Development and Standardization of Achievement Test

Hassan Moshtaghian Abarghoie [1], Yusuf Mahmoudi Khamiripoor [2], Hossein Hosseini [2], Bahram Esmaeili [3], jamshid Moshtaghian Abarghoie[4], Hojat Moshtaghian Abarghoie[5]

1. Master of Educational Research, Tehran University Education Area 4 Shiraz, IRAN
2. Department of Public Administration, M.sc of HR management, Payame Noor University, PO BOX 19395-3697 Tehran, IRAN 5.
3. M.sc of HR management, The Holy Prophet Higher Education Complex, Tehran, IRAN. 0098 (0)7117335087; fax: 0098 (0) 71122810906.
4. Master of Educational Administration, Sharif Technical College of Abarkouh,Yazd, IRAN
5. Department of Public Administration, PhD student of Human Resource Management, Payame Noor University, PO BOX 19395-3697 Tehran, IRAN

hojatmoshtaghian@gmail.com

Abstract: The purpose of this study was to develop and standardization Achievement test to measure student learning in biology program at the secondary school. Both classical and IRT models were used to address the research objectives of the study. The preliminary instrument consisted of 150 multiple-choice items that on a sample size of 300 students were performed. The final instrument was two parallel forms of 50 items that on the normative sample volume of about 938 Students were performed. Estimated reliability coefficient for internal consistency with test forms was 0.89, 0.88, respectively. On the basis of factor analysis both forms of the test was an overall factor saturated. Results showed there is no significant difference between mean scores of boys and girls. So standardized and percentile Norms for all subjects calculated. Findings from the IRT analysis showed that more than 92 percent of the items are significantly fitted to Three-Parameter Logistic Model. Test information function was a bell-shaped curve and over a wide ability range from -0.5 to +2.5 provides more information. Also the maximum information was provided at +1.5 from ability continuum.

[Hassan Moshtaghian Abarghoie, Yusuf Mahmoudi Khamiripoor, Hossein Hosseini, Bahram Esmaeili, Jamshid Moshtaghian Abarghoie, Hojat Moshtaghian Abarghoie. **Development and Standardization of Achievement Test.** Journal of American Science 2012 8(4):166-168]. (ISSN: 1545-1003). http://www.americanscience.org. 22

Keywords: Achievement test, Classical model, Item-Response Theory, Standardization

1. Introduction

The most educator of Educational Sciences in the definition of education, it knows that a series of regular activities in order to create desired changes in behavior of learners (Kimble, 1968). According to this definition cannot claim that learning has been made without the measurement and evaluation of changes. Nowadays Measurement and evaluation of educational activities accounted for a significant part of it .Research carried out shows that in each teaching session between 5 to 15 minutes of class time is spent to measuring and evaluating (Pasha Sharife, 1999). Measurement and evaluation not only provides some of information about characteristics of students to teachers, but also can affect students' learning styles and strategies and so affect their level and speed learning (Bloom, at all, 1971).

There are two general approaches in the learning process: *rote learning approach* that puts the emphasis on memorization of unrelated facts and *deep approach* to learning that involves; exploring the deliberate and active for fundamental principles and concepts, and problem solving (Pasha Sharife, 1999). The teacher-made tests, often, simple elements, surface and material unrelated to curriculum content are emphasized and largely ignores the more complex and deeper knowledge, so the students employ rote learning approach (Hooman,1993). In this regard, (Lefrancois, 2000) shows teachers that their training methods were conducted by surface measuring can be trained to rote learners."

Although apparently it is thought that the assessment is end of the educational activities of teacher, but the reality today is that often assessment and measurement determines the teacher training activities and students learning largely by their performance on achievement tests is shown (Hogan, 2003).

Note that the teacher assessment can have a significant impact on the training -learning process (Cizek, 1993), so it is better teachers try to choose well and variety learning objectives for their students and assessment methods appropriate to that goals use. In this regard, (Woolfolk, 2004) said If the tests determine what teachers actually teach and what students learn - that it truly is – so the way of

improving education, is the direct way but uphill: to assess important and fundamental abilities and habits.

One of the text books of high school students in Iran is life science. Science education can play an important role in the development of scientific thinking and the spirit of truth-seeking students, although the weakness of the Iranian students in this subject is relatively serious (Kyamanesh, 1995). No doubt to overcome this weaknesses, science education should be reformed. But the true reform teaching of science is not possible without a proper evaluation; Efforts to improve science education can be effective only when the examinations and other assessment methods should be strengthened. Therefore the new tests designing as an essential part of science education reform process are acknowledged (Deighton, 1971).

Theoretical foundations of development and standardization of achievement tests based on psychometric principles and procedures were constructed. Today tow methods, the classical test model and *Item-Response Theory* (IRT), for constructing tests and interpreting scores have served measurement specialists well (Gulliksen,1950 & Hambleton, 1989). However, in recent years, due to limitations of classical theory and advent of computers and software, its application reduced and use of IRT is prevalent.

Hambleton et al. (1978) have identified three major limitations of classical test theory: 1) classical test theory indices (item difficulty, item discrimination) vary with the ability of the group on which they are computed. 2) Comparisons of examinees on an ability measured by a test are limited situations in which the examinees are given the same or parallel test of, test items. 3) Classical test theory provides no basis for predicting what an examinee might do when confronted with an item. Furthermore, the basic concepts and definitions of classical test theory are untestable, they are simply assumed to be true. There is no way to empirically determine their relevance of classical test theory's assumptions to reality.

It is desirable to have (a) item statistics that are not group dependent, (b) test scores are not dependent on test difficulty, (c) test models that provide a basis for matching test items to ability level, (d) test models that not based upon implausible assumptions. There is now substantial evidence to suggest that these desirable properties, and others, can be obtained within the framework of another measurement theory, known as *item response theory* (Hambleton, 1989). In item response theory postulates that (a) underlying examinee performance on a test is a single ability or trait, and (b) the relationship between examinee performance on each item and ability measured by the test can be described by a monotonically increasing function. The function is called an *item-characteristic function*, and it provides the probabilities of examinees at various ability levels answering the item correctly. Examinees with more ability have higher probabilities for giving correct answer to items than lower ability examinees. Item-characteristic functions, as they are commonly called in one-dimensional test models, are typically described by one-, two-, or three-item parameters (Hambleton, 1989). The *three-parameter model* is the most complex IRT model. In this model there are three parameters that must be estimated: *item difficulty*, *item discrimination*, and *item guessing* parameters. (Urry,1977) compared the one-, two-, and three-parameter models. The results show that the three-parameter model best described the multiple-choice tests using the real test data.

Based on, the propose of this study was that in order to measure academic achievement and ability level of students in biological science, according to the instructional objectives and curriculums for this subject a test developed and standardized.

Since the constructions of test require too sure technical and structural characteristics, to answer the following research questions was also considered.

1- Do the test have been developed has Sufficient of reliability and validity, so that it can be used as a reliable and valid instrument?

2 - Whether the test content based on factor analysis is saturated of a general factor?

3 – Is there between boy and girl students performance at the test different?

4 – How the Developed test is fitted to three parameter logistic model in item response theory?

5 – Item parameters and ability parameter estimation in the test is how much?

7 – How are the test norms for public school students?

2. Material and Methods

2.1. Participants

Statistical universe of this research was all first grade students in public high schools in Shiraz in year 2008. The total volume of universe was 39,039 students. From this community were selected 300 students randomly for the experimental sample and 938 students by multistage sampling for the final stage of test standardization.

Criteria of sample size in the empirical stage were the difficulty and discrimination parameters, and for the standardization sample important criteria were:

1 - The sample size is adequate for factor analysis.

2 - The sample size for data analysis with the *three parameter logistic model* in the item-response theory is sufficed.

2.2. Instrumentation

The final research instrument was two parallel forms (Form A and Form B) of a Biology Achievement Test (B.A.T.). In each form "50" four-choice questions was distributed. "8" percent of the total items in each form were include the concepts of scientific method, "17" percent the concepts of cellular and molecular biology, "15" percent plant biology concepts, "19" percent nutrition and health concepts, "29" percent concepts of ecology and "12" percent genetics and reproductive. The test items were assessed the students' knowledge and cognitive processes.

For development of the test, first the team of expert teachers designed 150 questions based on *table of specification*. The questions generated were distributed in three parallel forms - in terms of content and objective-. In the empirical execution, development test was performed on a sample size of 300 students (100 students per any form) in secondary public high schools of Shiraz who were randomly selected.

After data collection, test items were analyzed based on the following criteria:

1 - Difficulty of each item to be at least "0.25" and at the most "0.75".

2 - The discrimination indices of any item not to be less than of "0.3".

3 – Biserial correlation coefficient of each item to be at least "0.25".

4 - Distractors of any item to be have power of sufficient absorption.

Based on mentioned criteria, "100" items were selected for the final test. We distributed final items in two parallel forms, A and B, and arranged from simple to difficult.

We administered the final forms of "B.A.T.," on normative sample. So that the test forms in each class were randomly assigned to subjects. Time to answer questions set was "60" minutes per form. Each correct answer scores "1" and each wrong answer was awarded a zero score. Total raw scores for each subject were obtained by the sum of the number of correct answers.

2.3. Methods

We used the following statistical methods to analyze data and answer research questions:

1 - For analysis of test materials, both classical model and IRT were used. In other words, retention or removal of a test items in the empirical stage dependence on the index of difficulty, discrimination and correlation with the total score of test and the fitness of data with the logistic model, Parameters estimation , also based on the IRT.

2 - Cronbach's Coefficient Alpha (Crombach,1951) of the test was estimated in the classical model and the *test information function* was estimated in the IRT model.

3 – For determine that the content of BAT how many general factors is saturated, the principal component (PC) analysis is used.

4 - For examination Gender bias of BAT, t-test for independent groups was used.

3. Results

3.1. Reliability:

Since in this study any item at all or nothing (with values 1 or 0) were scoring, the formula 20 Kuder – Richardson (1937) was calculated for reliability. Reliability coefficients for Form A and B of BAT respectively "0.89" (standard error 4.35) and "0.88" (standard error 4.33) were estimated (Table 1). Be noted that these coefficients for both forms is fairly impressive, and this indicates that the "B.A.T.," has internal consistency satisfactory.

Table1. Reliability coefficients and standard errors of "B.A.T.," forms

Test form	N. Of items	alpha	standard error
A	50	0.886	4.35
B	50	0.884	4.32

3.2. Validity

Although several specific methods for validated educational tests have been described, content validity in the measurement of academic achievement is important. Therefore in this study in addition to *construct validity* (via factor analysis) *content validity* has been used too.

Content validity of the test from the beginning, with clear and detailed definition of the test domain and preparing *table of specification* and formulate test items based on learning objectives and content, is ingrained in it. Moreover, content analysis of items by the relevant experts departments of education is confirms the validity of "B.A.T.,"

The results of factor analysis based on the principal component (PC) method Showed that "15.7" percent of the total variance in form A and 15.6 percent of the total variance in form B by an overall factor which is quite distinct from the rest was explained. Also, 68.7 percent of common variance between items in form A and 69.3 percent of the common variance between items in form B is determined by this overall factor (table 2).

Table2. The Eigen value, the percentage of variance, and the percentage of cumulative for 50-items of B.A.T.

factor	Eigen values		% of variance		Cumulative %	
	A	B	A	B	A	B
1	7.840	7.787	15.68	15.57	15.68	15.57
2	1.836	1.890	3.67	3.78	19.35	19.35
3	1.738	1.546	3.48	3.09	22.83	22.45

Also all items on both test forms have positive loading for the first factor and most of them (92 percent) have load factor higher than 0.3.

3.3. Differences between group

We used T-test for independent groups to investigate whether gender factor in the total score of B.A.T are involved or not? Results showed; probability that the mean difference between the two groups may be due to chance is very high (table III). Therefore, sufficient evidence to reject the null hypothesis that means the same is not being. So between boys and girls performance in B.A.T was no significant difference.

Table3. The difference between boys & girls in B.A.T.

Test form	Boys		Girls			
	M	SD	M	SD	df	t
A	22.73	10.63	22.97	8.07	425.3	0.78*
B	22.34	10.39	23.03	8.02	424.7	0.42*

*$p < 0.05$

No significant difference between mean scores of boys and girls, this confirms that the structure is measured by a set of "B.A.T.," items, independence of gender of subjects.

3.4. Norm

In the present study we used of a relative index (derived scores), based on the classical theory, in order to expression the results of "B.A.T.," as the set of standard. For this the raw test scores using a frequency distribution, were converted to the *percentile scores* and the *normalized standard scores* (Z-scores and T-scores with average "50" and standard deviation "10").

3.5. Equating

After was approved both forms of test have the same statistical indicators (number of items, mean, standard deviation, reliability coefficient, and standard error), in order to examinee score does not affect the form of test, We used of bout linear and nonlinear equating for forms A and B of "B.A.T.," For linear equating, then calculating the mean and standard deviation of each of the groups participating in the test forms, Form A was chosen as the anchor test and based on using the line equation, Refer to "(1),", equivalent score of B (x_B) was calculated.

$$x_B = \bar{x}_B + s_B / s_A (x_A - \bar{x}_A) \quad (1)$$

Where "x_B¯" and "x_A¯" are means and "s_B" and "s_A" are standard deviations.

For nonlinear equating raw score corresponding to percentile selected (2th, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 98th) of the subjects in each form of B.A.T., were equal assumed.

3.6. Assessing goodness of fit

One of the most important topics in the item - response theory, to appropriate and fitness of a mathematical function model with gathering data from a test run. If the model and experimental data do not establish a statistically acceptable fit, the results of the model will be unstable and unacceptable.

In this study, to assess goodness of fit the three-parameter logistic model with data from the B.A.T., ASCAL software was used. ASCAL (17) employs a joint maximum likelihood approach with prior distributions imposed for the same models. Based on the results from 100 questions of two forms of B.A.T., only 8 questions with the model were not fitted. It is also probably due to some simple questions, called the base level and a number of difficult questions, called the roof level have less discrimination power than other questions.

3.7. Estimation of item parameters

1- Item difficulty (b_i): represents the point on the ability scale at which an examinee has a $(1 + C_i)/2$ probability of answering item correctly (Hambleton, 1989).Range of items difficulty for form A of "B.A.T.," between "-0.989" (item 1) to "+1.954" (item 45) and for form B of "B.A.T.," between "-0.962" (item 2) to "2.162" (item 49) was estimated. The mean of items difficulty for "B.A.T.," was "+0.73". So it is a few hard.

2- Item discrimination (a_i): range of items discrimination for form A of "B.A.T.," between "0.404" (item 2) to 1.993(item 35) and form B of "B.A.T.," between "0.408" (item 4) to "2.182" (item 48) was estimated. The mean of items discrimination for "B.A.T.," was "+1.15".

3- Pseudo guessing parameter (c_i): a value that is smaller than the value that would result if examinees of low ability were to randomly guess the item. Range of guessing parameter of "B.A.T.," from minimum "0.10" (item 2

form B) to maximum "0.27" (item 27 form A) were estimated. The mean of items guessing for "B.A.T.," was "0.21".

3.8. Item and test characteristic curve

Item Characteristic Curve (ICC) is a mathematical function that relates the probability of success on an item to the ability measured by the item set or the test that contains the item (Hambleton & Swaminthan, 1985)

And test characteristic curve reflects the monotonic relationship between true score and ability scores for a particular set of test items (Hambleton, 1989) In this study ICCs take the form of three-parameter logistic distribution functions:

$$p_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} (i = 1,2,...n). \quad (2)$$

Where $p_i\ (\theta)$, is the probability that a randomly selected examinee with ability "θ" will answer item "i" correctly, "b_i", "a_i", and "c_i" are parameters characterizing item "i", and "D" is equal to "1.7". The variable n is used to define the number of items in the test.

Refer to "(2)," probability of correct response for fixed level of ability calculated, and true-score estimates on the basis of sum of probability $(\sum p_i(\theta))$ were reported in the test-score metric.

3.9. Item and test information function

Information functions have a prominent role in IRT. Thus, the test information function is related to the accuracy with which we can estimate ability.

In this research Item-information functions by use of Birnbaum's Equation [19], Refer to "(3)," for fixed levels of ability (θ) was calculated:

$$I_i(\theta) = \frac{2.89a_i^2(1 - c_i)}{[c_i + e^{1.7a_i(\theta - b_i)}][1 + e^{-1.7a_i(\theta - b_i)}]^2} (i = 1,2,...n). \quad (3)$$

Where $Ii\ (\theta)$ is the information provided by item "i" at "θ", and "a_i", "b_i", and "c_i" were defined earlier.

The test information function of three parameter logistic model is the sum of the item information functions over the items in a test [12]. So the test information function for "B.A.T.," was plotted. As in Fig. 1, the "B.A.T.," is informative and spreads the information over a wide ability range from "-0.5" to "+2.5" and the most information is in "+1.5".

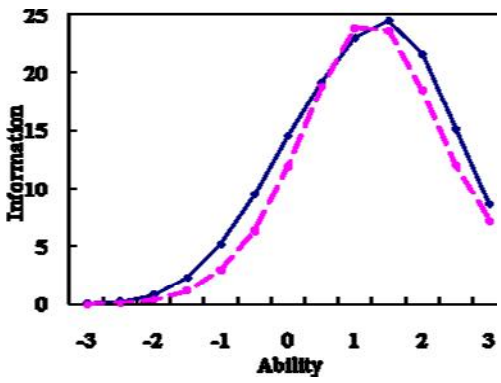

Fig. 1. Test information function for the "B.A.T.," (form A & B)

4. Discussions

The aim of this study was provide a tool to measure the level of training and knowledge of academic achievement of high school student in Biological Sciences. So two parallel forms of biological tests based on classical and IRT models were developed.

Cronbach's alpha reliability estimated showed that more than "88" percent of the variance in test scores related to the actual variance of the trait measured and "12" percent of remanence belongs to errors. So the set of "B.A.T.," questions has high and satisfactory internal consistency.

Although several methods have been described in validated educational testing, the content validity of academic achievement test is more important. In this study with respect to the exam questions were written by experienced teachers and other experts in education, on the basis of clear and detailed definition of the test domain, and formulate test items based on learning objectives and content, the validity of the "B.A.T.," is satisfactory.

The results of factor analysis showed the "B.A.T.," are saturated of a general factor. It can be concluded that the performance of students at "B.A.T.," largely influenced by an overall factor or ability that the test is basically constructed to measure it.

The t-test results showed that between boys and girls performance were not found significant differences in test. This means that the test is to unbiased sex groups and other word structures measured by the test subjects is independent of gender. This result is consistent with similar results (Kyamanesh, 1995 & Moshtaghian, 1999).

Findings from the analysis of questions based on the IRT model showed that the fit between experimental data and theoretical model is satisfactory. So as Hambleton and cook (1977) told, can conclude the three parameter logistic model is justified and valid

results. Results of three-parameter Logistic model for individual items also reflect the relationship between function in each of the test items with a single monotonic latent trait.

Results showed that the most items have strength discrimination power and are more useful for separating strong and weak examines, especially in the upper and middle range of the ability distribution. Whereas the high value of discrimination reveals that the most items relate to ability and a little depend on other factors (Thorndike, 1982).

On the basis of estimated difficulty parameter can be concluded that "B.A.T.," was relatively difficult.

The average of items guessing for "B.A.T.," was "0.21", that is smaller than "0.25", the value that would result if examinees of low ability were to randomly guess the item. Therefore the subjects have been ignored of blindly guessing, or distracters work well and can be attracted subjects with little ability (Hambleton, 1989).

Based on the test information function, "B.A.T.," spreads the information over a wider ability range (from "-0.5" to "+2.5") and the highest information is obtained at "+ 1.5". Since "information function is inversely proportional to the squared length of the asymptotic confidence interval for estimating ability from test score (Hambleton & Cook, 1977) and the standard error of measurement (SEM) is equal to the square root of the variance (Hambleton, 1989). We concluded that the "B.A.T.," provides more precise estimate of *True Score* of the subjects that their ability is up the middle point of continuum. It can be concluded that: the B.A.T is more reliable for the students with high ability levels.

Since each form of "B.A.T.," measure the same ability and have equal information functions, so is concluded that they are *weak parallel forms*.

Acknowledgements:

Authors thank to Abbas Bazargan for providing careful reviews and constructive suggestions for improving the research process.

Corresponding Author:
Dr. Hojat Moshtaghian Abarghoie
Department of Public Administration, PhD student of Human Resource Management, Payame Noor University, PO BOX 19395-3697 Tehran, IRAN
E-mail: hojatmoshtaghian@gmail.com

References

1. Birnbaum, A. "Some latent trait models and their use in inferring an examinee's ability". in *Statistical theories of mental test scores* F. M. Lord & M. R. Novick, Reading, MA: Addison-Wesley, 1968.
2. Bloom, S. B. and Hasting, J. T. And G. F. Madaus, *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.
3. Cizek, G. j. "Rethinking psychometrician's beliefs about learning," *Educational Researcher*, vol. 22, pp. 4-9, 1993.
4. Cronbach, L. J. "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, pp. 297–334, 1951.
5. Deighton, L. C. *The encyclopedia of Education.* New York: Macmillan, 1971.
6. Gulliksen, H. *Theory of mental tests*. New York: John Wiley and Sons, 1950.
7. Hambleton, R. K. "Principles and selected applications of item response theory," in *Educational Measurement,* 3nd Ed. L. Robert, linn, Ed. New York: Macmillan, 1989, pp. 147-200.
8. Hambleton, R. K. and Cook, Linda, L. "Latent trait models and their use in the analysis of educational test data," Journal *of educational measurement,* vol. 14, No. 2, pp. 75-93, 1977.
9. Hambleton, R. K. and Swaminathan, H. *Item response theory: Principles and application,* Boston: Nighff Publications, 1985.
10. Hambleton, R. K. Swaminathan, H. Cook, Eignor, L. D. and Gifford, J. "Developments in latent trait theory: Models, technical issues and applications," *Review of educational research*, vol. 48, pp. 467-510, fall 1978.
11. Hogan, T. P. *Psychological testing: A practical introduction*. John Wiley & Sons, International Edition, 2003, p 44.
12. Hooman, H. A. *Psychological and educational measurement and testing technique.* Tehran: Beautiful, 1993.
13. Kimble, G. A. *Hilgard and Marquis' conditioning and learning*, 2nd Ed. New York: Appleton-Century-Crofts, 1968.
14. Kuder, G. F, and Richardson, M. W. "The theory of estimation of test reliability," *Psychometrika,* vol. 2, pp. 151–160, 1937.
15. Kyamanesh, A. R. and et al "Fourth comprehensive assessment of the secondary education system," *Ministry of Education*, Tehran, Iran, Rep. Nov. 1995.
16. Lefrancois, G. R. *Psychology for teaching,* 10nd ed. Wadsworth, 2000, P. 489.
17. Moshtaghian, H. "Construction and standardization of since achievement test," M.A. thesis, Dept. Educational Psychology, Tehran Univ., Tehran, Iran, 1999.
18. Pasha Sharife, H. "Construction and standardization of Persian-Language achievement test,"p.15, 1999.
19. Thorndike, R. L. *Applied psychometrics*, Houghton Mifflin Company Boston, New Jersey, 1982.
20. Urry, V. W. "Tailored testing: A successful application of latent trait theory," *Journal of Educational Measurement*, vol. 14, no. 2, pp. 181-196, 1977.
21. Vale, C. D. and Gialluca, K. A. "ASCAL: A microcomputer program for estimating Logistic IRT item parameters," (Research Report ONR-85-4). St. Paul, MN: Assessment Systems Corporation, 1985.
22. Woolfolk, A. *Educational psychology,* 9nd ed. Pearson, International Edition, 2004, p. 555.

22/03/2012