Far East Journal of Mathematical Education

Published Online: April 2014
Available online at http://pphmj.com/journals/fjme.htm
Volume 12, Number 1, 2014, Pages 27-46

IMPROVING TESTS BY USING ITEM ANALYSIS WITH SPECIAL REFERENCE TO EIGHTH STANDARD STUDENTS OF THREE SAMPLED SCHOOLS OF GUWAHATI

Jayashree Pathak and Kanika Das

Department of Mathematics and Statistics
K. C. Das Commerce College
Guwahati-21, India
e-mail: jayashreepathak32@gmail.com

Department of Mathematics
Gauhati University
Guwahati-14, India
e-mail: daskanika2@gmail.com

Abstract

When tests are developed for instructional purposes to assess the effects of educational programs, or for educational research purposes, it is very important to conduct item and test analyses. Writing good test items is a creative exercise and its selection requires proper skill and care. Effective item writing as well as the simple statistical techniques of item analysis can improve classroom tests which help in building high reliability and validity into a test in advance. The researchers made a pilot study of 45 items given to 300 eighth standard students before finalizing the items for the final test. The average difficulty factor and the average discriminating index of the

Received: August 5, 2013; Accepted: November 7, 2013
Keywords and phrases: item, difficulty factor, discriminating index, upper group and lower group of students.
Communicated by Paul F. Messina

finalized 30 items are 0.56 and 0.52, variance is 74.92, reliability is 0.93, validity is 0.96 and standard error is 2.19. The results revealed that a test can be improved through the selection, substitution or revision of items. It is hoped that this instrument will help the teachers to determine the weaknesses of their students in learning mathematics.

Introduction

Mathematics Education today has become more important than ever. The goal of mathematics education is to provide all students with the ability to use mathematics to improve their own lives, to help them become aware of their responsibilities as citizens, and to help them prepare for their futures. Students come to the classroom with different learning styles, various levels of mathematics proficiency, language barriers, communication issues, and assorted backgrounds. In addition, student attitudes and personalities affect learning. Many students suffer from math anxiety, some students find maths boring or unnecessary and some students do poorly simply because they have a low self-esteem when it comes to maths. Other issues stem from the way in which students access mathematical content. Some students struggle to visualize or develop understanding for abstract concepts. Others students struggle to master mathematical procedures because they do not understand the concept of the rational for the steps of the procedure. Many students do not possess strategies for an unfamiliar word problem. Whatever the obstacle, it is essential that our educational systems try to meet the mathematical needs of all students before they fail. Mathematics education must begin at a very early age so that students develop the foundational understanding and skills necessary to achieve in mathematics. Adequate preparation in mathematics is essential in this ever-changing global environment if we expect our children to keep up with the world market, lead in technological advances, be prepared for national security, and provide a satisfying livelihood for themselves. Understanding and using data about student performance is very essential. Students' performance and progress can be known after taking a test of the subject. Without analyzing and discussing data, teachers are unlikely to identify the problems that need attention, take appropriate steps to

solve those problems, or know whether they are progressing toward the achievement of their goals. The systematic approach to test development was initiated by Binet and Simon in 1916. Since that time psychometricians have been concerned with the extent to which accurate measurement of a person's ability is possible. Making fair and systematic evaluations of others' performance can be a challenging task. Judgments cannot be made solely on the basis of intuition, haphazard guessing, or custom (Sax [16]). Teachers, employers and others in evaluative positions use a variety of tools to assist them in their evaluations. Tests are tools that are often employed to assist in student evaluations (Matlock-Hetzel [13]). As a basic unit of the test, the quality of each test item that constitutes the test plays an important role in deciding the nature and quality of the test. The nature of the test items should be diagnostic in such a way that the test takers' performance on these items should indicate the extent of understanding, misunderstanding, or lack of understanding of the content of the test depending on the responses of the test takers. The most commonly used tools in test item analysis are item difficulty and item discrimination. The present study investigated the item difficulty and item discrimination. It is customary to arrange items in order of difficulty so that test takers begin with relatively easy items and proceed to items of increasing difficulty. This arrangement gives the test takers confidence in approaching the test and also reduces the likelihood of wasting much time on items beyond their ability to the neglect of easier items they can correctly complete. Item difficulty is the percentage of students taking the test who answered the item correctly. The higher the difficulty index, the easier the item is understood to be (Wood [20]). To compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item. The proportion for the item is usually denoted as p and is called *item difficulty* (Crocker and Algina [4]). In the process of test construction a major reason for measuring item is to choose items of suitable difficulty level. If no one passes an item, then that particular item is not a good item. The same is true of items that every one passes. Such items cannot provide any information about individual difficulties as they do not contribute to the reliability and validity of the tests.

Most measurement experts agree that upon repeated testing an individual's observed score will vary even though his true ability remains constant. Test analysis examines how the test items perform as a set. Item analysis "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 and Levitov [17, p. 163]). It is a procedure to obtain a description of the statistical characteristics of each item in the test. This approach requires identification of single item which provide maximum discrimination between individuals on the latent trait being measured. The quality of the items in a test determines its validity and reliability. Item analysis thus provides an empirical basis for revising the test, indicating which items can be used again and which items have to be deleted or rewritten (Lange et al. [11]). Binet and Simon [2] who were among the first to systematically validate test items noted the proportion of students at particular age levels passing an item. This statistic was measuring the relative difficulty of the items for different age groups. The item difficulty index which is denoted by p is one of the statistics used in item analysis. Item difficulty is related to item variance and hence to the internal consistency of the test. Test constructors are usually concerned with achieving high test reliability, e.g., precision of measurement. Therefore, an item difficulty of .50 is considered to be the ideal value necessary to maximize test reliability. This is because half the examinees are getting the item correct and half the examinees are missing the item. The proportion missing an item is defined as $1 - p$ or q. Thus, when p is equal to .50, q is equal to .50. Hence test constructors have been advised (Gulliksen [9]) to select items with difficulty indices at or near .50. A second important item statistic in classical item analysis is item discrimination index. An item discrimination index provides a measure of how well an item contributes to what the test as a whole measures. If a test is given to a large group of people, then the discriminating power of an item can be measured by comparing the number of people with high test scores who answered that item correctly with the number of people with low scores who answered the same item correctly. In computing the discrimination index, D, first score of each student's test is ranked in order.

Next, the 27% of the students at the top and the 27% at the bottom are separated for the analysis. Wiersma and Jurs [19] stated that “27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis” (p. 145). There need to be as many students as possible in each group to promote stability, at the same time it is desirable to have the two groups be as different as possible to make the discriminations clearer. According to Kelly (as cited in Popham [15]) the use of 27% maximizes these two characteristics. Nunnally [14] suggested using 25%.

One of the main objectives of test analysis is to improve the internal consistency of the test under construction where internal consistency is defined as the extent to which all items are measuring the same ability. To ensure high internal consistency the random error in the test must be minimized. Therefore, internal consistency is directly dependent upon the correlation among the items in the test. When item-test correlations are high, the test is able to discriminate between high and low scorers and hence internal consistency is increased.

Several articles have been published concerning standards for item selection to maximize test validity and increase internal consistency. Flanagan [7] stated two considerations in selecting test items: (a) the item must be valid, that is, it should discriminate between high and low scorers, and (b) the level of item difficulty should be suitable for the examinee group. Gulliksen [9] agreed with Flanagan on these two points and added a third; items selected with $p = .50$ would produce the most valid tests. Several studies have been conducted to examine the effects of varying item difficulty on test development. Brogden [3] has dealt with simplified methods of obtaining indices of item discrimination. Because of the lack of computers in the early years of test development many psychometricians concerned themselves with devising tables to provide quick estimates of item discrimination. Kelley [10] found that in the computation of item discrimination only 54 percent of the examinee group (based on total test score) needed to be used. Considering the top 27 percent and the bottom 27

percent of the test scorers resulted in a considerable savings in computational time. Flanagan [7] developed a table of item discriminations to estimate the correlation between items and test score based on Kelley's extreme score groups of top and bottom 27 percent. In terms of discrimination index, .40 and greater are very good items, .30 to .39 are reasonably good but possibly subject to improvement, .20 to .25 are marginal items and need some revision, below .25 are considered poor items and need major revision or should be eliminated (Ebel and Frisbie [5]).

Fan [6] developed a table for the estimation of the tetrachoric correlation coefficient using the upper and lower 27 percent of the scorers, tetrachoric correlation is similar to the biserial correlation, where the correlation is between two variables, which are assumed to have a normal and continuous underlying distribution, but have been artificially dichotomized. Guilford [8] presented several shortcut tabular and graphic solutions for estimating various types of correlation coefficients to measure test item validity. These methods result in saving a considerable amount of time when one is forced to use hand calculations. Today these short cut methods can be used by classroom teachers who often do not have the aid of calculators or computers. However, many test constructors still use this method of item analysis even though computers are available with which more sophisticated item analytic techniques can be used.

Statement of the study

It is apparent that an empirical investigation seems warranted to determine whether item analysis would help to produce superior test of eighth standard students in terms of internal consistency and efficiency. It was for this reason that the present study was undertaken and titled as:

"*Improving tests by using item analysis with special reference to eighth standard students of three sampled schools of Guwahati*"

Objectives of the study

1. To identify deficiencies in the test.

2. To diagnose questionable items and discuss the problem with the class.
3. To revise or discard items in subsequent testing if the wording of the item is at fault.
4. To review or clarify items which are not properly understood by the students.
5. To determine whether the item as a whole was too easy or too difficult and therefore of little value.
6. To determine whether a good item dealing with a point that happened to have been effectively taught and well remembered should be retained.
7. To assess how well the students of eighth standard have mastered the different topics of mathematics and what are the broad areas of their weaknesses.

Delimitations of the study

As the study was a pilot study, there are delimitations.

1. The study was confined to the pupils of class VIII only.
2. The study was confined to only three English medium schools of Guwahati, Assam.
3. The study was confined to only 300 students.
4. The schools were affiliated private schools under SEBA of Guwahati.
5. The statistics obtained for examinees and items are sample dependent.
6. It is assumed throughout this study that the test under construction is unidimensional (all items are measuring only mathematical ability).
7. No special schools such as Blind school, Deaf and Dumb school etc. were taken.
8. Data analysis was confined to Difficulty Factor, Discrimination Index and Reliability coefficient.

Terms used

Item: The term item is used because all test questions are not actually questions; they are commonly statements that have multiple choice questions which have incorrect and correct results.

Item analysis: It is designed for multiple tests and can be useful in looking at individual items as well as an overall test. It is a very informative process.

Upper group students: The checked test scripts are arranged in descending order where the highest 27% of scripts are placed at the top and the students of this highest order is considered as upper group.

Lower group students: The checked test scripts are arranged in descending order where the lowest 27% of scripts are kept at the bottom and the students of this lowest order is considered as lower group.

Difficulty factor (p): Ratio of number of students who get the item correct to the total number of students attempting the item,

$$p = c/n,$$

where p is difficulty factor, c is number of students who selected correct response and n is number of students who attempted the item.

The higher the difficulty index, the easier the question is, so a value of 1 would mean all the students got the question correct and it may be too easy. P levels should be between .20 and .80 with the average being 0.5 so that reliability does not suffer (items which are too difficult or too easy decrease reliability). Items with difficulty levels between 0-0.2 and 0.8-1.0 are discarded as they are either too difficult or too easy respectively.

Discrimination index (D): The purpose of each test item is to separate those who show a high degree of skill, knowledge etc from those who have low skill, knowledge etc. Ideally each item on the test ought to contribute to this discrimination between those who have mastered or understood the content and those who have not.

In order to calculate the discrimination index, we must first divide the scripts into an upper and lower half,

$$D = (U - L)/S,$$

where U is number of students in upper group who responded correctly, L is number of students in lower group who responded correctly and S is number of students in larger group.

The discrimination index should be positive if an item is to be considered as a reasonable item. The value should be at least .30. Generally a high positive value indicated a good discriminating item. A low or negative value indicated that the item was too easy or more difficult. Such items should be rewritten or eliminated.

Sample of the study

Statistically analyzing multiple choice test items ensure that the items given to the students are effectively evaluating their learning. The purpose of a systematic approach to test design is to reduce error in test taking. For analyzing items of mathematics test paper a sample of 300 students of eighth standard was considered. The schools were divided into three categories A, B and C. Category A included a school whose results of mathematics of eighth standard annual examination was consistently more than 60% over a period of last three years. Category B included the school whose result was between 40% and 60% over the same period of the same examination. Category C included the school whose result in mathematics annual examination was below 40% for the last three years. The three schools considered for the sample were:

Category A	Nichols English Medium High School, Chatribari, Guwahati-8
Category B	Angels of God English Medium High School, Goswami Service, Guwahati
Category C	Y.W.C.A. English Medium High School, Chatribari, Guwahati

Research design and procedure

It was decided to select about 100 students from the school of each category. Mathematics syllabus of eighth standard was examined and analyzed thoroughly. Four areas of mathematics were selected namely, (a) Arithmetic - Fifteen arithmetic items were given to see whether the students are able to understand numbers, develop the meaning of operations and compute fluently. (b) Geometry - Ten geometric items were given to see whether the students are able to analyze characteristics of geometric shapes, make mathematical arguments regarding geometric relationships and use visualization to solve problems. (c) Algebra - Twelve algebraic items were given to see whether the students are able to understand patterns, relationships, functions and use algebraic symbols. (d) Statistics - Eight statistical items were given to see whether the students are able to learn to use appropriate statistical methods to analyze data, make inferences and predictions based on available data, understand and use basic concepts of statistics.

All the items were of multiple choices where four choices were given out of which one was the correct answer. Before administering the test to the students, detailed instructions were thoroughly discussed with the teachers of the concerned schools who have agreed to be present as invigilators at the time of test. Duration of 60 minutes was given which included 15 minutes for distribution of scripts, giving instructions to the students and finally collecting the scripts from the students. A total of 300 students participated in the test. The checked scripts were arranged in descending order where highest 27% of the scripts were placed on the top, next 46% were placed in the middle and the lowest 27% were placed at the bottom. The highest 27% and the lowest 27% contained 81 scripts each.

Next the difficulty factor and the discriminating index was found out. The results of the item analysis for 45 items are given in the table below.

Findings of the study

Item analysis findings (Total students - 162).

Table 1

Item Number	Upper Half (*U*)	Lower Half (*L*)	Difficulty Index (*P*)	Discrimination Index (*D*)
1	65	20	0.52	0.54
2	70	30	0.62	0.50
3	50	15	0.40	0.43
4	40	09	0.30	0.38
5	68	64	0.81	0.05
6	58	33	0.56	0.31
7	15	02	0.10	0.16
8	69	43	0.69	0.31
9	78	72	0.93	0.07
10	57	30	0.54	0.33
11	68	40	0.67	0.35
12	21	03	0.15	0.22
13	67	27	0.58	0.49
14	79	63	0.87	0.20
15	67	21	0.54	0.58
16	62	25	0.54	0.46
17	77	31	0.67	0.58
18	48	20	0.42	0.35
19	76	56	0.81	0.24
20	62	31	0.57	0.38
21	65	38	0.64	0.33
22	79	75	0.95	0.04
23	54	24	0.48	0.37
24	66	15	0.50	0.62
25	12	02	0.09	0.12
26	64	06	0.43	0.71
27	65	15	0.49	0.61
28	77	68	0.87	0.11
29	69	15	0.52	0.66

30	73	28	0.62	0.55
31	76	62	0.85	0.17
32	20	09	0.18	0.14
33	78	21	0.61	0.70
34	75	60	0.83	0.06
35	69	41	0.68	0.34
36	79	28	0.66	0.62
37	17	13	0.19	0.04
38	72	20	0.57	0.64
39	78	11	0.55	0.82
40	79	63	0.88	0.19
41	80	25	0.65	0.67
42	80	04	0.52	0.93
43	70	29	0.61	0.50
44	20	05	015	0.19
45	73	38	0.69	0.43

Analysis of the study

Item 1 - Item 15 consist of Arithmetic items. It is observed that the difficulty index of items 5, 9, 14 is very high which shows that these items are too easy for the students. Similarly for item numbers 7 and 12 the difficulty indices are 0.10 and 0.14 respectively which shows that these items are too difficult for the students. Moreover the discriminating index for these five items is less than .30. Hence these five items are discarded. Ultimately out of 15 Arithmetic items 10 items are considered for the final test. 66.67% of Arithmetic items were taken.

Item 16 - Item 25 consist of Geometry items. It is observed that the difficulty factors of items 19 and 22 are very high whereas item number 25, the difficulty factors is 0.27 which is very low. Again for these three items the discriminating index lie below 0.30 which results in keeping 7 geometrical items out of 10 items. 70% of geometrical items were considered.

For Statistical items numbering from 26-33, item number 28 and 31 have high difficulty factor and item number 32 has low difficulty factor. Also the discriminating index of these items is very low which leads in removing these three items. Consequently for the final test 5 items are considered out of 8 items. 62.5% of statistical items were considered.

For Algebra, twelve items numbering from 34 to 45 is given to the sampled students. It is found that items 34 and 40 have very high difficulty indices whereas item numbers 37 and 44 has very low difficulty index. Again the discriminating index for these four items is below 0.30 which results in rejecting the four items. 66.7% of algebraic items were selected.

From the column of discrimination index, it can be observed that 15 test items have values less than .30 which leads to deleting these items. Before discarding, reviewing or retaining any items, it was discussed with the students the reasons for their poor performance in some of the items. The students expressed that wording of few items was not clear. It was reviewed and those items were retained after alteration in the wordings. Again the students gave their views that some of the items were too difficult for their standard. In such cases these items were deleted for the final test. Ultimately final form consisted of 30 items after applying difficulty factor and discrimination index. The finalized 30 items are arranged serially.

After analyzing the items, it was discussed with the mathematics teachers of the three sampled schools regarding the teaching methodology used by them. It was suggested and advised to make the class more lively and interesting by frequent use of the interactive board and also using the technique of group discussion. Students learn more when working cooperatively rather than working in an isolated, competitive fashion. When students work successfully in a cooperative setting, their learning achievement increases, their self-confidence rises and they often have a better opinion of the subject. Group discussions help the students to discuss mathematics and share problem-solving methods with one another.

Finalized 30 items

Table 2

Item Number	Upper Half (U)	Lower Half (L)	Difficulty Index (P)	Discrimination Index (D)	$q_i = 1 - p_i$	$p_i q_i$
1	65	20	0.52	0.54	0.48	0.25
2	70	30	0.62	0.50	0.38	0.24
3	50	15	0.40	0.43	0.60	0.24
4	40	09	0.30	0.38	0.70	0.21
5	58	33	0.56	0.31	0.44	0.25
6	69	43	0.69	0.31	0.31	0.21
7	57	30	0.54	0.33	0.46	0.25
8	68	40	0.67	0.35	0.33	0.22
9	67	27	0.58	0.49	0.42	0.24
10	67	21	0.54	0.58	0.46	0.25
11	62	25	0.54	0.46	0.46	0.25
12	77	31	0.67	0.58	0.33	0.22
13	48	20	0.42	0.35	0.58	0.24
14	62	31	0.57	0.38	0.43	0.25
15	65	38	0.64	0.33	0.36	0.23
16	54	24	0.48	0.37	0.52	0.25
17	66	15	0.50	0.62	0.5	0.25
18	64	06	0.43	0.71	0.57	0.25
19	65	15	0.49	0.61	0.51	0.25
20	69	15	0.52	0.66	0.48	0.25
21	73	28	0.62	0.55	0.38	0.24
22	78	21	0.61	0.70	0.39	0.24
23	69	41	0.68	0.34	0.32	0.22
24	79	28	0.66	0.62	0.34	0.22
25	72	20	0.57	0.64	0.43	0.25
26	78	11	0.55	0.82	0.45	0.25
27	80	25	0.65	0.67	0.35	0.23

28	80	04	0.52	0.93	0.48	0.25
29	70	29	0.61	0.50	0.39	0.24
30	73	38	0.69	0.43	0.31	0.21
Total			16.84	15.5		7.15

Average difficulty factor of the 30 finalized items $= \sum_{i=0}^{n} p_i / 30$

$= \sum_{i=0}^{30} p_i / 30$

$= 16.84/30$

$= 0.56.$

Average discriminating index of the finalized 30 items $= \sum_{i=0}^{n} D_i / 30$

$= \sum_{i=0}^{30} D_i / 30$

$= 15.5/30$

$= 0.52.$

Average difficulty index of the selected items is 0.56 and the discrimination index is 0.52 which is very satisfactory.

Testing of reliability and validity

The importance of developing unidimensional tests is demonstrated most clearly in considering the concepts of test reliability and validity. For a test to be valid it must actually measure the trait it was intended to measure. For a test to be reliable it must provide similar results upon repeated measurement. It should be easier to estimate these two important aspects of a test when the test is unidimensional than when the test is multidimensional, hence the use of a unidimensional test in the present study.

Method of Rational Equivalence or Kuder-Richardson formula is used to test the reliability of the finalized 30 items. This formula enables us to get an estimate of the coefficient of reliability. It stresses the intercorrelations of the

items in the test and the correlations of the items with the test as a whole. The Kuder-Richardson formula for determining the test reliability in terms of the difficulty and intercorrelations of test items is

$$r_{tt} = \frac{n}{n-1}\left[\left(\sigma_t^2 - \sum_{i=1}^{n} p_i/q_i\right)/\sigma_t^2\right],$$

where r is the reliability coefficient of the whole, n is the number of items in the test, σ is the standard deviation of the test scores, p_i is the proportion of group answering the test correctly, and $q_i = 1 - p_i$.

Now σ_t = standard deviation.

Standard deviation of natural numbers 1 to 30 is given by:

$$\sigma_t = \sqrt{\frac{1}{12}(n^2 - 1)}$$

$$= \sqrt{\frac{1}{12}(900 - 1)}, \; n = 30$$

$$= \sqrt{\frac{1}{12}(899)}$$

$$\sigma_t^2 = \frac{1}{12} \times 899$$

$$\sigma_t^2 = 74.92$$

$$\sum_{i=1}^{n} p_i q_i = \sum_{i=1}^{30} p_i q_i \quad (i \to \text{1-30 for the finalized 30 items})$$

$$= 7.15$$

$$r_{tt} = \frac{n}{n-1}\left[\left(\sigma_t^2 - \sum_{i=1}^{n} p_i q_i\right)/\sigma_t^2\right]$$

$$= \frac{30}{29}\left[\frac{74.92 - 7.15}{74.92}\right]$$

$$= \frac{30}{29}\left[\frac{67.77}{74.92}\right]$$

$= (1.03)(0.92)$

$= 0.93,$ reliability is very high.

Validity of the study

The validity of a test concerns what the test measures and how well it does so. It tells us what can be inferred from test scores. The validity of a test cannot be reported in general terms. No test can be said to have high or low validity in the abstract. Its validity must be established with reference to the particular use for which the test is being considered. There are different types of validity like predictive, concurrent, content and construct.

For the present study instrinctive validity is found by taking the square root of its reliability. Validity with reference to test reliability is found to be $\sqrt{0.93}$ or 0.96, which is high.

Standard Error of Measurement

The reliability of a test may be expressed in terms of standard error of measurement. It is used for many testing purposes. The standard error of measurement is the standard deviation of the distribution of errors surrounding an individual's observed score. If standard error of measurement is small, then the precision of the measurement is greater. This statistic is often considered a more meaningful measure of an instrument's reliability (Magnusson [12, p. 82]). Based on the data for this study, the standard error of measurement was found to be small.

$$\text{Standard Error} = \sigma(1 - r_{tt})^{1/2}$$

$$= \sqrt{74.92}\sqrt{1 - 0.93}$$

$$= (8.66)(0.26)$$

$$\text{S.E.} = 2.19.$$

As it is found that the final 30 test items have both high reliability and high validity, and the standard error is small so they are accepted for the final test for the mathematics students of eighth standard.

Final Result

Table 3

Student	Item	Mean	S.D.	Average Difficulty Index	Average Discriminating Index	Reliability	Variance	S.E.
162	30	15.5	8.66	0.56	0.52	0.93	74.92	2.19

Conclusion

Developing a perfect test is the unattainable goal for everyone in an evaluative position. Looking at an item's difficulty and discrimination index will assist the test developer in determining what is wrong with individual items. Item and test analysis provide empirical data about how individual items and whole tests are performing in real test situations. The researchers made an attempt to show the uses of item analysis. Major purpose of item analysis is to improve tests by revising or eliminating ineffective items and to increase understanding of a test. The quality of the items in a test determines its validity and reliability. This paper addresses how educators can use item analysis test data as an invaluable tool to help inform mathematics instruction. The item analysis tool described and suggested in this article serves as a model that can provide powerful information to the classroom teacher and may develop into the basis for instruction design and lesson planning and modification. The tool reveals areas of strengths and weaknesses that may require needed changes in classroom instruction. It also provides data to help in assessing learning outcomes and course contents for students. Also, misconceptions in student thinking that constantly emerge in item analysis data may focus attention to the need for more effective teaching procedure. Therefore, item analysis data can uncover instructional weaknesses and clues for improvement. Through the application of item analysis procedures, researchers are able to obtain quantitative objective information useful in developing and judging the quality of a test and its items. Another important aspect of item analysis relates specifically to achievement tests which can provide important diagnostic information on what examinees have learned and what they have not. Pilot testing has

determined that the reliability index was 0.93 which is good index therefore the instrument is suitable to be used for real testing. The researchers suggest that item analysis should be used with a group of students as a pilot study before its application in the final test. This is because the question setter will be in a better position to know the loopholes whether any item is too easy or too difficult for the students and accordingly rectification and modification of the item can be made. It is hoped that this instrument will generate the result that will help the teachers to determine the weaknesses of their students in learning mathematics and will further improve their method of teaching to suit the learning ability of their students.

References

[1] L. W. Anderson, A comparison of classical item analytic procedures with affective data, A paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

[2] A. Binet and T. Simon, New methods for the diagnosis of the intellectual level of subnormals, H. H. Goddard, ed., Development of Intelligence in Children (the Binet-Simon Scale) (E. S. Kite, Trans., pp. 37-90), Williams and Wilkins, Baltimore, 1905/1916.

[3] H. E. Brogden, Variation in test validity with variation in the test distribution of item difficulty, number of items, and degree of their intercorrelation, Psychometrika 11 (1946), 197-214.

[4] L. Crocker and J. Algina, Introduction to Classical and Modern Test Theory, Holt, Rinehart and Winston, New York, 1986.

[5] R. L. Ebel and D. A. Frisbie, Essentials of Educational Measurement, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[6] C. T. Fan, Item Analysis Table, Educational Testing Service, Princeton, N.J., 1952.

[7] J. Flanagan, General considerations in selection of test items and a short method of estimating the product moment co-efficient from data at the tails of the distribution, J. Educational Psychology 30 (1939), 674-680.

[8] J. P. Guilford, Psychometric Methods, McGraw-Hill, New York, 1954.

[9] H. Gulliksen, The relation of item difficulty and inter-item correlation to test variance and reliability, Psychometrika 20 (1945), 79-91.

[10] T. L. Kelley, The selection of upper and lower groups for the validation of test items, J. Educational Psychology 30 (1939), 17-24.

[11] A. Lange, I. J. Lehmann and W. A. Mehrens, Using item analysis to improve tests, J. Educational Measurement 4(2) (1967), 65-68.

[12] D. Magnusson, Test Theory, Addison-Wesley, Reading, MA, 1966.

[13] S. Matlock-Hetzel, Basic Concepts in Item and Test Analysis, Texas A&M University, 1997.

[14] J. C. Nunnally, Educational Measurement and Evaluation, 2nd ed., McGraw-Hill Publishers Private Limited, New York, 1972.

[15] W. J. Popham, Modern Educational Measurement, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[16] G. Sax, Principles of Educational and Psychological Measurement and Evaluation, 3rd ed., Wadsworth, Belmont, CA, 1989.

[17] B. Thompson and J. E. Levitov, Using microcomputers to score and evaluate test items, Collegiate Microcomputer 3 (1985), 163-168.

[18] R. M. Thorndike, G. K. Cunningham, R. L. Thorndike and E. P. Hagen, Measurement and Evaluation in Psychology and Education, 5th ed., Macmillan, New York, 1991.

[19] W. Wiersma and S. G. Jurs, Educational Measurement and Testing, 2nd ed., Allyn and Bacon, Boston, MA, 1990.

[20] D. A. Wood, Test Construction: Development and Interpretation of Achievement Tests, Charles E. Merrill Books, Inc., Columbus, OH, 1960.