

Selection Practice of Best Test Item for Achievement Test

Dr. Indra Kumari Bajracharya*

Background:

Achievement is the stage of attainment by the students, generally expressed in terms of grade or scores. It is defined as performance of students in education tests based on scores. So, achievement is the learning outcome of a student. According to Good (1973) refers to academic achievement as knowledge attained or skill developed in the school subjects, usually designed by test scores or marks assigned by the teacher. Freeman (1965) defines a test of educational achievement as a test designed to measure knowledge, understanding, skill in a specific subject or group of subject. Trow (1956) defined academic achievement as attained ability or degree of competence in school tasks usually measured standardized test and expressed in grades or units based on norms, derived from a wide sampling of pupil performance. After a test has been administered and scored, a post hoc analysis is often performed in order to evaluate the test's effectiveness. This procedure often involves an analysis of the individual items on the test.

Item Analysis

Item analysis is a post administration examination of a test (Remmers, Gage and Rummel, 1967: 267). The quality of a test depends upon the individual items of a test (Freeman, 1962:113, Shrama, 2000: 197). A test is usually desirable to evaluate effectiveness of items. It provides information concerning how well each item in the test functions. An item analysis tells about the quality of an item. One primary goal of item analysis is to help improve the test by revising or discarding ineffective items. Another important function is to ascertain what test takers do and do not know. Item Analysis describes the statistical analyses, which allow measurement of the

* Associate Professor of Mahendra Ratna Campus, Tahachal, Kathmandu

effectiveness of individual test items. An understanding of the factors which govern effectiveness (and a means of measuring them) can enable us to create more effective test questions and also regulate and standardize existing tests. Item analysis helps to find out how difficult the test item (Gronlund, 1993: 103). Similarly it also helps to know how well the item discriminates between high and low scorers in the test. Item analysis further helps to detect specific technical flaws and thus provide further information for improving test items (Gronlund, 1993: 102). Similarly, it helps in selecting the best items for the final test, reject poor items and modify some of the items (Sharma, 2000:107). The item analysis procedure provides following information on each item

- The difficulty level of each item (how hard it is).
- The discriminating power of each item (whether or not good students get the item right more often than poor students)

The difficulty level of the items

The percentage of students who answer test items correctly in a test is called difficulty level of an item (Gronlund, 1993: 103). In other words, the difficulty level of item is defined as the proportion or percentage of the examinees or individuals who answered the items correctly (Singh, 1986: and Remmers, Gage and Rummel, 1967: 268). In a test, calculation of proportion or percentage of individuals choosing the right answer of the test item is called item difficulty (Wood, 1988: 377). According to J.P. Guilford " the difficulty value of an item is defined as the proportion or percentage of the examinees who have answered the item correctly" (Freeman, 1962:112-113 and Sharma, 2000:200). In this method, index of difficulty level of each item is determined on the basis of responses of all the examinees. This formula would be more accurate to determine the difficulty level of items from the entire sample (Garrett, 1981: 362, Patel, 1993: 162 and Gronlund, 1993:104). According to Frank S. Freeman "the difficulty value of an item may be defined as the proportion of certain sample of subjects who actually know the answer of an

item" (cited in Sharma, 200:200). The difficulty of an item can be determined in several ways (Garret, 1981: 362-363):

- By the judgment of competent people who rank the items in order of difficulty.
- By how quickly the items can be solved.
- By the number of examinees in the group who get the item right

The third procedure is the standard method for determining difficulty of each item in the objective test. The level of difficulty is represented by a numerical term, which may range from zero to 100 percent (Remmers, Gage and Rummel, 1967: 268). An item of test is not answered correctly by any of the examinees, it means the item is most difficult, the difficulty value is zero percent or proportion is also zero. If an item of test is answered correctly by every examinee, it means the item is very easy the difficulty value is 100 percent or proportion is one. An item answered correctly by 70 percent of the students is said to have a difficulty index of 70. A general rule of measurement of any item whose difficulty index is lower than 10 or higher than 90 is worthless measurement (Remmers, Gage and Rummel, 1967: 268). Since, difficulty refers to the percentage answering the item correctly, the smaller the percentage figures the more difficulty the item (Gronlund, 1993: 103). Thus if item is correctly answered by 90 percent of examinees, they are regarded easy whereas those items are difficult if they are correctly answered only by 5 percent of examinees. So, an items answered correctly by 100% or 0% examinees have no differentiating significance.

The formula for computing item difficulty (P-value) given by Gronlund, (1993: 103) and Garrett (1981:363) is present following formula :

$$P = \frac{R}{N} \times 100$$

Where

P = the percentage of examinees who answered items correctly.

R = the number of examinees who answered items correctly.

N = total number of examinees who tried the items.

The discriminating power of the items

The discriminating power of a test item refers to the degree to which success or failure of an item indicates possession of the ability being measured. In other words, the ability of the test items measures the better and poorer examinees of items (Remmers, Gage and Rummel, 1967: 268). According to Marshall Hales (1972) the discriminating power of the item may be defined as the extent to which success or failure on that item indicates the possession of the achievement being measured. In the same context, Blood and Budd (1972) defined the index of discrimination as the ability of an item on the basis of which the discrimination is made between superiors and inferiors. Similarly, the degree to which single items separates the superiors from the inferiors' individuals in the trait or group of traits being measured (Bean, 1953, cited in Sharma, 2000: 201). A discrimination index is meant to communicate the power of an item in separating the more capable items from less capable on some latent attributes (Wood, 1988: 377). The discriminating power is defined in the numerical term, which may range from +1 to -1 (Remmers, Gage and Rummel, 1967: 268). On the basis of discriminating power, items are classified into three types (Sharma, 2000: 201).

- **Positive Discrimination:** If an item is answered correctly by superiors (upper groups) and but not answered correctly by inferiors (lower group) such item possess positive discrimination.
- **Negative Discrimination:** An item answered correctly by inferiors (lower group) but not answered correctly by the superiors (upper groups) such item possess negative discrimination.

- **Zero Discrimination:** If an item is answered correctly by the same number of superiors as well as inferiors examinees of the same group. The item cannot discriminate between superior and inferior examinees. Thus, the discrimination power of the item is zero.

The formula for computing item discrimination given below Gronlund, (1993: 103) and (Ebel and Frisbie, 1991:231)

$$D = \frac{R_U - R_L}{N_U \text{ or } N_L}$$

Where D = Index of discrimination.

R_U = Number of examinees giving correct answers in the upper group.

R_L = Number of examinees giving correct answers in the lower group.

N_U or N_L = Number of examinees in the upper or lower group respectively.

Item Analysis procedure

The steps, which were used for item analysis, in the present study, propounded by Gronlund, (1993: 103) and Ebel and Frisbie(1991:225) are present below-

- All test papers were first orderly arranged from highest to the lowest score.
- Sum the numbers of examinees who give the correct responses and these numbers was divided the total number of examinees and multiply by 100 for each item. The result as percentage is the index of difficulty or P-value.
- 27% papers were selected with the highest scores and called this the upper group. Similarly, 27% papers were selected with the lowest scores, which were called the lower group.
- For each item, number of examinees who gave correct response in the upper groups was counted. The same was done in lower group.

- Subtract the lower group counts from the upper group count for the correct responses. Divide this difference by the number of examinees in one of the group (either upper or lower group both are in same size). The expressed result as a decimal is the index of discrimination or D- value.

Conclusions: There is no any uniform view on selecting items for the achievement test on the basis of P-value or indices of difficulty. A common practice is to select some items whose difficulty is at 50 percent level or close to, and other items with a wide range of degree of difficulty in terms of percent passing (Freeman, 1962:113). According to Ebel and Frisbe (1991:) a good test must have some easy items to test the low achievers and some difficult items to test the high achievers. According to Garret (1981: 164) the normal curve can be taken as a guide in the selection of difficulty indices. Thus, 50% items might have difficulty indices between .25 and .75. Similarly, 25% indices are larger than .75 and 25% are smaller than .25. The same criteria were adopted while selecting the items for the test on the basis of P-value by Patel (1993: 163).

Generally, items are considered better if discriminating power of items is higher. So, most of the items need to have discrimination index above .20 (Patel, 1993: 164). Remmers, Gage and Rummel, (1967) also has similar view. According to them, items with discrimination index above .20 are regarded as having sufficient discriminating power to be used in most tests in academic achievement. Garret (1981:363) also has similar view. But Ebel and Frisbe (1991:232-233) opine that achievement test items should have indices of discrimination of 0.30 or more. He has also classified items accordingly-0.40 and up is very good, 0.30 to 0.39 is reasonably good but possibly subject to improve, 0.20 to 0.29 is marginal items usually subjected to improvement and Below 0.19 is Poor items that are to be rejected or further improved by revision

References

1. Bean, K.L. (1953). *Construction of Educational and Personnel Tests*, New York: McGraw-Hill Book Co.F

2. Blood, D.F. and W.C. Budd, (1972). *Educational measurement and evaluation*. New York: Harper and Row.
3. Ebel, R.L. & Frisbie, D.A.(1991). *Essentials of Educational Measurement* (5th ed.). New Delhi: Prentice Hall of India Pvt. Ltd
4. Freeman Frank S. (1962). *Theory and Practice of Psychological Testing*, New Delhi: Oxbord & Ibh publishing.
5. Garrett Henry E. (1981). *Statistics in Psychology and Education*, India, Vakils, Feffer and Simons
6. Grounlund Norman E. (1993). *How to make Achievement Tests and Assessments*, Illinois:
7. John W. Best et.al. (2000). *Research in Education* (7th edition), New Delhi: Prentice Hall of India.
8. Marshall and Hales (1972). *Essentials of testing*, London: Addison-Wesley Publishing Company Ltd
9. Patel R. A.(1993). *Educational Evaluation Theory and Practice*, Delhi: Himalaya Publishing House.
10. Rammers H.H., Gage, N.L. & Rummel, J.I. (1967). *A Practical introduction to measurement and evaluation* (2nd ed.). Delhi: Universal Book Stall.
