Comparative Analyses of Difficulty and Discrimination Indices of Islamic Studies Courses as Quality Indicator of Examination

[1]Garba Alhaji Muhammad
Department of Psychology, School of Education,
Aminu Saleh College of Education,
Azare, Bauchi State
Email: naziabba93@gmail.com
GSM: 08036962442

[2]Abdullahi Uthman Imam
Department of Islamic Studies, School of Arts and Social Sciences,
Aminu Saleh College of Education, Azare, Bauchi State
Email: abdallahazare@gmail.com
GSM: 08060679705

DOI: 10.56201/ijee.v10.no2.2024.pg31.39

Abstract

This paper analyses the items of 2019/2020 Qur'an and Hadith Courses examinations of Islamic Studies department, school of undergraduate studies, Aminu Saleh College of Education, Azare, Bauchi State. The paper reviewed the concepts of assessment and testing, the qualities desired of a good assessment tool such as, validity and reliability. Also reviewed was item analysis with special reference to, item difficulty and item discrimination ability. The study employed an ex post facto design where responses of students in 2019/2020 Qur'an and Hadith Courses of Islamic Studies department, school of undergraduate studies, ASCOE, Azare were used. Sample of 18 items for each was used. The findings of the study revealed that 11(61%) out of 18 items of the Qur'anic examinations were of moderate difficulty while, only 7(39%) items in the Hadith examinations. On the average, both Qur'anic and Hadith examinations were found to be of moderate difficulty 0.56 and 0.44 respectively. The findings also revealed that, Qur'anic courses examinations were found to be of marginal discrimination (0.21) while Hadith courses examinations were poor (0.13) in discriminating between the high and low achievers. Finally, the paper concludes that Qur'anic courses examinations are more appropriate than the Hadith courses examinations. It is therefore recommended among others that, items found to be difficult, easy and or non-discriminating be reviewed or discarded.

Keywords: *Difficulty Indices, Discrimination Indices, Analysis, and Quality Indicators*

Introduction

Assessment of pupils and students performance is inevitable in education at all levels. No meaningful teaching and learning can take place in the absence of proper assessment of students before, during and after instruction. Assessment is a main step in the process of education by which the academic performance of students during a course attendance is tested. It can be considered as an educational tool which determines the competency of students in educational improvement as well as the gap between educational aims and the degree of learning. It is a process used in collecting information on events, objects and or people, particularly on human behavior which is used to evaluate the quality of work done. When assessment is applied to education, it is an all-embracing term covering any of the situations in which some aspects of pupils' education are measured by the teacher and the success of their instructional practices. Assessment is a means whereby the teacher obtains information about knowledge gains, behavioural changes and other aspects of the development of learners (Oguneye, 2002). It involves the deliberate effort of the teacher to measure the effect of the instructional process as well as the overall effect of school learning on the behaviour of students. Assessment covers all aspects of school experience both within and outside the classroom. It covers the cognitive as well as the affective and psychomotor aspects of learning.

Testing is a fundamental part of the teaching-learning process used not only as a basis for assessment at the end of the teaching and learning process but to guide teaching, and aid in the development of curriculum, as well as in the determination of learning difficulties, level of mastery and differences among students. Testing has been fully accepted in modern societies as the most objective method of decision making in schools, industries and government establishments. It is now used for admission, recruitment, promotion, placement, evaluation, guidance research and teaching purpose among others (Emaikwu, 2011). In Nigeria, achievement at any level of education is awarded with certification of those who successfully completed a course of study with good academic records. Educational test results are important parameters by which society voices the product of its educational system. As such, educational institutions are expected to conduct tests that will enable them establish the desired characteristics of their examinees. Therefore, developing an appropriate assessment strategy is a key part of effective sustainable curriculum development. It is against this background; this study aims to find out indices that will inform all stakeholders the functionality or otherwise of the instrument. Item analysis is yhr valuable and powerful technique available to teaching professionals and instructors for the guidance and improvement of their instructions. This is so because it enables instructors to increase their test construction skills, identify specific areas of course content which need greater emphasis or clarity, and improve other classroom practices.

Item Analysis

Item analysis is a process which examines students' responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. It also "investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test" (Thompson & Levitov, 1985, McCowan & McCowan 1999). Murphy and Davidshofer (1988) argued that item analysis can eventually augment the level of the way test constructer and users understand the test. Close examination of individual items is

of paramount importance to understanding why a test shows a specific level of validity and reliability. A good item analysis therefore can be very good informative when tests fail to show expected level of validity and reliability. Meanwhile it is a clear indicator of the reasons why a test is valid and reliable or otherwise.

According to Anastasi and Urbina (1997) items could be analyzed qualitatively, in terms of their content and form, and quantitatively, in terms of their statistical properties. Qualitative analyses include the consideration of content validity as well as effective evaluation of item writing procedures. However, quantitative analysis dealt principally with the measurement of item difficulty and item discrimination. Both the validity and reliability of any test depend ultimately on the characteristics of the items. Therefore, high validity and reliability can be built into a test in advance through item analysis.

Item Analysis Statistics

An item analysis involves many statistics that can provide useful information for determining the validity and improving the quality and accuracy of test items. These statistics are used to measure the ability levels of examinees from their responses to each item and they are as follows:

i) Item Difficulty: Item difficulty is a measure of the difficulty of an item. For items (that is, multiple-choice items) with one correct alternative worth a single point, the item difficulty (also known as the item difficulty index, or the difficulty level index, or the difficulty factor, or the item facility index, or the item easiness index, or the p -value) is defined as the proportion of respondents (examinees) selecting the answer to the item correctly.

ii) Item Discrimination: The item discrimination (or the item discrimination index) is a basic measure of the validity of an item. It is defined as the discriminating power or the degree of an item's ability to discriminate (or differentiate) between high achievers (i.e, those who scored high on the total test) and low achievers (i.e, those who scored low), which are determined on the same criterion.

iii) Test Item Distractor Analysis: It is an important and useful component of test item analysis. A test item distractor is defined as the incorrect response options in a multiple choice test item. It was observed that, there is a relationship between the quality of the distractors in a test item and the student performance on the test item, which also affect the student performance on his/her total test score. The performance of these incorrect item response options can be determined through the test item distractor analysis frequency table which contains the frequency, or number of students, that selected each incorrect option.

The item difficulty and discrimination indices were determined using percentage sample, scripts were arranged in descending order of performance of the examinees and the first 27% of the scripts called upper group 'NU' and the last 27% of the scripts called lower group 'NL' were taken using the formulae:

$$\text{Difficulty Index (P)} = \frac{R_U + R_L}{N_U + N_L}$$

$$\text{Discrimination Index (DI)} = \frac{R_U - R_L}{N_U + N_L}$$

Where R_U is the number of examinees in the upper group who got the item correctly, R_L is the number of examinees in the lower group who got the item correctly and N_U and N_L are the number

of examinees in the upper and lower groups respectively. The use of percentage sample was informed based on the findings of Osarumwense and Oyedeji (2015) which revealed that there was no significant difference between the item difficulty and discrimination indices obtained by using the two methods. Therefore, recommend that test developers and analysts should use percentage sample of examinees to compute test item difficulty/discrimination index in order to reduce cost and time as both methods were found to be the same. Justifying the use of 27%, Wiersma and Jurs (1990) stated that 27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis.

Specifically, the criteria used in estimating difficulty and discrimination are as follows as put forward by (Haladyna, Downing & Rodriguez, 2002):

Table 1: Item difficulty and discrimination was based on the interpretation.

Difficulty		**Discrimination**	
Range value	**Interpretation**	**Range value**	**Interpretation**
0.00-0.14	Very Difficult	0.40 and above	Very Good
0.15-0.29	Difficult	0.30-0.39	Reasonably Good
0.30-0.69	Moderately Difficult	0.20-0.29	Marginal
0.70-0.84	Easy	0.19 or Less	Poor
0.85-100	Very Easy		

Validity

Obimba (1989), defines validity as the extent to which the result of a test indicates that the test measures what it is required to measure, all of what it is required to measure and nothing else but what it required to measure. In other words, validity is the extent to which a test is truthful, accurate or relevant in measuring a trait it is supposed to measure. A test intended for example to measure knowledge in economics, the test items must be about Economics and should not include item in any other subject area. In other words a test that is valid for a subject or purpose may be invalid for another subject or for another purpose. Validity therefore, indicates test truthfulness and appropriateness as a measuring instrument. If a test cannot measure accurately what it is expected to measure then it is useless.

Again, validity is a matter of degree because test results possess it to some extent and in varying degrees. No test is completely valid or totally invalid. Test result may for instance have low, moderate, or high validity. Validity is broadly dividedin to content, criterion-related and constructs validity (Sidhu 2005 and Denga 2003).

Content Validity

Content validity as the name indicates is a form of validity which evaluates the relevance of the test items individually and as a whole. Each item should be a sample of knowledge or performance which the test purports to measure and should constitute a representative of sample variable to be tested. **Face Validity and logical validity**

Criterion – Related Validity

Criterion-related validity employs empirical techniques in studying the relationship between scores on a test and some outside criterion. For instance test scores may be correlated with some future measure of success. Two types of criterion validity ware generally identified (Denga, 2003). ***Predictive validity and concurrent validity.***

Construct Validity

A construct is an unobservable (covert) psychological trait, (attribute, characteristic or quality), the existence of which in the individual is assumed to account for some aspect of his observable (overt) behavior. The existence of a construct in the individual is inferred from his observed behavior(s). Examples of constructs are intelligence, anxiety, happiness, critical thinking, honesty, preference musical aptitude etc (Anastasi, 1990; Anikweze, 2005). Construct validity therefore refers to the degree to which a test performance reflects the existence of a psychological trait (contract) in an individual. Test results are analyzed to determine their construct validity. This is done by first identifying the construct which the test results are indicative of, and in what ways the result concurs with the nature of that construct as to be indicative of its existence in the individuals.

Reliability

Broadly speaking, reliability in the language of psychometrics refers to consistency in measurement. It suggests trustworthiness. To the extent that decisions of any kind are to be made, wholly or in part, on the basis of test scores, test users need to make sure that the scores are reasonably trustworthy. Urbina (2004), reliability is based on the consistency and precision of the results of the measurement process. In order to have some degree of confidence or trust in scores, test users require evidence to the effect that the scores obtained from tests would be consistent if the tests were repeated on the same individuals or groups and that the scores are reasonably precise.

Types of Reliability

There are different types and degrees of reliability. A reliability coefficient is an index of reliability, a proportion that indicates the ratio between the true score variance on a test and the total variance (Cohen 2009). Basically, they are: Test-retest, Alternate (or equivalent) forms, Split-half, Scorer reliabilities,

Objectives of the Study

The study was guided by the following objectives, to compare the:

1. items difficulty indices of Qur'an and Hadith courses of the 2019/2020 academic session.
2. overall mean difficult index of Qur'an and Hadith courses of 2019/2020 academic session.
3. items discrimination indices of Qur'an and Hadith courses of the 2019/2020 academic session.
4. the overall mean discrimination index of Qur'an and Hadith courses of 2019/2020 academic session.

Hadith course of Islamic studies department ASCOE, Azare ISL 103, 206, and 305
102, 304, and 404

Methodology

The study was descriptive, cross-sectional conducted at the Islamic Studies Department, Aminu Saleh College of Education, Azare, Bauchi State. The study recruited item analysis reports

of the ISL 102,103, 206, 304, 305 and 404 examinations for 2019/2020 academic session. ISL 102, 304 and 404 and Qur'anic courses while ISL 103, 206 and 305 are Hadith courses. Each examination consisted of 6 items that comprised of 18 items each for both Qur'an and Hadith courses. Data obtained was entered in MS Excel 2019 and analyzed and score of the students was categorized into the high scoring (H) group (top 33%), mid scoring (M) group (middle 34%) and the low scoring (L) group (bottom 33%) respectively, after arranging the scores in descending order.

The result is presented on the basis of providing answers to the research questions raised for the purpose of this research as follows:

Objective One: To compare the items difficulty indices of Qur'an and Hadith courses of the 2019/2020 academic session.

Table 2: Result of the item difficulty indices of Qur'an and Hadith courses of 2019/2020 academic session.

Range of Difficulty Index	Quran Courses Frequency	%	Hadith Courses Frequency	%	Description
0.00 – 0.14	0	0	1	6	Very Difficult
0.15 – 0.29	4	22	7	39	Difficult
0.30 – 0.69	11	61	6	33	Moderately Difficult
0.70 – 0.84	3	17	4	22	Easy
0.85 – 1.00	0	0	0	0	Very Easy
Total	18	100	18	100	

The difficulties index categories were set as 0.00-0.14 (Very difficult), 0.15-0.29 (Difficult), 0.30-0.69 (Moderately Difficult), 0.70-0.84 (Easy) and 0.85-1.00 (Very Easy). Table 1 revealed that from the Quranic courses, only four items representing 22% had item difficulty index less than 0.30, three items that is 17% had difficulty index between 0.70-0.84 regarded as difficult and easy items respectively, the remaining eleven items representing 61%, had item difficulty index within the appropriate range of 0.30-0.69 that is moderate difficulty. The four difficult items are items 1, 3, 6, and 9 with the item difficulty indices of 0.22, 0.19, 0.21 and 0.25 respectively. The implication of these results is that the four items that is items 1, 3, 6, and 9 were too difficult for the examinees. Less than 30% of the examinees were able to get the items correctly.

Hadith courses: Out of the 18 items, eight items representing 44.44% had item difficulty index less than 0.30 and they are items 1, 4, 11, 13, 14, 16, 17 and 18, four items that is 22% had difficulty index between 0.70-0.84 regarded as easy items, while only 6 items representing 33%, had item difficulty index within the appropriate range of 0.30-0.69 that is moderate difficulty.

Objective Two: To compare the overall mean difficult index of Qur'an and Hadith courses of 2019/2020 academic session.

Table 3: Result of the overall mean difficulty index of Qur'an and Hadith courses on 2019/2020 academic session.

Variables	N	Mean Difficulty Index
Qur'an Courses	18	0.56
Hadith Courses	18	0.44

Table 3: shows that Hadith courses are more difficult, but on the overall, both papers (Qur'an and Hadith) were having good or excellent difficulty level for attracting 0.56 and 0.44 as their respective mean difficulty indices.

Objective Three: To compare the items discrimination indices of Qur'an and Hadith courses of the 2019/2020 academic session.

Table 4: Result of the items discrimination indices of Qur'an and Hadith courses on 2019/2020 academic session.

Qur'an Courses				Hadith Courses			
Items	RU	RL	Discrimination Index (DI)	Items	RU	RL	Discrimination Index (DI)
1	69	30	0.25	1	68	43	0.16
2	77	24	0.34	2	67	66	0.01
3	77	44	0.21	3	58	28	0.19
4	50	23	0.18	4	65	44	0.14
5	68	43	0.16	5	72	48	0.16
6	67	66	0.01	6	65	46	0.12
7	58	28	0.19	7	68	43	0.16
8	65	44	0.14	8	67	66	0.01
9	72	48	0.16	9	58	28	0.19
10	65	46	0.12	10	71	38	0.21
11	64	18	0.30	11	67	18	0.19
12	67	33	0.22	12	71	38	0.21
13	64	18	0.30	13	67	18	0.19
14	64	21	0.28	14	65	46	0.12
15	70	27	0.28	15	67	66	0.01

16	75	41	0.22	16	65	46	0.12
17	71	38	0.21	17	67	66	0.01
18	67	18	0.19	18	65	44	0.14

Table 4 showed that from the 2019/2020 Qur'an Examination, two (2) items equivalent to 11.11% of the items had "reasonably good" DI (0.30 – 0.39), they are items 2 and 11. Eight (8) items (44.44%), that is, items 1, 3, 12, 13, 14, 15, 16 and 17 had "marginal" DI (0.20-0.29). However, eight (8) items (44.44%) that is items 4, 5, 6, 7, 8, 9, 10 and 18 were found to have "poor" DI (< 0.20).

Accordingly, from same table 4, regarding Hadith courses, two items corresponding to 11.11% of the items had "reasonably good" DI (0.30 – 0.39), i. e. items 10 and 12. Sixteen items (88.89%) had "marginal" DI (0.20-0.29) and they are items 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 13, 14, 15, 16, 17, and 18. Lastly, no item was found to be "very good" DI (0.40 and above) or "poor" DI (0.19 or less).

Objective Four: To compare the overall mean discrimination index of Qur'an and Hadith courses of 2019/2020 academic session.

Table 5: Result of the overall mean discrimination index of Qur'an and Hadith courses on 2019/2020 academic session.

Variables	N	Mean Discrimination Index
Qur'an Courses	18	0.56
Hadith Courses	18	0.44

Table 5 above revealed the mean discrimination index for the Qur'an examinations was found to be 0.21 which is desired while that of Hadith examinations was found to be 0.13 regarded as poor. Impliedly, the hadith examinations did not discriminate well between the high and low achievers.

Conclusion

Based on the findings, it is concluded that, Qur'anic courses examinations have more moderate difficulty items and more discriminatory between the low and high achievers than the Hadith courses examinations and proved to be more appropriate to students.

Recommendations

Two sets of recommendations were offered in this study, recommendations from the study and recommendations for further studies.

Recommendations from the Study

The following recommendations were made based on the findings of the study:

I. Test developers of such examination should bear in mind that whenever one or more items in a test are too difficult or too easy, then the validity of the examination(s) is at stake.

Thus, analysis of students' response to test items should be done as pilot testing before the final version of the examinations. By so doing will help to improve on item selection by eliminating unreliable items, substituting for poor items, or by recasting poorly stated questions for better effect.

II. The detection of items with poor discrimination ability should be an important factor to be considered in any examination. This is because, if an assessment tool failed to discriminate between the upper and the lower group students, then obviously performance of the students will be influenced by some irrelevant factors.

III. The study is expected to raise the awareness of staff, departments, schools and managements to the importance of including item analysis results in the routine exams' evaluation in the academic boards. To achieve this, schools through their educational development units should provide training workshops to the staff members on how to interpret effectively reports of item analysis.

Recommendations for Further Studies

I. Since the present study investigated Qur'an and Hadith courses, another study on other courses should be carried out for possible detection of item or test problems.

II. It is recommended that further studies in other courses should be carried out in our institutions to provide more empirical evidence.

References

Anastasi, A. &Urbina, S. (2009). *Psychological Testing.*7th Edition. New Delhi: PHI Learning Private Ltd

Emaikwu, S. O. (2011). Issues in Test Item Bias in Public Examinations in Nigeria and Implication for Testing.*International Journal of Academic Research in Progressive Education and Development vol (1), 175-187*

Morphy, K. R and Davidshofer, C. O. (1988).*Psychological Testing*: Principles and Applications. New Jersey: Englewood Cliffs.

Obimba, F. U (1989).*Fundamentals of Measurement and Evaluation in Education and Psychology*.Owerri: Tatan Publishers Limited.

Oguneye, W. (2002).*Continuous assessment: Practice and prospects*. Lagos: Providence Publishers.

Sidhu, K, S.(2005). New Approaches to Measurement and Evaluation New Delhi: Sterling Publishers Ltd.