): 612-624
ISSN: 1553-3969

Split-half method of estimating reliability index: a comparison of the odd-even and first-and second-half ways of splitting test items.

OSSAI, Peter Agbadobi Uloku (Ph.D) (corresponding author)
Email: agbadobipeter@g.mail.com
Cell Phone Number: 08063721891
Senior Lecturer, Department of Guidance and Counselling, Delta State University, Abraka, Nigeria. The author's bachelor's degree was in mathematics education while master and doctorate degrees were earned in measurement and evaluation.

ENWEFA, Chiekem
email:Chiekemenwefa@gmail.com
Cell Phone Number: 07035269009
Lecturer 1, Department of Guidance and Counselling, Delta State University, Abraka, Nigeria. The author had his bachelor's degree in mathematics education. The master and doctorate degrees were both earned in measurement and evaluation.

ABSTRACT

Sequel to the limited knowledge of the split-half method of estimating reliability co-efficient by many post-graduate students of Delta State University, Abraka, this study sought to compare the odd-even and first- and second-half ways of splitting mathematics test items into equivalent halves. The purpose was to enlighten our post-graduate students on the choice of the method of splitting test items into two equivalent halves. A sample of 105 students were selected from the final year secondary school students in Edo and Delta states of Nigeria. The two states are in the south-south geo-political zone of Nigeria. They also made up the defunct Bendel State. A total of thirty-five students was drawn from a secondary school in Edo State. The same number of students was selected from each of the two schools selected from Delta State. The mathematics test items examined by a regional examination body, the West African Examinations Council for May/June 2009, 2010 and 2011 were used as instrument for the study. A correlational survey design was employed in the study. Although the two means of splitting the

mathematics test items showed significant values for (except for the first- and second-half means for 2009), the r-values for the odd-even means were considerably higher compared to the first- and second-half means. The use of the Spearman Brown's prophecy formula to step up the half tests further confirmed the results. Bearing in mind that reliability estimates are sample-dependent, the researcher suggested that further studies be done using other states in Nigeria. It may also be advisable for other measuring devices such as NABTEB and NECO to be employed.

Key words: reliability, split-half, comparison, different, halves.

INTRODUCTION

Testing is an inevitable exercise in the teaching/learning process. That is why learning is incomplete until the achievement of learners is assessed. The assessment of learners leads to test scores as Ossai (2016) stated that test scores come after the administration of a test which may be teacher-made or standardized.

Considering the important uses of a test in terms of classification, certification, diagnosis prediction and recruitment (Okorodudu, 2012; Anastasi & Urbina, 2007), one has to emphasize that any measuring instrument should not be faulty. Psychometricians agree (Best & Kahn, 2007) that reliability and validity are essential to the effectiveness of any data-gathering process.

This work is limited to the examination of reliability, with emphasis to the examination procedure for establishing index of reliability. Zhu and Han (2011) viewed reliability as a measure that is reliable to the extent that independent but complete comparable measures of the same traits or construct of a given object agree. According to the authors, reliability is dependent on how much variation in scores that is associated with random or chance errors. In line with this view of reliability by these authors, Best and Kahn (2007) opined that a test is reliable to the extent that it measures whatever it is measuring consistently. Reliability also means consistency (Livingston, 2018).

Factors Affecting the Reliability of a Test

The National Teachers' Institute (NTI, 2000), Livingston (2018) and Disha (2020) spelled out

some of the factors that influence the reliability of test scores as:

i) Length of the test;
ii) Homogeneity of items;
iii) Difficulty value of item;
iv) Discrimination index;
v) Test instructions and clarity of the questions
vi) Item selection,
vii) The state of the testee;
viii) Environmental conditions;
ix) Guessing and chance errors;
x) Interval between the occasions the test is taken.

The NTI (2000) suggested some measures to be taken in order to make test scores reliable. These include simple and clear language in setting questions and giving instructions, using an objective method of scoring, guarding against cheating and eliminating guessing. Others are moderate difficulty levels of items, having a reasonable number of test items, a conducive testing environment and not prolonging the interval between testing occasions.

Relating reliability to sampling, Livingston (2018) opined that when a testee makes his responses to test items, he/she is faced with only a sample of the questions or problems that could have been included. Another edition of the test, according to the scholar, presents a different sample of questions or problems to the teetee. Livingston also posited that relevant increase in the number of questions or problems in a test brings about a better sample of a testee's performance; likewise increase in the number of qualified raters will lead to a better sample of raters' judgements of the responses from the testees.

The point made by Livingston (2018) regarding the result of increasing the number questions and the number of raters to enhance the performance of testees is analogous to the relationship established among measurement, population, sample and mean by Gravetter and Wallnau (2004). Gravetter and Wallnau posited that when a sample is seen as a measurement of a population, then a sample mean is a measurement of a population mean. The scholars explained that if the means from different samples are almost identical, then the sample mean provides a measure of the population that is reliable. The scholars added that the sample mean provides an unreliable measure of the population mean should there be considerable differences from one sample to another.

Anastasi and Urbina (2007) identified types of reliability thus: test-retest, alternate-form, split-half, Kuder-Richardson and Scorer. The treatment of the split-half reliability is the focus of this study.

The split-half method involves a single administration of the instrument (Nworgu, 2006). According to the scholar, from the single administration, two sets of scores are obtained by splitting the test into two equivalent halves. Nworgureported that the equivalent halves can be odd-even or first-half versus second-half; the different ways of dividing the test will affect the index of reliability Anastasi and Urbina (2007) posited that the first-half and second-half method of splitting the items has a shortcoming arising from differences and difficulty levels and such factors as practice, fatigue, boredom and warming up on the part of the testees. The scholars opined that the odd-even method is regarded as the most adequate in most cases. Gronlund (1981) pointed out that the inability of the split-half method to the changes in the individual from one occasion to another is one of the limitations.

Experience as a researcher and as a lecturer as well as a rater of students' dissertations and theses at the Faculty of Education, Delta State University, Abraka, Nigeria, indicates that, more often than not students use the odd-even style of splitting a test in their studies. This decision is often taken without regard to the short-comings identified by Anastasi and Urbina (2007). A closer look at some of these measuring devices employed by our post-graduate students in estimating internal consistency of test items shows that the item difficulty and item discrimination indices are within the range of testees. The students in question also fail to indicate why they have chosen the odd-even method of splitting a test. Thus, this researcher has deemed it necessary to compare the indices arising from using the odd-even and first-half versus second half systems of splitting test items in cases where the split-half reliability is applied.

Statement of the Problem

Post-graduate students of the Faculty of Education, Delta State University, Abraka, Nigeria often conduct reliability estimates regarding the internal consistency of test items. When they use the split-half method of establishing reliability, they often use the odd-even system of splitting the test. They take this decision

without consideration for the item difficulty levels of the test items. Experience has shown that the test items they use have moderate levels of difficulty, yet they opt for the odd-even system without giving any reason for their choice. Available literature shows that another way to divide a test into two equivalent halves is the first half and second half method. Where the difficulty level of items is moderate, the method is comparable to the odd-even system of splitting a test. Which of the two systems produces a higher reliability co-efficient? A comparison of the two systems of splitting a test into two equal halves has become necessary to enlighten our post-graduate students.

Research questions and hypotheses

The study addressed the following research questions and hypotheses:

1. What is the index of relationship between the two halves of the 2009 mathematics test items using odd-even method of splitting?
2. What is the index of relationship between the two halves of the 2009 mathematics test items using first half and second half method of splitting?
3. What is the index of relationship between the two halves of the 2010 mathematics test items using odd-even method of splitting?
4. What is the index of relationship between the two halves of the 2010 mathematics test items using first half and second half method of splitting?
5. What is the index of relationship between the two halves of the 2011 mathematics test items using odd-even method of splitting?
6. What is the index of relationship between the two halves of the 2011 mathematics test items using first half and second half method of splitting?

Hypotheses

1. There is no significant relationship between the two halves of the 2009 mathematics test items using odd-even method of splitting.
2. There is no significant relationship between the two halves of the 2009 mathematics test items using first half and second half method of splitting.
3. There is no significant relationship between the two

halves of the 2010 mathematics test items using odd-even method of splitting.

4. There is no significant relationship between the two halves of the 2010 mathematics test items using first half and second half method of splitting.
5. There is no significant relationship between the two halves of the 2011 mathematics test items using odd-even method of splitting.
6. There is no significant relationship between the two halves of the 2011 mathematics test items using first half and second half method of splitting.

Methodology

A correlational survey design was employed in the study with a population of final year secondary school students chosen from government-owned senior secondary schools in Edo and Delta States. These are two states (from the defunct Bendel State) that make up the six states in the south-south geo-political zone of Nigeria.The mathematics test items examined by the West African Senior School Certificate Examinations (WASSCE) for May/June 2009, 2010 and 2011 constituted the instrument used in the study. The WASSCE is a regional examining body in West Africa made up of Nigeria, Gambia, Ghana and Sierra Leone. A standardized mathematics instrument, the WASSCE General Mathematics/Mathematics (core) 1 (May/June) 2009, 2010 and 2011 consists of 50 items each, to be responded to within 90 minutes. Thirty-five students selected from a secondary school in Edo State responded to the mathematics test items for May/June 2009. The same number of students was chosen from each of the two secondary schools sampled from Delta State.

The selectionof students was done through simple random sampling technique of balloting. For each of these years, students' responses were divided into two parts for the purpose of analysis using the split-half method of establishing reliability. The researcher subjected students'responses for each of the years to odd-even and first half – second half means of estimating reliability index under the split-half method. Thus, the product moment correlation technique was applied to the two means of estimating the index of reliability.The two r-values for each of the years were compared to

find out which was higher. A test of significance was also conducted for each of the years to know which of the means had a better measure of consistency. Spearman Brown's prophecy formula was not applied since the study focused on the two halves of each of the instruments.

Data Analysis and Presentation of Results

Research question one:

What is the index of relationship between the two halves of the 2009 mathematics test items using odd-even method of splitting?

Table 1: showing mean, standard deviation and the index of relationship between the odd-even halves of the test.

Variables	N	$\bar{x}$	SD	r
Odd	35	8.8	2.3	.507
Even	35	9.4	2.9	

Table 1, shows that the index of relationship between the two halves is .507. The odd half has a mean of 8.8 and a standard deviation of 2.3while the even half has 9.4 and 2.9 respectively, as the mean and standard deviation.

Research question two

What is the index of relationship between the two halves of the 2009 mathematics test items using first-half and second half method of splitting?

Table 2, showing mean standard deviation and the index of relationship between the first half and the second half of the test?

Variables	N	$\bar{x}$	SD	r
First half	35	9.5	2.8	.153
Second half	35	11.3	5.7	

Table 2, shows that the index of relationship between the two halves of the test is .153. The first half has a mean of 9.5 and a standard deviation of 2.8 while the second half has a mean of 11.3 and a standard deviation of 5.7

Research question three

What is the index of relationship between the two halves of the 2010 mathematics test items using odd-even method of splitting?

Table 3: showing mean, standard deviation and the index of relationship between the two halves of the test.

Variables	N	$\bar{x}$	SD	r
Odd	35	7.5	3.5	
				.595
Even	35	8.7	3.0	

Table 3: shows that the index of relationship between the two halves of the test is .595;the mean and standard deviation for the odd half are 7.5 and 3.5 respectively. The mean and standard deviation for the even half are 8.7 and 3.0, respectively.

Research question four

What is the index of relationship between the two halves of the 2010 mathematics test items using first half and second half method of splitting?

Table 4: showing mean, standard deviation and the index of relationship between the two halves of the test.

Variables	N	$\bar{x}$	SD	r
First half	35	9.3	3.3	
				.437
Second half	35	7.0	3.5	

Table 4 shows that the index of relationship between the two halves of the test is .437. The mean and standard deviation of the first half are 9.3 and 3.3, respectively. Likewise, the second half has a mean and standard deviation of 7.0 and 3.5 respectively.

Research question five

What is the index of relationship between the two halves of the 2011 mathematics test items using odd-even method of splitting?

Table 5: showing mean, standard deviation and the index of relationship between the two halves of the test.

Variables	N	$\bar{x}$	SD	r
Odd	35	8.7	3.7	
				.733

Even	35	11.8	4.6	

Table 5 shows that the index of relationship between the two halves of the test is .733;the mean and standard deviation for the odd half are 8.7 and 3.7 respectively. For the even half, the mean and standard deviation are 11.8 and 4.6, respectively.

Research question six

What is the index of relationship between the two halves of the 2011 mathematics test items using first half and second half method of splitting?

Table 6: showing mean, standard deviation and the index of relationship between the two halves of the test.

Variables	N	$\bar{x}$	SD	r
First half35	10.5	4.1		
				.451
Second half35	10.2	4.9		

Table 6 shows that the index of relationship between the two halves of the test is .452 while the mean and standard deviation for the first half are 10.5 and 4.1 respectively. Likewise, 10.2 and 4.9 are the mean and standard deviation, respectively for the second half.

Hypotheses testing

Hypothesis one:

There is no significant relationship between the two halves of the 2009 mathematics test items using odd-even method of splitting.

Table 7: showing index of relationship between the two halves of the test.

Variables	N	r	Significant
Odd	35		
.507	.002		
Second half	35		

Table 7 indicates that r is .507. Given an alpha level of .05, the p-value is .002; the null hypothesis is rejected since the p-value is less than the alpha level. This implies that a significant relationship exists between the two halves of the test.

Hypothesis two:

There is no significant relationship between the two halves of the 2009 mathematics test items using first half and second half method of splitting.

Table 8: showing index of relationship between the two halves of the test.

Variables	N	r	Significant
Odd	35		
.153	.379		
Second half	35		

Table 8 shows that r is .153. Given an alpha level of .05 of significance, the p-value is .379; since the p-value is greater than the alpha level, the null hypothesis is upheld. That is, there is no significant relationship between the two halves of the test.

Hypothesis three:

There is no significant relationship between the two halves of the 2010 mathematics test items using odd-even method of splitting.

Table 9: showing index of relationship between the two halves of the test.

Variables	N	r	Significant
Odd	35		
		.595	.000
Even 35			

From table 9, r is .595 while p-value is .000; given an alpha level of .05, the hull hypothesis is rejected since the p-value is less than the level of significance. This means that there is a significant relationship between the two halves of the test.

Hypothesis four:

There is no significant relationship between the two halves of the 2010 mathematics test items using first half and second half method of splitting.

Table 10: showing index of relationship between the two halves of the test.

Variables	N	r	Significant
First half 35			
	.437	.009	
Second half	35		

Table 10 indicates that r is .437. At an alpha level of .05, the p-value is .009; since the p-value is less than the alpha level of .05, the null hypothesis is rejected. Thus, there is a significant relationship between the two halves of the test.

Hypothesis five:

There is no significant relationship between the two halves of the 2011 mathematics test items using odd-even method of splitting.

Table 11: showing index of relationship between the two halves of the test.

Variables	N	r	Significant
Odd	35		
		.733	.000
Even	35		

From table 11 r is .733; at an alpha level of .05, the p-value is .000. Since the p-value is smaller than the alpha level of significance, the null hypothesis is rejected. This implies that a significant relationship exists between the two halves of the test.

Hypothesis six:

There is no significant relationship between the two halves of the 2011 mathematics test items using first half and second half method of splitting.

Table 12: showing index of relationship between the two halves of the test.

Variables	N	r	Significant
First half	35		
.451	.007		
Second half	35		

Table 12 indicates that r is .451 while the p-value is .007; at an alpha level of .05, the null hypothesis is rejected because the p-value is less than the alpha level of significance. This means that a significant relationship exists between the two halves of the test.

Discussion of findings

Literature suggests that one of the best ways of splitting test items when using the split-half method to establish reliability co-efficient is the odd-even means. As discussion on the findings from this study shows the findings appear to be in that direction.

Research questions one and two shows r-values to be respectively .507 and .153 for the odd-even and the

first-second halves means of splitting a test. The difference between the two indices is conspicuously in favour of the odd-even means. The trend continued in the responses to the 2010 mathematics test items where r-values are .595 and .437 for the odd-even and first-second halves means, respectively. Likewise, the r-value for the odd-even means in the responses to the 2011 mathematics test items was .733, as against .451 for the first-second half means. The different values for the two ways of splitting a test corroborate the view of Nworgu (2000) that the different ways of splitting a test will affect the co-efficient of reliability. Specifically, Anastasi and Urbian (2007) posited that first-second half means of splitting a test is affected by item difficulty index and such factors as practice, fatigue, boredom and warming up on the part of the examinees.

The hypotheses tests for the two ways of splitting the mathematics test items showed that a significant relationship exists between the odd-even halves of the test items for 2009, 2010 and 2011. Same applies to the first-second halves splitting except for 2009 test items where the test of hypothesis was not significant in the first-half method of splitting. Even though the tests of significance showed almost the same result for the two means of splitting the test, the effect sizes were markedly different. The sizes of the relationship for each of the years in the order of odd-even and first-second halves are .251 and .023 for 2009, .354 and .190 for 2010 and .537 and .203 for 2011. There is an appreciable difference in sizes of relationship when odd-even and first-second halves of splitting the mathematics test items are compared. For instance, the size of the odd-even half is more than ten times that of the first-second half in the 2009 test items. Similarly, the size of relationship for the odd-even half is about two times that of the first-second half in the 2010 test items. Likewise, the strength of relationship for the odd-even half is about five times that of the first-second half in the 2011 mathematics test items. The application of Spearman Brown's Prophecy formula further shows the appreciable difference in the two ways of splitting the test.For 2009 , the r-values for the entire test were 0.67 and 0.27 for the odd-even and first-second halves respectively. In the same order, the r-values for the whole test are 0.75 and 0.61 for 2010;

likewise, the r-values for the entire test in the 2011 mathematics test items were 0.35 and 0.62 respectively.

The apparent low indices in the first-second half means of splitting the mathematics test items may not be caused by the factors observed by Anastasi and Urbina (2007). The testing environment and the sample-dependent nature of test scores may have contributed. More often than not, paper-and pencil tests in Nigeria are written in harsh conditions. Candidates may have to wait for administrative protocols to be completed before settling down for a test (which spills over to any time of the day). Zhu and Hen (2011) observed that examination candidates do better in the morning than in the afternoon. Apart from this the test scores may depend on the sample used. Danner (2016)posited that an instrument is capable of yielding measurements of different levels of reliability when different samples are used. Perhaps, this may apply in this study.

Conclusion and Recommendation

The study compared the odd-even and first- and second-half ways of splitting mathematics test items into two equivalent halves when the split-half method of estimating reliability co-efficient is used. Although, nearly all the r-values for the two ways were significant, the r-values for the odd-even means of splitting the test items for the three years were considerably higher than those of the first- and second-half means of splitting.

This researcher will not hastily conclude that the odd-even option is better than the first – and second-half option. This is because test scores can be sample-dependent. Admittedly, candidates may contend with the challenges of fatigue and testing environment in their responses to the second part of first-and second-half means of splitting, suffice it to say that reliability index is subject to the sample used (Danner, 2016).

To guide the post-graduate students of Delta State University, Abraka, it is hereby suggested that studies be conducted in this area using different samples from other states in Nigeria. Other instruments such as the National Examinations Council (NECO) and the National Business and Technical Examination Board (NABTEB).

REFERENCES

Anastasi, A. & Urbina, S. (2007). *Psychological testing* (7th ed.). New Delhi:

Prentice-Hall of India.

Best, J.W. & Kahn, J.V. (2007). *Research in education* (9th ed.).New Delhi:
Prentice-Hall of India.

Danner, D. (2016). Reliability – the precision of a measurement. *GESIS Survey*
Guidelines. Mannheim, Germany: GESIS-Leibniz Institute for the Social Sciences, doi: 10.15465/gesis-org-en-011.

Gavetter,F.J. & Wallnau, L.B. (2004). *Statistics for the behavioural sciences* (6th
ed.). Australia: Thomson Wadsworth.

National Teachers' Institute (2000). *Education.* Kaduna: NTI

Nworgu, B.G. (2006). *Educational research.* (2nd ed.). Nsukka: University Trust
Publishers.

Okorodudu, R.I. (2012). *Understanding educational and psychological*
measurement and evaluation. Abraka: University Press.

Ossai, P.A.U. (2016. Test scores and skewness: implications to examiners and students.*Journal of Educational Research and Policies*, 11(2), 156-160.

Zhu, J. &Han, L. (2011). Analysis of the main factors affecting the reliability of
test papers. *Journal of Language Teaching and Research,* 2(1), 236 – 238.