Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.9, No.34, 2018
www.iiste.org

Analysis of Test Items Used in an Achievement Test in Physics at Secondary Level

Ph.D. Scholar (Education), Shahzad Ahmad
Federal College of Education, H-9, Islamabad.

Research Associate, Sadia Jamil
Federal College of Education, H-9, Islamabad.

Abstract
Aim of present study is the analysis of test items constructed in an achievement test in the subject of physics for class IX. It is used to study the difficulty level, discrimination power and distracter analysis of each test item. The achievement test was comprised of 40 multiple choice questions. Test was administered by researcher on a sample of 250 students from Islamabad Model Schools in Sector G-9. Findings of the study showed that most of the test items fall in acceptable range of difficulty index and discrimination power. In keeping view, the findings of the study, researcher comes to know that most of the test items prove to be satisfactory regarding difficulty index and discriminating power. However, ten out of forty test items were discarded due to high or very low level of difficulty and poor discrimination power. This work is equally significant for the researchers and for the subject teachers in preparing achievement tests in order to assess performance of students with optimum level of difficulty index and discriminating power.
Keywords: Item Difficulty, Discrimination Index, Distracter Analysis and Reliability.

1.1 Introduction

Item analysis is a statistical technique which is used to ascertain the usefulness of test items in preparing achievement tests. In order to develop quality assessment test and precisely useful multiple-choice items, item analysis plays an important role both in contributing to the objectivity of the test and to highlight the areas where students are conceptually weak.

According to Brown and Frederick (1971), Item analysis has two main purposes: Firstly, to identify imperfect test items and secondly, to highlight particular content areas which learners have or have not fully mastered. The main function of item analysis procedure is to measures usefulness of each test items in term of its difficulty level and the capacity to differentiate between high and low achievers of particular test. In nutshell, item analysis helps in selecting and keeping the best test items for final draft of the test and discarding poor test items. There is also scope for reviewing and modifying the ill constructed test items.

In general, once test items have been prepared, the worth of such items can be assessed using number of procedures representative of item analysis a) the difficulty level of the test items b) the capacity of the individual test item to differentiate, and c) the distracter analysis. Difficulty concerns with how many number of persons recommending the answer correctly. Discrimination can be observed by making comparison of persons getting a particular items correct with the total test score. Finally, the distracter analysis concerns with the quality of distracters.

1.2 Purpose of the study

The purpose of the study is to determine the items difficulty level, discrimination power and distracters analysis of the test items used in an achievement test in physics at secondary level.

Objectives of the study

The objectives of the study are as follow:

1. To find out the difficulty level of individual test item.
2. To find out the discrimination power of individual test item.
3. To study the distracters analysis of test items.
4. To find out the test reliability of the test items.

1.3 Significance of the Study

Successes of any education system rely solely on the well-established system of examination. Examination on one hand play the role for promoting students to next grade and on the other hand provides scope for assessment of curriculum, program and performance of teachers. The significance of the study may be understands as

- It enables subject teacher to know about the too easy and too difficult test items (Items difficulty).
- It enables him to differentiate between high achievers and low achievers (Discrimination power).
- It enables him to construct useful distracters (Distracters analysis).

- It also provides scope for checking whether the test under consideration is reliable or unreliable (Test reliability).

1.4 Instrument of the Study

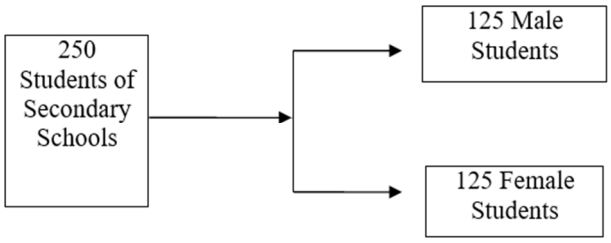
The instrument of the study was comprised of 40 test items for data collection. Initially test was consisting of 60 test items from unit on motion and rest in physics for class IX by Punjab text book board, Lahore. The test is developed by researcher himself and approved by experts in the subject of physics at secondary level. All experts ensured that various levels of Bloom's Taxonomy have got due weightage. Test items which found ambiguous, providing clues and of language issues were discarded. All necessary instructions were provided to the participants to ensure reliability.

1.5 Population of the study

The population of the study comprised of all students of class 9^{th} in Islamabad Model Schools, G-9 sector, Islamabad.

1.6 Sample of the Study

Simple random sampling technique was used to obtain a sample of 250 students at secondary level. Sample comprised of the students of both genders.



1.7 Data Collection

Achievement test in physics was administered by the researcher himself to the sample of 250 students. Initially students were informed about the particular unit for achievement test and concerning directions for its administration.

1.8 Data Analysis

Student's scores on test items were arranged by using MS-Excel in descending order. In order to start the item analysis top 30% scorers and bottom 30% scorers of the entire sample were selected. Upper and lower group of students each comprised of 75 students was used. Nunnally (1972) suggested 25% while SPSS (1999) uses the highest and lowest one third (33%).

1.8.1 Item Difficulty (F)

It deals with how difficult is the test item. It is indicated by the percentage of the pupils who got the item right. It is recommended that test item would be neither too easy nor too difficult.

It involves following basic steps:

1) Arrange the papers in order from the highest to the lowest scores (say 250 papers).
2) Select 30% papers with the highest scores (high achievers) - 75 papers.
3) Select 30% papers with the lowest scores (low achievers) - 75 papers.
4) 40% papers in the middle (100 papers) would not be taken in account.
5) Calculate the correct responses of high achievers and low achievers on each item.
6) Apply the formula and calculate F (Facility Index)

F= NR/NT

Where

NR= Number of students who got the item right.

NT=Total number of students.

7) F is acceptable when it ranges from 0.30 to 0.90.
8) Value more than 0.90 indicates that item is very easy.
9) Value less than 0.30indicates that item is very difficult.

Example

Table:1 High Achievers=75, Total papers=250, Low Achievers=75

	N	A	b	C	d	Omit
High Achievers	75	13	53	03	06	0
Low Achievers	75	12	43	07	13	0

"b" is the correct answer (b=96).

F= NR/NT

=96/150

=0.64

Table:2 Evaluation of Item Difficulty for Item Analysis

Item Difficulty Index	Item Evaluation
Above 0.90	Very easy item
0.62	Ideal item
Below 0.20	Very difficult item

Source: Instructional Assessment Resources (IAR 2011)

1.8.2 Discrimination Power (D)

It refers to the degree to which test item discriminates between pupils with high and low achievement. One purpose of testing is to discriminate between high and low achievers.

It involves following basic steps:

1) Arrange the papers in order from the highest to the lowest scores (say 250 papers).
2) Select 30% papers with the highest scores (high achievers) - 75 papers.
3) Select 30% papers with the lowest scores (low achievers) - 75 papers.
4) 40% papers in the middle (100 papers) would not be taken in account.
5) Calculate the correct responses of high achievers and low achievers on each item.
6) Apply the formula and calculate D

D=NH-NL/n

Where

n= Number of high or low achievers.

NH=Number of high achievers who got the item right.

NL= Number of low achievers who got the item right.

7) D is acceptable when it ranges from 0.30 to 1.0.
8) Value more than 0.40 indicates 100% discrimination.
9) Value less than 0.30 indicates incapability of the item to discriminate.

Example

Table: 3 High Achievers=75, Total papers = 250, Low Achievers=75

	N	a	b	c	d	Omit
High Achievers	75	09	56	09	01	0
Low Achievers	75	24	22	25	04	0

"b" is the correct answer (b=78).

D=NH-NL/n

=56-22/75

=0.45

Table:4 Evaluation of Discrimination Indexes for Item Analysis

Index of Discrimination	Item Evaluation
0.40 and above	Very good item; accept
0.30 to 0.39	Reasonably good but subject to improvement
0.20 to 0.29	Marginal item usually need and subject to improvement
Below 0.19	Poor item to be rejected or improved by revision

Source: Ebel (1972) in Ovwigho (2013)

1.8.3 Distracters Analysis

The last but not least stage in item analysis procedure is to assess the quality of distracters for each test items. These are incorrect alternatives alongside correct option in case of multiple choice items. The quality of good distracters is to "distract" examinees which are unaware of the right response.

Distracter analysis provides the information to instructor to observe that how many number of testers in the upper group and bottom group tick out each option on a multiple-choice item. Distracter fails to do its job when it is not selected by examinees at all. In other words the usefulness of distracter relies solely on its selection by some examinees. Otherwise it is ineffective and must be revised or replaced. When distracter attracts more examinees from the bottom group than the upper group it is negative discrimination. Similarly when distracter

attracts more examinees from upper group than the bottom group it is positive discrimination. Distracters should demonstrate negative discrimination.

Example

Table: 5 High Achievers=75, Total papers = 250, Low Achievers=75

	N	A	B	C	d	Omit
High Achievers	75	09	56	09	01	0
Low Achievers	75	24	22	25	04	0

1.9 Findings and Discussion

1.9.1 Item Difficulty

Table: 6 Includes the item difficulty indices of each test item.

Item No.	Item Difficulty	Item No.	Item Difficulty
1	**0.71**	**21**	**0.36**
2	**0.74**	**22**	**0.28**
3	**0.64**	**23**	**0.74**
4	**0.69**	**24**	**0.67**
5	**0.24**	**25**	**0.52**
6	**0.56**	**26**	**0.42**
7	**0.63**	**27**	**0.58**
8	**0.66**	**28**	**0.45**
9	**0.12**	**29**	**0.92**
10	**0.66**	**30**	**0.67**
11	**0.76**	**31**	**0.78**
12	**0.30**	**32**	**0.80**
13	**0.80**	**33**	**0.90**
14	**0.40**	**34**	**0.70**
15	**0.31**	**35**	**0.58**
16	**0.77**	**36**	**0.70**
17	**0.49**	**37**	**0.32**
18	**0.22**	**38**	**0.36**
19	**0.66**	**39**	**0.58**
20	**0.16**	**40**	**0.60**

In accordance with Instructional Assessment Resources (IAR), numbers of test items were classified in to following three categories of item difficulty as indicated in table.

Table:7 Distribution of items in term of level of difficulty

Item Difficulty Index (P)	Number of Test Items
Easy (Above 0.90)	1
Moderate (0.20-0.90)	37
Difficult (Below 0.20)	2

1.9.2 Item Discrimination

Table: 8 Includes the coefficient of item Discrimination of each test item.

Item No.	Item Discrimination	Item No.	Item Discrimination
1	0.36	21	0.34
2	0.22	22	0.09
3	0.13	23	0.42
4	0.42	24	0.30
5	0.04	25	0.45
6	0.45	26	0.26
7	0.38	27	0.24
8	0.36	28	0.29
9	0.08	29	0.14
10	0.45	30	0.36
11	0.61	31	0.33
12	0.16	32	0.64
13	0.22	33	0.18
14	0.05	34	0.38
15	0.12	35	0.34
16	0.21	36	0.46
17	0.37	37	0.28
18	0.24	38	0.29
19	0.38	39	0.33
20	0.01	40	0.48

Table below includes the classification of test item into different levels of discrimination regarding discrimination coefficient by Ebel (1972) in Ovwigho (2013).

Table: 9 Distribution of items in term of level of discrimination

Discrimination Index (P)	Number of Test Items
Very Good (Above 0.40)	8
Reasonably Good (0.30-0.39)	12
Marginal (0.20-0.29)	10
Poor (Below 0.20)	10

1.9.3 Distracters Analysis

Table: 10 Includes the responses of subjects on each test item.

Item No.	Correct Responses	Item Responses							
		a		b		c		d	
		Upper	lower	upper	lower	upper	lower	upper	lower
1	B	00	00	67	40	00	02	08	33
2	C	01	01	01	05	64	47	09	13
3	B	13	12	53	43	03	07	06	13
4	C	06	12	00	15	68	36	01	12
5	D	11	15	18	21	26	22	20	17
6	C	05	14	20	22	49	35	01	04
7	C	02	13	07	16	61	34	05	12
8	C	03	10	00	10	64	35	08	20
9	B	36	53	12	06	21	10	06	06
10	C	08	39	01	01	65	35	01	00
11	C	05	05	02	14	67	47	01	09
12	A	29	17	23	32	17	19	06	06
13	A	69	52	01	04	05	12	00	07
14	B	20	17	32	28	13	17	10	13
15	A	28	19	14	12	16	15	17	29
16	B	06	16	66	50	00	03	03	06
17	B	22	42	51	23	01	06	01	04
18	D	18	24	11	15	20	28	26	08
19	C	02	18	02	13	67	33	04	11
20	B	34	08	13	12	08	09	20	46

Item No.	Correct Responses	Item Responses							
		a		b		c		d	
		Upper	lower	Upper	Lower	upper	lower	upper	lower
21	C	17	24	15	32	40	14	03	05
22	D	17	13	08	12	25	32	25	18
23	C	01	04	03	21	72	40	00	10
24	C	07	17	01	02	62	39	05	17
25	B	09	24	56	22	09	25	01	04
26	A	42	22	21	30	07	15	05	08
27	B	12	10	53	35	03	13	07	17
28	A	45	23	16	37	09	10	05	05
29	D	00	03	00	04	00	04	75	64
30	B	05	11	64	37	02	14	04	13
31	B	04	21	71	46	00	02	00	06
32	C	02	22	01	04	72	48	00	01
33	D	00	02	00	05	00	07	75	61
34	B	07	22	67	38	01	06	00	09
35	C	11	18	03	12	59	28	02	17
36	C	01	09	04	17	69	36	01	13
37	C	20	22	11	20	35	14	09	19
38	C	05	09	13	17	38	16	19	33
39	D	09	13	08	13	02	18	56	31
40	C	05	12	04	23	63	27	03	13

In above tables there are 160 response options comprising of 40 plausible and 120 implausible one. The main purpose of distracter analysis is to distinguish between plausible and implausible options. There are 8 test items found to be confusing as maximum subjects select implausible options rather than true one. All most all distracters are mostly selected by subjects of lower group i.e. negative discrimination.

1.10 Test Reliability

An instrument is considered reliable if the instrument produce same result every time when use to evaluate identical measurement. A Kuder Richardson formula, KR20 and KR21 for analyzing test items, which is based on item difficulty, was used to analyze internal consistency of the test. The value of KR20 and KR21 range between 0 to 1.The closer the value to 1 the better the internal consistency.

$$\rho_{KR20} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{\sigma^2}\right)$$

Where
K= Number of test items
$\mathbf{P_j}$=Number of subject in the sample who answered item j correctly
$\mathbf{q_j}$= Number of subject in the sample who didn't answered item j correctly
$\boldsymbol{\sigma}^2$=Variance of the total scores of all the subject taking the test

$$\rho_{KR21} = \frac{k}{k-1}\left[1 - \frac{\mu(k-\mu)}{k\sigma^2}\right]$$

Where
K= Number of test items
μ= Mean score
$\boldsymbol{\sigma}^2$=Variance of the total scores of all the subject taking the test
The value KR_{20} =0.7966 and KR_{21} = 0.7515 shows that the test has high reliability.

1.11 Conclusions

Results of this study signify the importance of item analysis for constructing a useful, reliable and valid test. Its main focus was to manage item bank for individual discipline during an academic calendar so that a collection of items within acceptable range of difficulty index, discrimination power and suitable distracters obtained. In the present study an achievement test on a unit from subject of physics comprised of 40 test items was prepared and administered to check the level of difficulty, discrimination and standard of distracters. It clearly highlighted the poor, confusing items which need improvement.

References

Boyle, J. D., & Radocy, R. E. (1987). *Measurement and Evaluation of Musical Experiences*. MacMillan Publishing Company.

Institute for Interactive Media and Learning (2013). *Multiple Choice Questions*. Retrieved on 3 February, 2017 from www.iml.uts.edu.au/assessment/types/mcq.

Ebel, R.L. (1972) *Essentials of Educational Measurement* (1st Edition). New Jersey: Prentice Hall.

Wiersma, W. Jurs. SG (1990). *Educational Measurement and Testing (2nd Ed.). Boston: Allyn and Bacon. emailed to Fast track students and RN-BSN students*.

Nunnally, JC. 1972. *Educational Measurement and Evaluation*. 2nd edition. New York. McGraw-Hill.

Ebel, Robert L. *Measuring Educational Achievement.* Englewood Cliffs, New Jersey, Prentice Hall, 1965.

Brown, Frederick G., "*Measuring Classroom Achievement*", Halt Richard and Winston, U.S.A.", 1981, pp. 101-110, 224p.

Instructional Assessment Resources. (2011*). Item Analysis*. Retrieved November 9, 2013 from University of Texas at Austin, Instructional Assessment Resources, IAR Web site: http://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php

Ovwigho, B. O. (2014). Empirical Demonstration of Techniques for Computing the Discrimination Power of a Dichotomous Item Response Test. *Journal of Educational and Social Research*, *4*(1), 189.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160.