Journal of Science Education
Volume: II, Issue II, 2020, 73-90

Validation and Analysis of an Achievement Test in the Subject of Biology at the Secondary Level

Anjum Asmat[*]
Farah Latif Naz[**]

Abstract

A reliable and valid test is essential to measure students learning outcomes. The present study was designed to construct a valid Biology test and to analyze the achievement of 9th-grade science students. The sample of the present study was two hundred and nine (209) students who were selected from nine Government schools and private schools by using a simple random sampling technique. For the validation process, two parallel forms of achievement tests were constructed from the subject of Biology. Each form contains thirty-five MCQs. Items were selected from the textbook of Biology of grade 9th and administered to two hundred and nine students (Male and Female) in different private and government secondary schools for boys and girls in Multan city. The validity and reliability of tests were also ensured. Scoring of items was done in a dichotomous manner i.e either correct or incorrect. "Z" test was applied to see the difference between the mean performance of private schools and Government schools, and it was found statistically significant. In the case of male and female performance, the "z"-test was found to be insignificant. Content validity was achieved following a table of specifications. The correlation coefficient between the two forms was found to be 0.78. Kuder Richardson-21 was also used to compute the reliability of the test. Item analysis was done on four criteria i.e. facility index (FI) discrimination index (D), phi coefficient, and point biserial correlation (r_{pbis}). Based on all four criteria, fourteen items were rejected from parallel form no.1 and fifteen items were rejected on parallel form test no. two. It is recommended to use more than one criteria to develop and a valid achievement test so that a good pool of items may be generated.

Keywords: achievement test, biology, reliability, traditional analysis, validity

---

[*] Corresponding Author: PhD Scholar, Bahauddin Zakariya University, Multan. Pakistan.
Email: 1anjum.asmat@gmail.com

[**] Lecturer, Department of Education, Bahauddin Zakariya University, Multan, Pakistan.
Email: Farhlatif@bzu.edu.pk

Introduction

The subject of biology has been given great importance in all fields of life. It is the study of interrelationships between living organisms and their environment. It has a great contribution to the technological growth of the nation and is considered a prerequisite subject in various fields (Ahmed, 2008). Such fields include biotechnology, pharmacy, medicines, forestry, agriculture, nursing, nanotechnology, and many other domains (Ahmed & Abimbola, 2011). Due to the significant role of biology, much emphasis has been made on biology instruction, especially at the secondary school level. To check the student's learning, achievement tests are being developed for science subjects as well as for other subjects so that student’s competence in a specific subject can be measured.

According to Best and Khan (2006), achievement tests are used to measure the learning of an individual according to their capability to perform the test. In the words of Wiersma and Jurs (2006), achievement tests are being been administered to find the understanding of students in a specific area of knowledge and skill. In other studies, Gabriel and Olubunmi (2009) explain that, For obtaining information, tests are regarded as the most popular tool in the teaching system. Besides this, at a specific grade level, these tests are being administered for placing, advancing, and retaining the students.

For this reason, content validity is essential for all academic achievement tests content validity is important. Validity means that “the individual’s scores from an instrument are meaningful, and enable you, as the researcher, to draw good conclusions from the sample you are studying to the population” (Cresswell, 2005). Content validity “pertains to the degree to which a certain measure duly reflects the particular topics or subjects emphasized in the classroom curriculum” (Alexander, 2006, p. 308).

In this regard, a table of specifications and expert judgment has been used to see the validity of the test. Gronlund and Linn (2000), state that a table of the specification has a content of a course and instructional objectives. Another attribute of an achievement test is reliability. Thorndike and Christ (2009) said that In Parallel forms reliability, we should give two tests to each student that are separated by time interval.

With regards to constructing the test, the most popular forms of objective type tests were MC questions. Popham (2018) provided guidelines for the construction of selected-response items. For this purpose, a systematic selection of test items (MCQs) must be ensured for the construction of a reliable and valid test. This emphasizes the significance of item analysis for question banks. Item analysis is largely used for constructing reliable test items of multiple-choice responses. In the Item analysis process, we use to collect, summarize, and get the information from the responses of students to check the quality of test items (Ananthakrishnan, 2002). Further, another widely used procedure to determine the reliability of test items is the traditional item analysis. The traditional item analysis essentially determines the test homogeneity.

In addition to this, In the Item analysis, the index of difficulty is the first item attribute to be determined. Gotteman & Schwarz (2011) state, "High p-values mean the item is easy and low p values mean the item is difficult" (p. 24). Furthermore, item discrimination is an important index. The discrimination index of an item is defined as "the is the degree to which it discriminates between students of high and low achievement" (Linn & Gronlund 2000, p, 321).

Statement of the Problem

A reliable and valid test is significant in measuring students learning outcomes however; most of the available tests measure the instructional objectives. Therefore, this study was designed to construct a Biology achievement test of which the validity and the reliability can be ensured and which can be used to determine the achievement scores of students in terms of learning outcomes of students in Biology at the secondary level.

Objectives of the Study

The objective of this study was to provide valid and reliable test items of Biology at the secondary level through traditional item analysis. Particularly, the study has the following sub-objectives:

1. To check the overall performance of students on test
2. To know the significance of the difference between means of sample scores.
3. To analyze the characteristics of the test
4. To evaluate test items through the traditional method

Hypothesis

For the present study, null hypothesis was developed to see the significant difference

H_{01}. There is no significant difference between the performance of male students and female students on an Achievement test of Biology.

H_{02}. There is no significant difference between the Performance of Private schools and Govt. Schools.

Literature Review

Science is very important to solve the problems of our daily life and overcome the new challenges of modern society (Tytler, 2014). According to Omosewo (2009), all learning institutions must think seriously and should take part in the development of the nation through science education. Many of the developing countries can achieve progress in science and technology because of science education.

Moreover, Science education mainly consists of three subjects like chemistry, physics, and biology and there has been less enrolment in these subjects because of the less knowledge of teachers (Okebukola, 2004). The student's performance in these subjects has not been found encouraging and this alarming situation needs to be investigated (Aina, 2013). For instance, Information and communication technology will also not be possible without science education.

In this regard, Biology is an important subject and preference must be given to this subject because it helps individual to understand the environment concerning himself (Ali, Toriman & Gasim, 2014). Moreover, Nwosu (2006) observed that science concepts and principles teach the students and develop the ability to solve daily life problems in a better way and hence provide a platform for solutions in every walk of life. Previous and recent research showed the indispensability of biology; therefore, much priority should be given to the subject of biology especially at the primary and secondary level of school.

For this purpose, achievement testing in science and other subjects has prime importance to measure few aspects of the intellectual competence of individuals. An achievement test is an instrument provided after a teaching-learning session**.** Tatum (2010) described that achievement tests are constructed to know the degree of knowledge and efficiency of students in a specific field or area. In the words of Best and khan (2016), tests are especially helpful to determine individual and group status in academic learning (Best & Khan, 2016).

In another study, Chatterji (2003) states that achievement tests are essential to evaluate the process of the learner and these are being used to measure the status of the individuals in a particular area of knowledge and skill. An achievement test is a tool that has been done by someone successfully by using his efforts and skill. (Shaheen, et al 2015).

In achievement testing, validity and reliability are essential aspects for the validation of the test. Validity is the degree to which a test measures what it is supposed to measure. According to Shuttleworth (2008), "Validity refers to the strength of the final results and whether they can be regarded as accurately describing the real world" (p.2). Content validity is the assessment of items that are adequate in terms of quantity and quality to measure the characteristics that are supposed to be measured (Buyukozturk, 2011). Another type of validity is face validity which describes the extent to which examinees believe that the instrument is measuring what it is supposed to measure. Ary et al; (2010) stated that it was the degree to which a test seems relevant and important (p.228).

In reliability, we see the stability, consistency of items, and whether they free of error. It provides always provides the same results as students when it is re-administered. This is measured by Coefficient Alpha or KR#20 and KR#21. According to Fraenkel & Wallen (2009), "reliability and Validity always depend on the text in which an instrument is used". Based on context, an instrument may or may not provide consistent scores. Literature provides three methods to find the reliability coefficient; the parallel-forms method; test-retest method, and the internal consistency methods. Gay, Mills, & Airasian (2012) assert that test-retest reliability is the degree to which scores on the same test are consistent over time. In the Parallel Forms Method, they measure the same topics or objectives although, only the specific items are not the same. It was observed that a strong sign of reliability is that it has a high correlation coefficient between two tests. (Fraenkel& Wallen, 2009).

Methodology

In this study, a quantitative approach was used.

Population & Sampling

All the secondary school students of the Government and private educational institutions of Multan city constitute the population of this study. Participants were sampled through simple random sampling. Data collection was done from two hundred and nine respondents out of five hundred students of age ranged from 14-16 from nine selected schools.

The participant consisted of ninety-six boys (45%) and one hundred and thirteen girls (54%).

Tool Development

In this study, two parallel kinds of tests were developed to find the achievement of students in biology. The tests were developed from the textbook of the biology of 9th grade. There were thirty-five MCQs in each kind of test. There were four options in each MCQ. In the initial steps, a table of the specification was constructed. The validity of tests was ensured through subject specialists.

Results

The two tests were administered to two hundred and nine students. Scores were presented in Form one and Form two. To pass the test, 20 marks were decided as passing criteria. On the whole, 164 students could score 20 marks. In terms of percentage 78%. Students obtained passing marks in parallel form #1 and 62% in parallel form#2 of the test.

Table 1
Mean score and standard deviation parallel form of tests
N=209

Test	Mean	St.Dev.
Parallel Test#1	24.2	5.58
Parallel Test#2	21.2	5.89

Table 1 shows that the values of mean and SD of scores of students on two parallel forms of tests. In parallel form test #1, the mean score was M=24.2, S.D=21.2, whereas, in parallel form test #2, the mean score was M=21.2, S.D=5.89.

Table 2
Gender and school-wise Z-Test of parallel form test # 1

Group	N	X	SD	z-value	Sig.
Male	96	24.97	5.20	1.57	Insignificant
Female	113	23.71	5.85		
Govt	151	23.19	5.51	5.26	Significant
Private	58	27.17	4.72		

Table 2 indicates the difference of scores in sub-groups after the administration of parallel form test#1. When a statistical test was applied to the scores of the male and female group, a nonsignificant difference was found. Similarly, when a statistical test was applied to the scores of govt. and private school students, a significant difference was found at 0.05 level of significance.

Table 3
Gender and school-wise Z-Test of parallel form #2

Group	N	X	SD	Z	Significance
Male	96	21.40	5.07	0.38	Insignificant
Female	113	21.1	6.42		
Govt	151	19.3	4.89	8.71	Significant
Private	58	26.17	5.21		

Table 2 indicates the difference of scores in sub-groups after the administration of parallel form test#2. When a statistical test was applied to the scores of the male and female group, a nonsignificant difference was found. Similarly, when a statistical test was applied to the scores of govt. and private school students, a significant difference was found at 0.05 level of significance.

Analysis of Test Characteristics

To find the quality of the test, content validity and reliability are initially used. To measure the test objectives, content validity is used. A table of the specification or a test blueprint, that provides a guideline to teachers in the alignment of objectives, instruction, and assessment (Notar, Zuelke, Wilson, & Yunker, 2004). In this study, Kunderson Richardson#21 was used to administer the test-reliability coefficient. The reliability coefficient of parallel test#1 was found as 0.78.and it was 0.77 of parallel test#2. The correlation coefficient between scores of parallel forms #1 and 2 was 0.79. It shows a high positive correlation and a hence higher level of parallel form reliability.

Item Analysis

The following four criteria were used to analyze items:

a. facility index (FI),
b. Discrimination index (D),
c. phi- coefficient (ɸ), and
d. point bi-serial.

Initially, tests were checked, and scores were granted. Then tests were arranged in descending order for high achievers and low achievers consisting of 27 % of the sampled students.

Facility Index (FI)

The facility index was found by summing up correct responses of the high achievers and low achievers.
Then formula of facility index was applied to find the value of "F".

It was decided to accept those items which fall in range (30% - 70%)

Discriminating Index (D)

To differentiate high achievers from low achievers, a discriminating index (D) was computed. All the scored answer scripts of both parallel forms of tests were arranged in descending order. marks. Applying this method, there were 56 students identified as high achievers and 56 students as low achievers. Such items that were less than 0.20 value of "D" were rejected.

Coefficient (Ø)

To find out the item discrimination, Phi-coefficient was used. The following formula was used to find the coefficient.
Phi-coefficient $\emptyset = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$

Point Biserial correlation

For item analysis, the point bi-serial method was also used. The point-biserial correlation is the correlation between the right/wrong,0 or 1. A point-biserial value of at least 0.15 is proposed though it is reported that "good" items have point-biserial above 0.25. To ensure the quality of items, point-biserial correlation is recommended.

In Point biserial correlation was used to calculate using the mean of each item and mean of the whole sample, standard deviation, the proportion of correct, and proportion of incorrect. All these values were put in the above-mentioned formula. In this way, r_{pbis} of each item was calculated. According to set criteria of r_{pbis}, such items whose value of r_{pbis}, was less than 0.25 were rejected

Table 4

Traditional item analysis of parallel form test#1

Item no.	Group	Correct response	Incorrect response	Total correct responses (HA+LA)	FI (%)	D	ϕ	r_{pbis}
1	High	54	3	111	98.6*	-0.05*	-0.16*	-0.11*
	Low	57	0					
2	High	52	5	87	79.9*	0.29	0.19*	0.28
	Low	35	22					
3	High	57	0	113	99*	0.017*	0.09*	0.04*
	Low	56	1					
4	High	55	2	84	73.7*	0.46	0.51	0.45
	Low	29	28					
5	High	48	9	77	65.55	0.33	0.35	0.29
	Low	29	28					
6	High	54	3	84	72.3*	0.42	0.47	0.39
	Low	30	27					
7	High	50	7	76	74.6*	0.42	0.44	0.40
	Low	26	31					
8	High	49	8	81	70.8	0.29	0.32	0.27
	Low	32	25					
9	High	46	11	76	63.2	0.28	0.35	0.20
	Low	30	27					
10	High	54	3	91	86.1*	0.29	0.37	0.33
	Low	37	20					
11	High	46	11	62	58.3	0.52	0.52	0.45
	Low	16	41					
12	High	57	0	113	99.5*	0.01*	0.09*	0.22*
	Low	56	4					
13	High	55	2	96	86.1*	0.24	0.33	0.29
	Low	41	16					
14	High	53	4	89	79.4*	0.30	0.36	0.32
	Low	36	21					
15	High	42	15	66	50.2	0.31	0.31*	0.23*
	Low	24	33					
16	High	47	10	59	59.3	0.61	0.61	0.51
	Low	12	45					
17	High	54	3	78	64.1	0.52	0.56	0.40
	Low	24	33					
18	High	55	2	84	77.9*	0.45	0.51	0.45
	Low	29	28					
19	High	56	1	85	72.5*	0.47	0.54	0.42
	Low	29	28					
20	High	57	0	89	85.1*	0.43	0.52	0.58
	Low	32	25					

21	High	37	20	45	43.5	0.52	0.52	0.43
	Low	8	49					
22	High	49	8	65	59.3	0.58	0.58	0.43
	Low	16	41					
23	High	57	0	83	75.1*	0.54	0.61	0.53
	Low	26	31					
24	High	55	2	81	76.5*	0.51	0.56	0.46
	Low	26	31					
25	High	47	10	58	49.8	0.63	0.63	0.49
	Low	11	46					
26	High	51	6	70	69.3	0.56	0.57	0.49
	Low	19	38					
27	High	14	14	63	52.2	0.40	0.40	0.32
	Low	37	37					
28	High	53	4	83	78.9*	0.40	0.44	0.42
	Low	30	27					
29	High	57	0	80	79.9*	0.59	0.65	0.61
	Low	23	34					
30	High	26	31	37	34.5	0.26	0.28	0.25*
	Low	11	40					
31	High	43	14	78	71.3*	0.14*	0.15	0.08*
	Low	35	22					
32	High	51	6	75	71.3*	0.47	0.54	0.44
	Low	24	38					
33	High	43	14	61	57.4	0.43	0.43	0.33
	Low	18	39					
34	High	49	8	60	46.9	0.66	0.66	0.48
	Low	11	46					
35	High	32	25	46	47.4	0.32	0.32	0.29
	Low	14	43					

Table 4 indicates that nineteen items were very easy and considered poor based on facility index and hence they were rejected. Four items were found poor based on the discrimination index and decided to be eliminated completely. Five items were found poor based on phi-coefficient and decided to be eliminated as these five Items were found poor based on r_{pbis}. On the whole, based upon all four criteria, 21 items were rejected and item no.5, 8, 15, 16, 17, 21, 22, 25, 26, 27, 33, 34, and 35 were retained in parallel form test#1.

Table 5
Traditional item analysis of parallel form test #2

Item No.	Group	Correct response	Incorrect response	Total correct responses (HA+LA)	FI (%)	D	ϕ	r_{pbis}
1	High	49	8	78	69.86	0.35	0.37	0.23*
	Low	29	28					
2	High	55	2	88	83.7*	0.39	0.45	0.44
	Low	33	24					
3	High	55	2	86	77.5*	0.42	0.57	0.40
	Low	31	26					
4	High	42	15	63	59.8	0.37	0.37	0.30
	Low	21	36					
5	High	45	12	61	47.4	0.59	0.51	0.40
	Low	16	41					
6	High	49	8	70	55.02	0.49	0.50	0.39
	Low	21	36					
7	High	51	6	68	62.7	0.60	0.60	0.46
	Low	17	40					
8	High	53	4	82	65.1	0.42	0.46	0.38
	Low	29	28					
9	High	52	5	63	57.4	0.72	0.72	o.51
	Low	11	46					
10	High	53	4	80	76.6*	0.46	0.49	0.42
	Low	27	30					
11	High	31	26	56	50.2	0.11*	0.11*	0.09*
	Low	25	32					
12	High	38	19	50	48.8	0.46	0.45	0.36
	Low	12	45					
13	High	42	15	62	55.9	0.39	0.39	0.34
	Low	20	37					
14	High	43	14	64	60.2	0.39	0.39*	0.34
	Low	21	36					
15	High	57	0	85	76.1*	0.50	0.58	0.46
	Low	28	29					
16	High	49	8	71	62.2	0.47	0.48	0.37
	Low	22	35					
17	High	48	9	64	55.5	0.56	0.56	0.46
	Low	16	41					
18	High	45	12	70	58.3	0.35	0.36	0.26
	Low	25	32					
19	High	46	11	75	60.2	0.30	0.31	0.28
	Low	29	28					
20	High	38	19	68	61.7	0.14*	0.14*	0.13*
	Low	30	27					
21	High	51	6	82*	77.5	0.35	0.39	0.30
	Low	31	26					

22	High	42	15	72	58.3	0.22	0.22	0.22*
	Low	30	27					
23	High	46	11	66	53.1	0.46	0.46	0.37
	Low	20	37					
24	High	52	5	83*	76.1	0.37	0.41	0.31
	Low	31	26					
25	High	43	14	73	65.1	0.23	0.24	0.19*
	Low	30	27					
26	High	53	4	83*	77	0.40	0.45	0.36
	Low	30	27					
27	High	47	10	60	50.2	0.60	0.59	0.47
	Low	13	44					
28	High	33	24	46	33.5	0.35	0.35	0.26
	Low	13	44					
29	High	54	3	89*	79.4	0.33	0.40	0.31
	Low	35	22					
30	High	32	25	47	40.2	0.30	0.30	0.29
	Low	15	42					
31	High	40	17	49	37.3	0.54	0.54	0.42
	Low	9	48					
32	High	50	7	72	63.2	0.50	0.50	0.44
	Low	22	35					
33	High	43	14	62	48.8	0.42	0.42	0.31
	Low	19	38					
34	High	48	9	64	54.5	0.56	0.56	0.41
	Low	16	41					
35	High	52	5	80	67.9	0.42	0.46	0.38

Table 5 indicates that eight Items were considered poor based on the facility index and hence were rejected. Two Items were found poor based on ($D<0.19$) and decided to be eliminated. Two Items were found poor on the basis($\phi<0.19$) and were decided to be eliminated. Five items were considered poor based on $r_{pbis}\leq 0.25$ and decided to be eliminated. Based on all four criteria, twenty items were retained in parallel form test #2. These items include item no. (4, 5, 6, 7, 8, 12, 13, 14, 16, 17, 18, 19, 22, 27, 28, 30, 31, 32, 33 and 34).

Discussion

In achievement tests, classroom instruction must synchronize with the test items. For validity, teachers need to know the statistical procedures so that the test construction process can be improved (Mozaffer & Farhan 2012). Most of the teachers use traditional item analysis to check the quality of multiple-choice items in this regard, the difficulty, discrimination indices, phi co-efficient, and point biserial are important tools to know the worth of the items. In the present study, the findings of Form one revealed that nineteen items were rejected based on the facility index. Tekin (1996) suggested that the items of which the discrimination index is 0.19 and below should not be used. Therefore, four items were excluded whose discrimination index was less than 0.19. Five items were rejected based on phi-coefficient. According to Varma (2006), FI(%) (item difficulty) is not the indication of item quality, for this purpose, point biserial correlation (item discrimination value) must be used to find the quality of items. Similarly, results of previous studies showed that a good item has a point biserial correlation above 0.25. In this study, five items were rejected based on r_{pbis} and were identified as poor items and need to be decided to be reviewed for further administration.

In the same way, findings of Form two on the set criteria, under FI (%), eight (23%) items were rejected and 77% were accepted. This study is also matched with the study of Suruchi & Rana (2014), in that the majority (75%) and (78%) of the items were acceptable as far as the difficulty was concerned. In the case of the "D" value, two items were rejected and ninety-eight were acceptable. This study is also matched with Pande et al. (2013), where researchers evaluated the quality of MCQs in Physiology and found 75% of the items within the acceptable range and having excellent discrimination. And lastly in the case of phi-coefficient two items were rejected. This finding is also in line with the findings of several studies for example Hingorjo & Jaleel (2012). Who found that most of the items (i.e. more than 50%) were within an acceptable level of item difficulty and discrimination. The Correlation Coefficient was found to be 0.79 between scores on Form "1" and "2". It showed a high and positive correlation between parallel forms. The reliability coefficient of parallel form#1 was found 0.78 and the reliability coefficient of parallel form#2 was 0.77. So according to Fraenkel and Wallen (2009), it is confirmed that a test with a reliability coefficient of 0.70 and above, is normally considered satisfactory in terms of reliability. "Z" test was applied to find the difference between subgroups i.e Gender-wise and school-wise. In achievement tests, different methods have been used to validate the test. For this purpose, content validity is more important in academic achievement tests. In this study content validity was checked by experienced teachers and by a Table of specifications, which was based on detailed learning objectives and content, Biology achievement test validity was found to be satisfactory.

Conclusion

In this study, most of the students qualified for parallel test# 1 while on parallel test # 2, fewer students qualified for the test. In both parallel forms of tests (1 and 2), the difference between mean scores of male and female students was statistically non-significant whereas, the difference between mean scores of students of private and government school students was found significant. High reliability was found between both forms of tests. In traditional analysis of tests, the norm of acceptance for values of FI (%) was between 30% and 70%, and the value of “D” was more than 0.30. For r_{pbis} it was more than 0.25. In parallel form test#1, based on this criterion, nineteen items were rejected based on the facility index. Based on “D” values, four items were rejected. Based on the “ϕ coefficient”, five items were rejected. Based on “r_{pbis}” coefficient, five items were rejected.

Similarly, for parallel form test#2, b ased on set criteria, eight items below “Facility index” were rejected. Based on “D” values, two items were rejected. Based on the “ ϕ coefficient”, two items were rejected. Based on “r_{pbis}” coefficient, five items were rejected.

References

Ahmed, M. A. (2008). *Influence of personality factors on biology lecturers' assessment of difficulty levels of genetics concepts in Nigerian colleges of education* (Unpublished doctoral thesis). University of Ilorin, Ilorin, Nigeria

Ahmed, M. A., & Abimbola, I. O. (2011). Influence of teaching experience and school location on biology teachers' rating of the difficult levels of nutrition concepts in Ilorin, Nigeria. *Journal of Science, Technology, Mathematics and Education* (*JOSTMED), 7*(2), 5261.

Alexander, P. A. (2006). *Psychology in Learning and Instruction*. New Jersey: Pearson.

Ali, A. R., Toriman, M. E., &Gasim, M. B. (2014). Academic achievement in biology with suggested solutions in selected secondary schools in Kano State, Nigeria. *International Journal of Education and Research, 2*(11), 215-224.

Ananthakrishnan, N. (2002) Item analysis-validation and banking of MCQs. in: Ananthakrishnan, N., & Sethuraman, K.R., Kumar S, [eds]. *Medical Education principles and practice*. 2nd ed., Pondicherry. pp. 119-37.

Ary D. Cheser,L., Asghar, R., Sorensen, C.K. (2010). *Introduction to research in education (8thed*). Belmont, CA: Wadsworth.

Best, J. W., & Kahn, J. V. (2006). *Education research*, 10th Ed. New Delhi: PHI Learning Private Ltd, 10-12.

Buyukozturk, S. (2011). Sosyalbilimleriçinverianalizi el kitabı (13rd ed.). Ankara: Pegema Yayıncılık.

Chatterji, M. (2003). *Designing and using tools for educational assessmen*t, Boston: Allyn and Bacon.

Cresswell, J.W. (2005). *Educational research: planning, conducting and evaluating quantitative and qualitative research* (2nd ed). Pearson Merrill Prentice Hall.

Fraenkel, J. R. & Wallen, N. E. (2009). *How to design evaluate education research*. New York: McGraw Hill Companies (7th ed).

Gabriel, O. A., & Olubunmi, A. (2009). Comprehensive scientific demystification of Kigelia Africana: A review. *African Journal of Pure and Applied Chemistry*, *3*(9), 158-164.

Gay, L. R., Mills, P., & Airasian, P. (2012). Educational research: Competencies for analysis and application. Upper Saddle River, NJ: Pearson Gibson

Gotteman, D.M. & Schwarz, S. W. (2011). Juvenile justice in the US: Facts for policymakers. National Center for Children in Poverty, New York, NY https://doi.org/10.7916/D8Q81N10

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index, and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*(2), 142.

Hornby, J. M., (2004). Enhanced production of farnesol by Candida albicans treated with four azoles. *Antimicrobial Agents And Chemotherapy*, *48*(6), 2305-2307.

Linn, R.L. (2000). Assessment and accountability. *Educational Researcher*, 9 (2), 4-16.

Mozaffer, R. H., & Farhan, J. (2012). Analysis of one-best MCQs: The Difficulty index, discrimination index, and distractor efficiency. *Journal of Pakistan Medical Association, 62*, 142–147. Retrieved from http://jpma.org.pk/PdfDownload/3255.pdf [PubMed], [Web of Science ®],

Nwosu, A. A. (2006). Biology education for the new millennium. In E. A. C. Okeke (Ed) *Educational reform in Nigeria for the new millennium.*

Notar, C.E., Zuelke, D. C., Wilson, J. D. & Yunker, B. D. (2004). The table of specifications: insuring accountability in teacher-made tests. *Journal of Instructional Psychology*, 31, 115-129.

Okebukola, F. (2012). The views of Nigerian teachers in public and private primary schools on the teaching of early literacy in English. *Literacy*, *46*(2), 94-100.

Omosewo, E. O. (2009). Views of physics teachers on the need to train and retrain Physics teachers in Nigeria. *African Research Review*, *3*(1).

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., &Agrekar, S. H. (2013). Correlation between difficulty and discrimination indices of MCQs informative exam in physiology. *South-East Asian Journal of Medical Education*, *7*(1), 45-50.

Pham, H., Besanko, J., & Devitt, P. (2018). Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple-choice question. *MedEd Publish*, *7*.

Shaheen, M. N. U. K., Kayani, M. M., & Shah, N. H. (2015). The teaching of Science at Secondary Level: An Analysis of Teachers' Classroom Practices. *International Journal of Innovation in Teaching and Learning (IJITL)*, *1*(1).

Shuttleworth, M. (2008). Validity and Reliability. Experiment Resources. Retrieved from http://www.experiment-resources.com/validity-and-reliability.html

Suruchi, S., & Rana, S. S. (2014). Test item analysis and the relationship between difficulty level and discrimination index of test items in an achievement test in biology. *PIJR*, *3*(6), 56-8.

Tatum, B. C. (2010). Accelerated education: Learning on the fast track. *Journal of Research in Innovative Teaching*, *3*(1).

Tekin, H. (1996). *Measurement and evaluation in education*. Yargi Publications, Ankara.

Thorndike, R. M., & Thorndike-Christ, T. (2009). *Measurement and evaluation in psychology and education* (Eight ed.). Essex, England: Pearson Education.

Tytler, R. (Eds.). (2014). *The age of STEM: Educational policy and practice across the world in science, technology, engineering, and mathematics*. Routledge.

Verma, K. S. (2006). Impact on real and reactive power pricing in open power market using unified power flow controller. *IEEE Transactions on Power Systems*, *21*(1), 365-371.

Wiersma, W. & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Boston, MA: Allyn and Bacon.

Citation of this Article:
Asmat, A. & Naz, F.L. (2020). Validation and analysis of an achievement test in the subject of biology at the secondary level. *Journal of Science Education*, *2*(2), 73-90.