Journal of Elementary Education
Vol.24, No. 1 pp. 61-70

Development of an Achievement Test in the Subject of Health and Physical Education at Intermediate Level in Pakistan

Zahida Aziz Sial* & **Khalid Mahmood Khan****

Abstract

The present study was aimed at Development of an Achievement test in the subject of Health and Physical Education at Intermediate level. For this purpose, an achievement test comprising 60 items (test was divided into two parallel forms containing 30 items each) was constructed from the text book of Health and Physical Education for class 11. Seven institutions (Govt. Colleges and Govt. Higher Secondary Schools) were taken by using random sampling technique and sample of 190 students was taken by using the same sampling technique for study from the population. The test was administered to 190 students (M/F) in different Govt. Colleges and Govt. Higher Secondary Schools of Multan District. The answer sheets were scored and results were tabulated. Two items were rejected on the basis of facility index (F%). Four (04) items were improved on the basis of facility index (F%). Eleven (11) items were rejected on the basis of discrimination index (D). Eight (08) were rejected on the basis of phi-coefficient (ϕ). After the refinement of items through both methods, the selected items became the basis for the standardization of an achievement test in the subject of Health and Physical Education at Intermediate level.

Keywords: achievement test; rasch model; Facility index (F); discriminatory index; (D) phi-coefficient (ϕ)

* Assistant Professor, Department of education, B.Z.U Multan
** Subject Specialist Govt. College for Elementary Teachers Rangeel Pur Multan

Introduction

The importance of tests, especially achievement tests, at any level of education cannot be ignored as they are the systematic procedures for measuring sample and students' knowledge. They provide a proper feedback to teaching at all stages of learning. They become the major source for improving standards of education at any stage. According to Gay (1996), "tests provide the teachers and other officials with important information regarding the individual's and group's proficiency, provided that they are properly constructed. They also measure the current status of learners in the given area of knowledge or skills". It has been defined by Sax (1997) in the following way:

> "A test is a task or series of tasks used to obtain systematic observation presumed to be representative of educational or psychological traits or attributes."

Keeping in view the past experiences and future needs in the field of test construction, the patterns of questions in our tests have changed. Now -a- days more weight age is given to objective type items in all papers of almost every level. This is really a good step towards the evaluation of students' achievement. In objective type items the most commonly used type is the multiple choice items/ questions.

To know the standards of test items, it has become important to analyze the test items and item analysis has become the backbone of test construction. Test construction may only be fruitful through the well thought, careful and sophisticated process of item analysis. Item analysis is the process to determine validity of individual items. In words of Anastasi & Urbina (2007), "the reliability as well as validity of any test depends on the characteristics of its item. They can be built into a test in advance through item analysis. Tests can be improved through the selection, substitution or revision of item". Therefore, it is a set of procedures that provides us with the estimate of validity of each item.

Ebel & Frisbi (1991) have stated the importance of item analysis in the following way:

> Item analysis indicates which items are difficult, easy, and moderately difficult or easy, so it provides an index of difficulty value of each item….

It means item analysis also indicates why a particular item has not functioned effectively and how this might be modified. So, different experts have adopted different procedures for item analysis. Many techniques of item analysis have been used and developed. Item analysis for norm-referenced tests can be conducted in two ways. According to Ebel & Frisbi (1991), in traditional item analysis, the following aspects are considered:

i. For each item value of facility index (F) is calculated. Item having values of (F) between .20 and .80 is retained and other items are rejected.

ii. For each item, the discriminating power (D) is calculated statistically. Items having the value of (D) more than .20 is retained and other items are rejected. D, may be positive and negative or 0 value. It can range from – 1 to + 1. Items having value of (D) at .40 or greater are very good items. Items having value of (D) at .30 to 39 are reasonable. Items having value of (D) at .20 to .29 are needed some revision. Items having value of D below .19 are poor items.

The effectiveness of distracters is also observed keeping in view the established norms for the acceptance and rejection of the distracters.

The other method of item analysis was launched by a Danish Mathematician, George Rasch. It can be applied to measure the ability and attitudes in various disciplines such as Health, Education, Psychology, Social Sciences and Economics. (Penta et al, 2005). It is based on two expectations regarding the outcomes of a person attempting an item in a test. Firstly, that a person with higher attainment will have a greater probability of success on any item from that topic than a person with lower attainment. Secondly, that any person should always be more likely to answer correctly the easier item on that topic than a hard one. The method expresses the probability of success on item mathematically in terms of two parameters, one related to person's attainment and the other related to item's difficulty

Trimzi (1984) also describes its process and importance in these words:

> Rasch calibration sets out to place the measurements of person attainments and item difficulty on same scale and uses the same units for both. The procedure uses the variable to locate the position of person attainment measures, which correspond for the possible raw scores on the test.

It means that the Rasch Analysis of items not only analyses the items but also categorizes the students' calibration.

Rasch method of item analysis is one of the novel methods in the field of education. It has minimized many defects of traditional item analysis. Application of this model provides diagnostic information regarding how well the criterion is met. It also provides information about how well items are questioned on assessment work to measure the ability or trait (http://en.wikipedia.org/wiki/Rasch model).

Physical education is the only source, which provides the remedies for these diseases (blood pressure, tension and diabetes) (Asif, 2006). Keeping in view the importance of Health and Physical Education as a subject, the researcher decided to develop the achievement test at Higher Secondary School level employing the Rasch Model of Item Analysis. The present study may be pioneer in the field of Health and Physical Education in Pakistan.

Statement of the Problem

Development and Rasch Analysis of an Achievement Test in Health and Physical Education at Intermediate level.

Objectives of the Study

The objectives of the study were:

i. To construct an achievement test in the subject of Health and Physical Education for 11^{th} class.
ii. To determine the difficulty level of each item.
iii. To evaluate the quality of items with the help of Rasch model.
iv. To compare the results of both traditional and new technique of item analysis (Rasch Model)
v. To assess the reliability of the test.

Research Methodology

The present study was quantitative in approach. The following procedure was adopted as a research methodology.

Population

All students from Govt. colleges and Govt. higher secondary schools of District Multan were taken as the population for the present study.

Sampling

Seven institutions (Govt. Colleges and Govt. Higher Secondary Schools) were taken by using random sampling technique and sample of 190 students were taken by using the same technique of sampling for study from the population. The detail of the sample was as

Number of Boys	=	99
Number of Girls	=	91
Total	=	190

Statistical Analysis and Interpretation

The data were analyzed by using different statistical methods such as mean, median, S.D, Skewness, Quartiles (Q1, Q3) Range of the score.

Test A and test B were analyzed through two methods i.e. Traditional Method and Rasch Model of Analysis. One aim of the present study was to discriminate between difficult and easy items. For this purpose traditional method of item analysis was adopted. In order to get accurate and reliable results for each item, value of facility index (F), discriminating power (D) and Phi (ϕ)-co-efficient were calculated. Item analysis was conducted by using Rasch Model. For this purpose, Prox item calibration and Prox person measurement tables were prepared. Both person attainment and item difficulty were shown on the same line for comparison. Item person performance (IPP) was recorded and shown through a master table. Item characteristic curves (ICC) and person characteristic curves (PCC) were drawn. Mean performance of the students was calculated. Reliability of the test was also calculated.

Results of the study

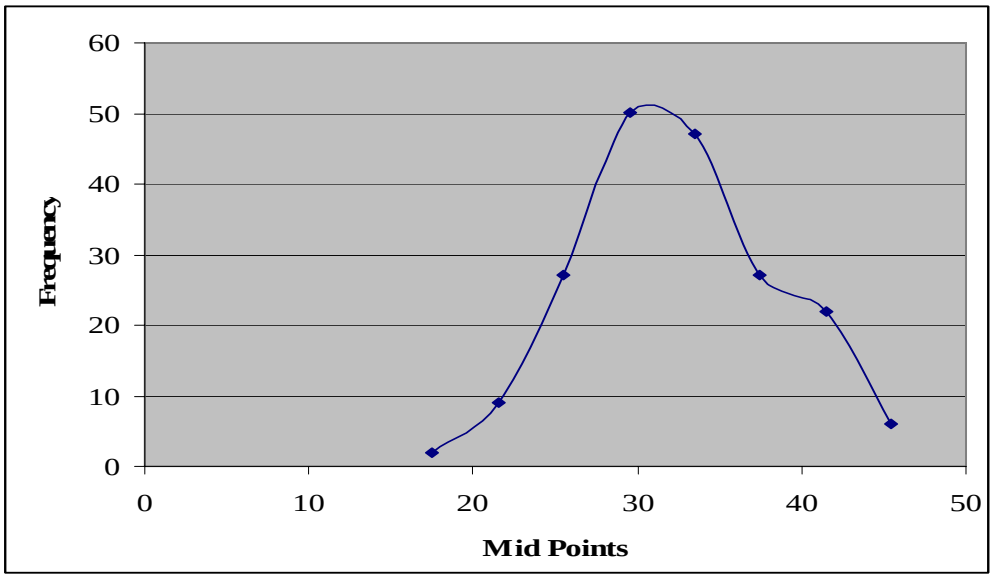
The following results were drawn from the study

Table 1: Performance of the students in Total Test

Class Interval	Frequency (f)	Mid Points
16-19	2	17.5
20-23	9	21.5
24-27	27	25.5
28-31	50	29.5
32-35	47	33.5
36-39	27	37.5
40-43	22	41.5
44-47	6	45.5
Total	190	

Graph 1 shows clear picture of the students in total test that two students fall in class interval 16-19, Nine students fall in class interval 20-23, Twenty-seven students fall in class interval 24-27, Fifty students fall in class interval 28-31, Forty-seven students fall in class interval 32-35, Twenty-seven students fall in class interval 36-39, Twenty-two students fall in class interval 40-43 and Six students fall in class interval 44-47.it means scores were spread symmetrically and this is proof of test reliability value which was 0.74 and 0.79 respectively.

Graph 1



Quantitative Analysis of the Total Test

Table 2 shows quantitative analysis of the sample in total. Since median was apparently greater than mean hence there seemed a slight negative skewness (i.e. S_k = -0.06) in the sample distribution. 25% of the total sample fell under the score 28.26 (i-e Q_1 = 28.26) and 75% under the score 36.14 (i.e. Q_3 = 36.14). The respondents below Q_1 represented low achievers and above Q_3 high achievers. Total score ranged from 17 to 47.

Table 2: Quantitative Analysis of the Total Test

N	Mean	Median	S.D	Skewness	Q_1	Q_3	Range = Max-Min
190	32.44	32.60	7.21	-0.06	28.26	36.14	47-17 = 30

Table 3 Gender Wise Quantitative Analysis of the Total Test

Table 3 reflects the quantitative analysis of the sample on total test on the basis of gender. Mean of female respondents was apparently less than median and mean of male was greater than Median which indicated a little negative skewness (i.e. S_k = -0.03) in the female and positive skewness (i-e S_k = +0.12) in the male sample distribution. Out of the total sample 25% of female and male fell under the score 29.03 and 27.07 respectively (i-e Q_1 = 29.03 for female and Q_1 = 27.07 for male) and 75% under the score 40.31 and 33.08 respectively (i-e Q_3 = 40.31 for female and Q_3 = 33.08 for male). Score of the female respondents ranged from 22 to 47 and of male 17 to 40.

Table 3: Gender Wise Quantitative Analysis of the Total Test

Sr. No.	Category	N	Mean	Median	S.D	Skewness	Q_1	Q_3	Range = Max-Min
1	Female	91	34.80	34.87	6.68	-0.03	29.03	40.31	47-22=25
2	Male	99	30.55	30.30	6.32	+0.12	27.70	33.08	40-17 = 23

Table 3 reflects the quantitative analysis of the sample on total test on the basis of gender. Mean of female respondents was apparently less than median and mean of male was greater than Median which indicated a little negative skewness (i.e S_k = -0.03) in the female and positive skewness (i-e S_k = +0.12) in the male sample distribution. Out of the total sample 25% of female and male fell under the score 29.03 and 27.07 respectively (i-e Q_1 = 29.03 for female and Q_1 = 27.07 for male) and 75% under the score 40.31 and 33.08 respectively (i-e Q_3 = 40.31 for female and Q_3 = 33.08 for male). Score of the female respondents ranged from 22 to 47 and of male 17 to 40.

Analysis of items through Traditional Method

Following calculations were done for analyzing test items.

i. Facility index (F)
ii. Item discrimination (D)
iii. Phi-Coefficient (ϕ)

Facility index (F)

Value of (F) was calculated by following these steps:

1. 27% of the papers with the highest-score were selected as a high scoring group.
2. 27% of the papers with the lowest-score were selected as a low scoring group
3. Then, the value of "F" was deduced with the help of the following formula.

$$F = \frac{N_h + N_l}{2n} \times 100$$

Item Discrimination (D)

Item discrimination refers to the ability of an item to differentiate among students. Item discrimination was calculated by using the following formula:

$$D = \frac{N_h - N_l}{n}$$

Phi-Coefficient (ϕ)

A more refined and accurate way of calculating discrimination index is through phi-coefficient (ϕ). In this method, we include correct as well as the incorrect responses of the high achievers and low achievers. Item discrimination calculated with the help of this formula, gives a more vivid picture of two groups.

$$\text{Phi-coefficient } (\phi) = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Where

a is the number of correct responses in the top 27%.

b is the number of incorrect responses in the top 27%.

c is the number of correct responses in the bottom 27%.

d is the number of incorrect responses in the bottom 27%.

	Correct	Incorrect
Top 27%	**a**	**b**
Bottom 27%	**c**	**d**

Results and discussion

On the basis of findings, following major results were drawn.

Two (2) items were rejected on the basis of facility index (F%)and Four (4) items needed improvement on the basis of facility index (F%).,Fifty four (54) items were very good items on the basis of facility index (F%)Eleven (11) items were rejected on the basis of discrimination index (D)Forty nine (49) items were good on the basis of discrimination index (D) Eight (8) items were rejected on the basis of phi-coefficient (ϕ). Fifty two (52) items were good on the basis of phi-coefficient (ϕ)Nine (09) distracters were rejected as they not more effective. Ten (10) distracters were to be rejected as they attracted by high achievers more than low achievers. While comparing gender wise responses of sample group, Scores of the female respondents ranged from 22 to 47 and of male 17 to 40.mean of female and male students was 17.67 and 15.11.difference of means and score range show the better performance of female students.

Test has positive test reliability value i.e. 0.79 by KR # 20 method and 0.73 by KR # 21 method. After the refinement of items through both methods, the selected items became the basis for the standardization of an achievement test in the subject of Health and Physical Education at Intermediate level.

Recommendations

On the basis of the results, following recommendations were made:

1. The number of students can be increased in gender wise and domicile wise.
2. The number of items can be increased
3. Test may be used for standardization.
4. Item analysis techniques and Rasch Model can be included as compulsory concepts into the course out line of teacher training programmes.

5. Traditional item analysis techniques may also be introduced to the students, teachers and examiners through seminars, debates and workshops.
6. Teacher may use "The Rasch Model" for the calibration of their tests in addition to the traditional methods of item analysis and test calibration.
7. Item bank for all subjects may be developed with the coordination of BISE and BZU Multan.

References

Anastasi, A. & Urbina, S. (2007). *Psychological testing*. India: Dorling Kindersley (Pvt.). Ltd.

Asif, M. (2006). *A text book of health & physical Eeducation for degree classes*. Nowshera.: New Classic Publishers.

Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement.* New Delhi: Prentice – Hall.

Gay, L.R. (1996). *Educational research*. New Jersey: Prentice Hall, Inc.

Penta, M. et al. (2005). *Application of the Rasch Model*. Sprimont: Mardaga Retrieved from http://www.raschanalysis.com

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation.* Washington: International Thomson Publishing Company.

Trimazi, S.H. (1984). *Calibration of an achievement test in Mathematics for year 8: An empirical study*. Unpublished Doctoral Thesis. Mount Lawley Perth: Western Australian College of Advanced Education. http://en.wikipedia. org/wiki/rasch model.