Mathematical Problems of Computer Science 44, 7–21, 2015.

New Approach for Test Quality Evaluation Based on Shannon Information Measures

Mariam E. Haroutunian and Varazdat K. Avetisyan

Institute for Informatics and Automation Problems of NAS RA
e-mail: armar@ipia.sci.am, avetvarazdat@gmail.com

Abstract

There are two currently popular statistical frameworks for addressing test data analyzing: Classical Test Theory (CTT) and Item Response Theory (IRT). Each of these approaches has its advantages and disadvantages. The detailed description of CTT was given in the previous paper of V. K. Avetisyan [1]. In this paper the description of IRT models are provided to show the complex mathematical apparatus used. In this investigation we suggest a new model of test quality evaluation based on information measures such as Shannon entropy, conditional entropy and average mutual information.We show that this approach is easier and more informative.

Keywords: Test quality, IRT models, Test information function, Shannon entropy, Average mutual information.

1. Introduction

Test development is a very difficult and time-consuming process. It consists of the following main steps: (a) establishing the purpose of a test, (b) defining the construct of interest, (c) drafting test items, (d) conducting a pilot test, and (e) analyzing the item response data [9]. These steps are cyclical and the analysis of item response data leads to the identification of good items and informs revision of bad ones. Revised items are then subjected to additional rounds of pilot testing and analysis. Several authors provide detailed information about each of these steps [2].

The analysis of item response data is an important step in the development of quality characteristics of test and test items. Item analysis is a procedure for quantifying various characteristics of test items. It helps us to identify items that are too easy or excessively difficult for examinees, and ability to distinguish between low- and high-scoring examinees.

There are two currently popular statistical frameworks for addressing test data analyzing: Classical Test Theory (CTT)[3] and Item Response Theory (IRT) [4]. Both theories enable to predict outcomes of psychological tests by identifying parameters of item difficulty and the ability of test takers. Both are concerned to improve the reliability and validity of psychological tests [5]. Both of these approaches provide measures of validity and reliability. There are some identified issues in the classical test theory that concerns with calibration of item difficulty, sample dependence of coefficient measures, and estimates of measurement error which in turn is addressed by the item response theory.

CTT is regarded as true score theory. The central model of the classical test theory is that the observed test scores (TO) are composed of a true score (T) and an error score (E), where the true and the error scores are independent. A mathematical expression of the model is as follows: $TO = T + E$. Traditionally, methods of analysis based on classical test theory have been used to evaluate tests. The focus of the analysis is on the total test score; frequency of correct responses (to indicate question difficulty); frequency of responses (to examine distracters); reliability of the test and item-total correlation (to evaluate discrimination at the item level) [1, 8]. Although these statistics have been widely used, one limitation is that they relate to the sample under scrutiny and, thus, all the statistics that describe items and questions are sample–dependent [10].

IRT is a general statistical theory about the examinee item and the test performance and how performance relates to the abilities that are measured by the items in the test. IRT models show the relationship between the ability or trait (symbolized θ)measured by the instrument and an item response. IRT may be regarded as roughly synonymous with the latent trait theory. In IRT each item on a test has its own item characteristic curve that describes the probability of getting each particular item right or wrong given the ability of the test takers [11]. The item response may be dichotomous (two categories), such as right or wrong, yes or no, agree or disagree or, it may be polytomous (more than two categories).The IRT score is often called an ability, trait, or proficiency. The probability of a correct response is expressed as a function of θ. When the probability is calculated for a specific value of θ, it can be interpreted as the probability of a correct response for an examinee randomly selected from a group of examinees with that value of θ.

Two of the fundamental advantages of IRT models are unidimensionality and local independence. The unidimensionality means that a set of items and/or a test measure(s) only one latent trait (θ), and local independence means that there is no statistical relationship between examinees responses to the pairs of items in a test, once the primary trait measured by the test is removed. What follows are various models that make different assumptions about that relationship.

2. IRT Models

Within the general IRT framework, many models have been formulated for modeling of the relationship between the trait measured by the test and item responses. and applied to real test data [12]:

Unidimensional Dichotomous Models

- Normal Ogive Model
- One-Parameter Logistic Model (Rasch Model-1PLM)
- Two-Parameter Logistic Model (2PLM)
- Three-Parameter Logistic Model (3PLM)
- Nonparametric Model

Unidimensional Polytomous Models

- Partial Credit Model
- Generalized Partial Credit Model
- Rating Scale Model
- Graded Response Model
- Nominal Response Model (Nominal Categories Models)

Multidimensional Dichotomous Model
Compensatory Three-Parameter Logistic Model.

One-Parameter Logistic Model (1PLM a.k.a. Rasch Model)[8] has been developed by the mathematician George Rasch [4]. In this model the probability of a randomly chosen examinee at an ability level θ obtaining a correct answer on item i can be expressed as

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}, \tag{1}$$

where $P_i(\theta)$ is the probability of a randomly chosen examinee at ability level θ answering the item i correctly, e is an exponential constant whose value is about 2.718, D is a scaling factor whose value is 1.7, and b_i is the difficulty parameter of item i.

$P_i(\theta)$ is a monotonic increasing function, which means that the probability of a correct answer increases as the latent trait increases. An examinee with a high level of ability has a greater chance of answering an item correctly as compared to an examinee with low ability.

Item characteristic curve (ICC) illustrates the relationship between the latent trait and the probability of a correct answer. Figure 1 illustrates an item characteristic curve for the One-Parameter Logistic Model Item. Person ability is on the x-axis of this figure, and the probability of a correct response is on the y-axis. Item difficulty is also referred to as the location parameter because it affects how far the curve is shifted left or right along the x-axis. **The difficulty parameter** b tells how difficult the item is. Its value equals the θ value where the slope of the function is steepest. Low difficulty values shift the entire curve to the left, and high difficulty values shift the whole curve to the right. That is, small values of item difficulty represent easy items because moving the curve to the left increases the probability of a correct response at a given level of ability. Conversely, large values of item difficulty represents difficult items because moving the curve to the right lowers the probability of a correct response for a given level of ability. This interpretation of difficulty is opposite to that of item difficulty in a classical item analysis, but it is in the more intuitive direction: small values correspond to easy items, and large values– to difficult items. Because item difficulty affects the location of the item characteristic curve, one can obtain the item difficulty parameter value directly from a plot of the item characteristic curve. For the Rasch model, difficulty is the point on the x-axis where the probability of a correct answer is 0.5. You can find item difficulty from a plot (Figure1) by drawing a horizontal line from the value 0.5 on the y-axis until it reaches an item characteristic curve. Then, draw a vertical line down to the x-axis to obtain the difficulty value.

Figure 2 shows two 1PLM items with different b parameters. Item 1 is more difficult than item 2; for any given value of y, the probability of getting item 1 right is lower than the probability of getting item 2 right.

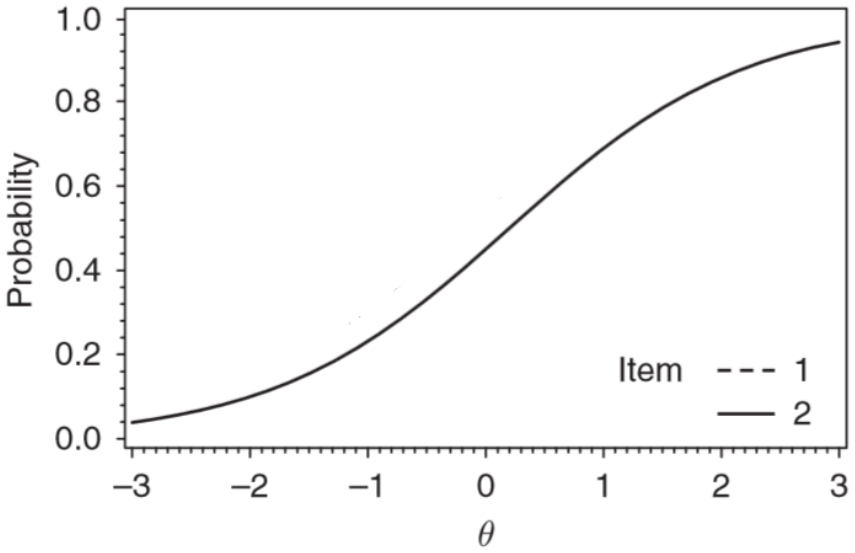

Fig. 1. 1PLM item characteristic curve.

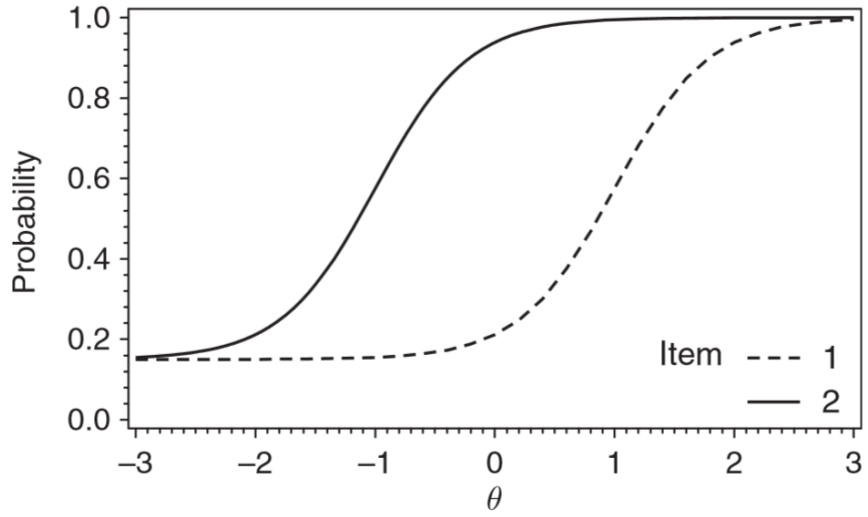

Fig. 2. 1PLM items with different b-parameters. Item1 is more difficult.

Two-Parameter Logistic Model (2PLM) is a generalization of the 1PLM. In this model test item is characterized by two parameters: the difficulty b parameter and the discrimination a parameter. Instead of having a fixed discrimination of 1 across all items as in 1PLM, in the 2PLM, each item has its own discrimination parameter. Thus, the model is mathematically expressed as

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \tag{2}$$

where a_i is the discrimination parameter of item i.

The discrimination parameter, or slope a, tells how steeply the probability of correct response changes at the steepest point on the curve. Figure 3 shows two items with different a parameters; item 1 has a higher discrimination than item 2. Thus, item 1 can better differentiate (discriminate) between an examinee with a moderately high θ and a moderately low θ.

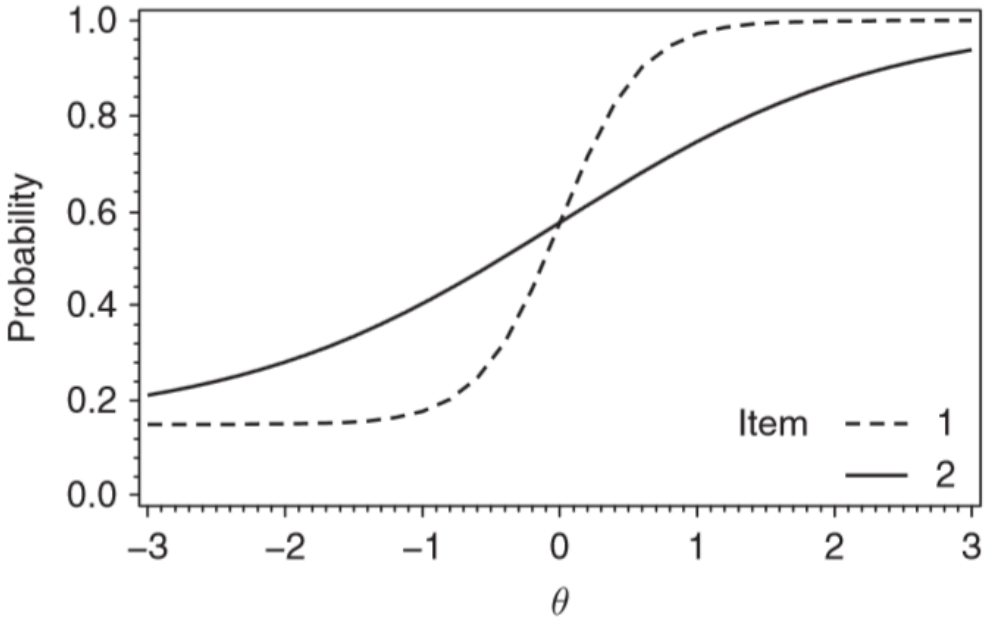

Fig. 3. 2PLM items with different a-paramaters.Item1 is more discriminating.

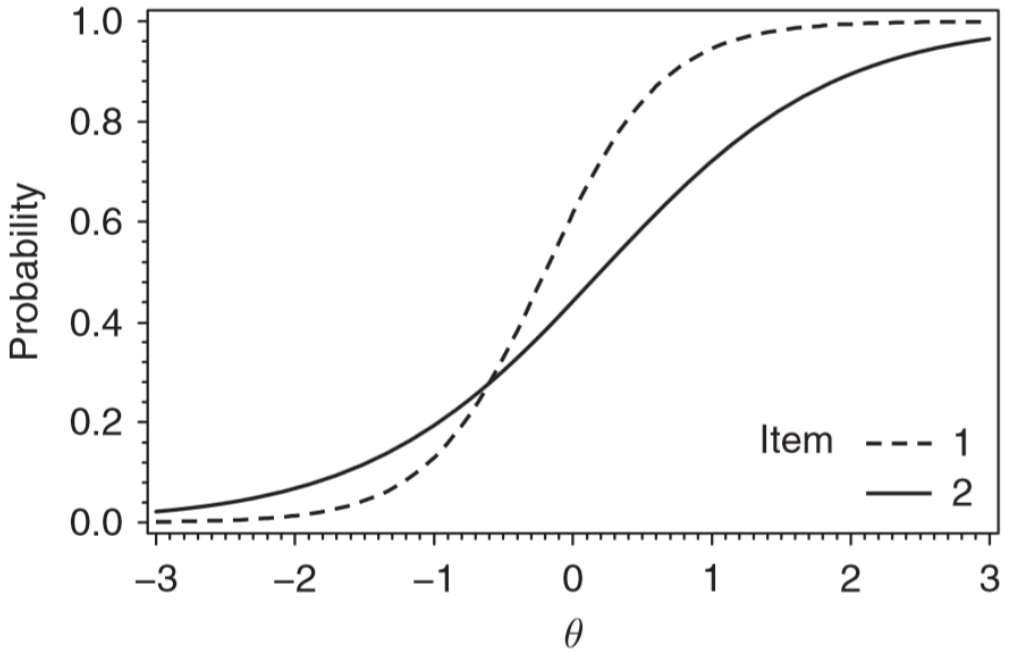

Fig. 4. 2PLM items. The items have different a-parameters and different b-parameters.

In Figure 4 items 1 and 2 are 2PLM items. They have different a and b parameters, but both functions approach a lower asymptote of zero.

Three-Parameter Logistic Model (3PLM) allows an ICC to have non-zero lower asymptotes. This model is more suitable for response data with those items in which examinees at the extremely low proficiency level may get the items correctly by chance; for

example, a multiple choice item. In this model

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \tag{3}$$

where c_i represents the probability that examinees at extremely low levels of the trait answer item i correctly. This third item parameter c_i is often called either the pseudo-chance-level parameter or the guessing parameter, although "pseudo-chance-level parameter" is theoretically more appropriate (Hambleton, Swaminathan, & Rogers, 1991). The 2PLM is a special case of 3PLM when $c = 0$, and 1PLM is a special case of 2PLM when $a = 1$. The lower asymptote parameter c, provides the probability that an examinee with a very

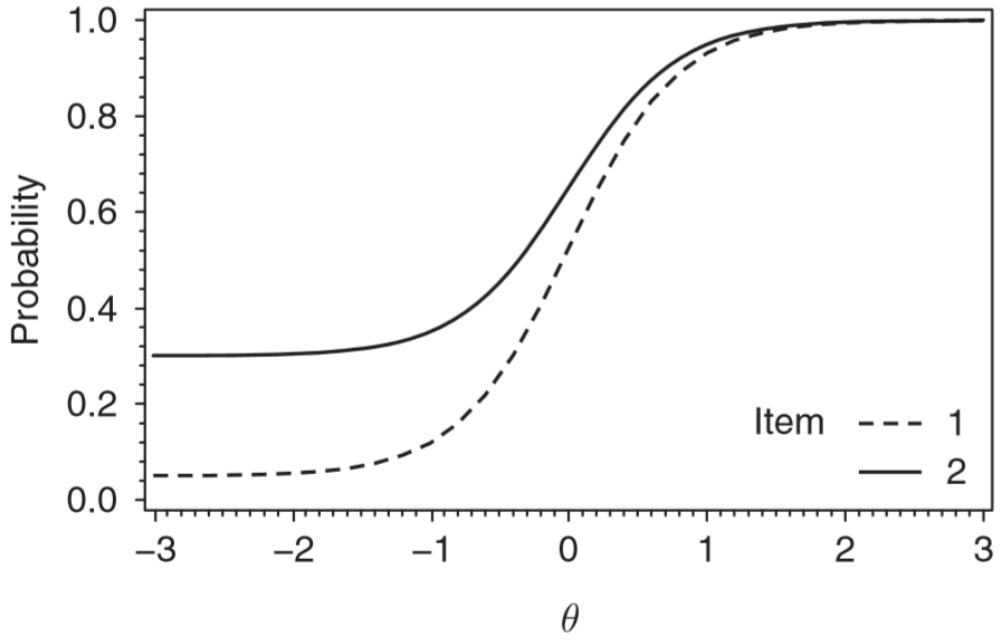

Fig. 5. 3PL items with different c-paramaters. Item 2 has a higer c-parameter.

low level of θ will answer the item correctly. The b-parameter is the point on the ability scale, where an examinee has $(1 + c)/2$ probability of a correct answer. The a- parameter is proportional to the slope of the ICC at the point b on the ability scale.

Polytomous models. Items that have more than two categories are labeled polytomous, or polychotomous. Several models have been proposed and used for polytomous items.

Partial Credit Model (PCM) is an extension of the 1PLM. Equation (1) for the 1PLM can be rewritten as

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}} = \frac{\exp(D(\theta - b_i))}{1 + \exp(D(\theta - b_i))} = \frac{P_{i1}(\theta)}{P_{i0}(\theta) + P_{i1}(\theta)}, \tag{4}$$

where $P_{i1}(\theta)$ is the probability of a randomly chosen examinee, whose proficiency level is θ, scoring 1 on item i, and $P_{i0}(\theta)$ is the probability of a randomly chosen examinee, whose proficiency level is θ, scoring 0 on item i.

Thus, the probability of a person at θ, scoring x over $x - 1$ can be computed as

$$\frac{P_{ix}(\theta)}{P_{ix-1}(\theta) + P_{ix}(\theta)} = \frac{\exp(D(\theta - b_{ix}))}{1 + \exp(D(\theta - b_{ix}))}, \qquad x = 1, 2, ..., m_i, \tag{5}$$

where $P_{ix}(\theta)$ and $P_{ix-1}(\theta)$ refer to the probabilities of an examinee at θ, scoring x and $x-1$, respectively. It should be noted that the number of item difficulty parameters are now m_i (one less than the number of response categories) in (5). The probability of a randomly chosen examinee, who has ability θ, scoring x on item i can be expressed as

$$P_{ix}(\theta) = \frac{\exp \sum_{k=0}^{x}(D(\theta - b_{ik}))}{\sum_{h}^{m_i} \exp \sum_{k=0}^{h}(D(\theta - b_{ik}))}, \quad x = 1, 2, ..., m_i. \tag{6}$$

The function of (6) is often called a score category response function (SCRF).

Figure 6 shows option characteristic curves for a partial credit item with four categories. In this figure, there are three places, where the curves for two adjacent categories intersect. The first intersection occurs at an ability level of about -1.7. This value is the step parameter for transitioning from a score of 0 to a score of 1. The next intersection point is at about 0.8, and it represents the point where the probability of scoring in category 1 is the same as scoring in category 2. The remaining step parameter is located at about 1.8 on the

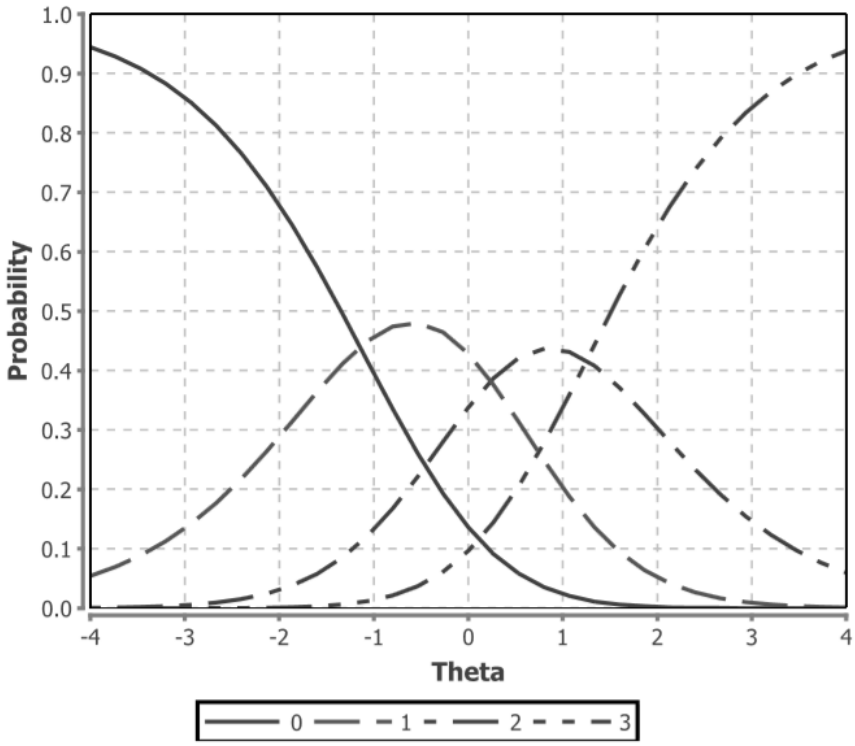

Fig. 6. Option characteristic curves for a partial credit item with four categories.

latent scale. Beyond this level of ability, examinees are most likely to obtain the highest possible item score. There are two aspects to consider when evaluating threshold parameters for a polytomous item response model. The first is the spread of threshold estimates. A nice feature of the item graphed in Figure 6 is that the step parameters range from about -1.7 to $+1.8$. A consequence of this spread of step parameters is that the item will provide information throughout most of the scales range.

Generalized Partial Credit Model (GPCM) [14] is a generalization of the PCM with a parameter for item discrimination added to the model. Muraki [14] expressed the model

mathematically as follows:

$$P_{ix}(\theta) = \frac{\exp \sum_{k=0}^{x}(Z_{ik}(\theta))}{\sum_{h}^{m_i} \exp \sum_{k=0}^{h}(Z_{ik}(\theta))}, \tag{7}$$

where

$$Z_{ik}(\theta) = Da_i(\theta - b_i + d_{ik}), \tag{8}$$

and d_{ix} is the relative difficulty of score category x of item i. Although Muraki followed the same way of parameterization for item and score category difficulty as Andrichs rating scale model, the item difficulty parameters for each score category can be simply rewritten as

$$b_{ix} = b_i - d_{ix}, \tag{9}$$

and so the equation (8) will become

$$Z_{ik}(\theta) = Da_i(\theta - b_{ix}), \tag{10}$$

The only difference between the PCM and GPCM is the additional discrimination parameter for each item (a_i).

There are various approaches in the construction of tests using IRT. Some approaches use the two-dimensions that plot item discriminations and item difficulties. Other approaches use a three-dimension for the probability of test takers with very low levels of ability getting a correct response. Other approaches use only the difficulty parameter (one dimension) such as the Rasch Model. All these approaches characterize the item in relation to the probability that those who do well or poorly on the exam will have different levels of performance.

3. Information Function

Characteristic curves are useful tools for illustrating the relationship between the probability of a response and person ability, but there is another function that describes the quality of measurement more directly. The item information function, $I_i(\theta)$, describes an items contribution to measurement of person ability such that the greater the amount of information, the more precision there is in estimating the person ability.
Item (Fisher) information function [8] mathematically is expressed as

$$I_i(\theta) = \frac{(P_i'(\theta))^2}{P_i(\theta) \cdot Q_i(\theta)}, \tag{11}$$

where $Q_i(\theta) = 1 - P_i(\theta)$ is the probability of incorrect answering to item i and is defined by

$$Q_i(\theta) = \frac{1}{1 + e^{1.7(\theta - \beta_i)}}. \tag{12}$$

The item information function of the 1PL model is actually quite simple:

$$I_i(\theta, b_i) = P_i(\theta, b_i)Q_i(\theta, b_i). \tag{13}$$

It is easy to see that the maximum value of the item information function is 0.25. It occurs at the point, where the probabilities of correct and incorrect responses are both equal to 0.5. In other words, any item in the 1PLM is most informative for examinees, whose ability is equal to the difficulty of the item. As ability becomes either smaller or greater than the item difficulty, the item information decreases. This relationship is illustrated in Figure 7. The item in this figure has a difficulty of 0.04, and the information function reaches its peak, when the ability level is 0.04. Thus, the item in Figure 7 provides the best measurement of people with an ability level of 0.04. It also provides useful information for examinees with ability between -1.0 and 1.0. Outside this range, the item does not contribute much to measurement of the latent trait. Additional items are needed to improve measurement precision at other parts of the scale.

For the 2PL model, the item information function becomes

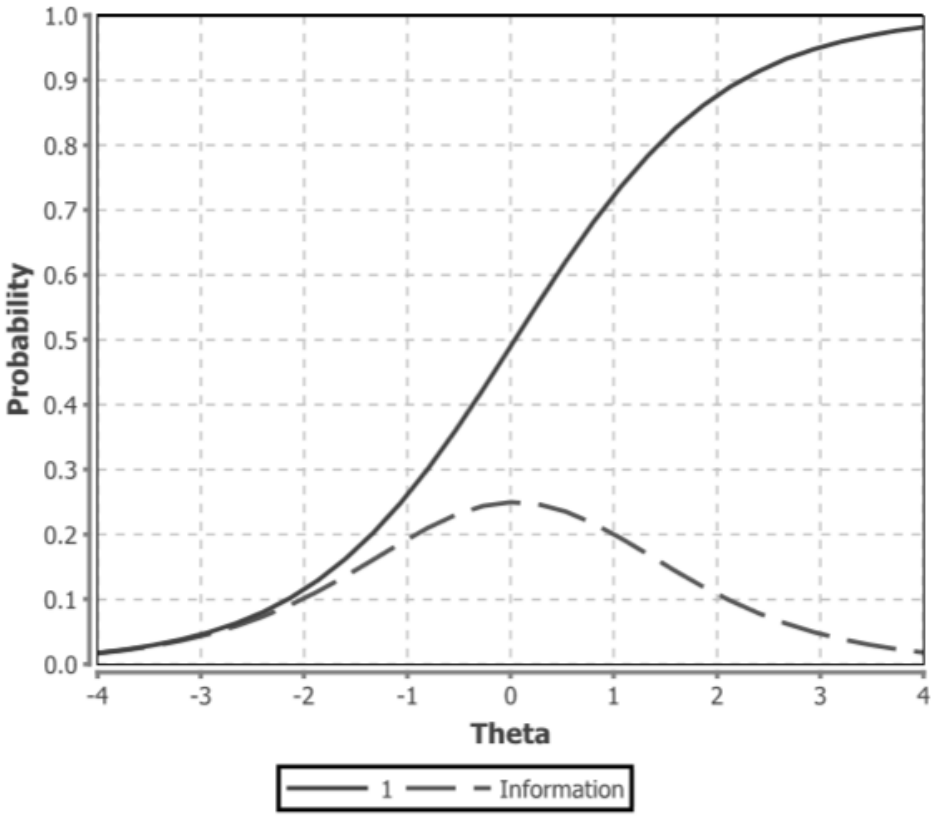

Fig. 7. An item characteristic curve and item information function.

$$I_i(\theta, b_i, a_i) = a_i^2 P_i(\theta, b_i) Q_i(\theta, b_i). \tag{14}$$

In the 1PL model, all item information functions have the same shape, the same maximum of 0.25, and are simply shifted along the ability axis such that each item information function has its maximum at the point, where ability is equal to item difficulty. In the 2PL model, the item information functions still attain their maxima at item difficulty. However, their shapes and the values of the maxima depend strongly on the discrimination parameter. When discrimination is high (and the item response function is steep), the item provides more

information on ability, and the information is concentrated around item difficulty. Items with low discrimination parameters are less informative, and the information is scattered along a greater part of the ability range.

The item information function of the 3PL model is

$$I(\theta, a, b, c) = a_i^2 \frac{Q(\theta)}{P(\theta)} \Big[\frac{P(\theta) - c}{1 - c} \Big]^2. \tag{15}$$

Item information functions combine in a simple way to summarize measurement precision for an entire test. Specifically, **the test information function (TIF)** is the sum of item information functions:

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta). \tag{16}$$

Test information guides the test development in IRT. It is used to design a test with maximal precision at important parts of the scale. An item from the Rasch family of models will contribute maximum information at the point where person ability equals item difficulty. As a result, if you need more information at a particular point on the scale, then choose items with that level of difficulty.

Maximizing information at an important region of the scale allows to minimize measurement error in that region because of the relationship between information and measurement error. The standard error (SE) of the examinee ability estimate is inversely related to information, SE $(\theta) = 1/\sqrt{I(\theta)}$. Small values of information translate to large standard errors, and large values of information result in small standard errors.

As Figure 8 shows in locations where there is more information, the standard error is lower.

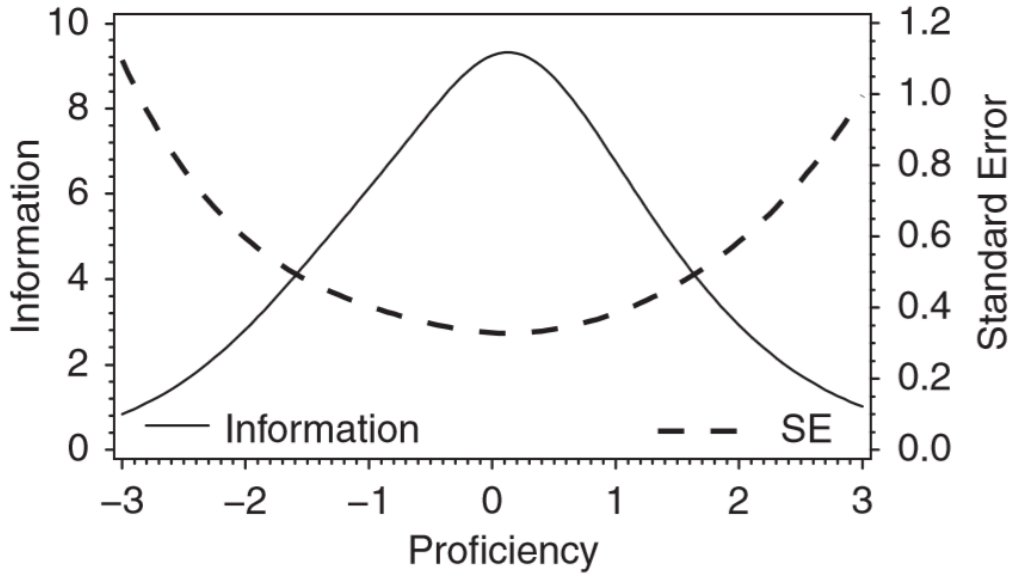

Fig. 8. Relationship between test information and standard error of the proficiency estimate.

4. Application of Shannon Information Theory Elements for Test Item Quality Evaluation

We suggest a new approach for test quality evaluation based on the information measures. Particularly, **Shannon entropy, conditional entropy, average mutual information [16]** are used for estimation. Suppose that the test consists of N items, each item can be considered as a random variable (RV). These variables we denote by $X_1, X_2, .., X_N$. Each RV can have two values: 1 for correct answer and 0 for incorrect with the corresponding probabilities p and $1-p$, where p is the frequency of correct answers in the total number of examinees:

$$X_i = \begin{cases} 1 & \text{with probability} \quad p_i, \\ & \qquad\qquad\qquad\qquad i = \overline{1, N} \\ 0 & \text{with probability} \quad 1 - p_i, \end{cases}$$

Shannon entropy of RV X_i

$$H(X_i) = -\sum_{x_i} p(x_i) \log p(x_i) \tag{17}$$

can be rewritten as

$$H(X_i) = -p_i \log p_i - (1 - p_i) \log(1 - p_i) = H(p_i). \tag{18}$$

We use logarithms of base 2.

The $H(p)$ function is illustrated in Figure 9.

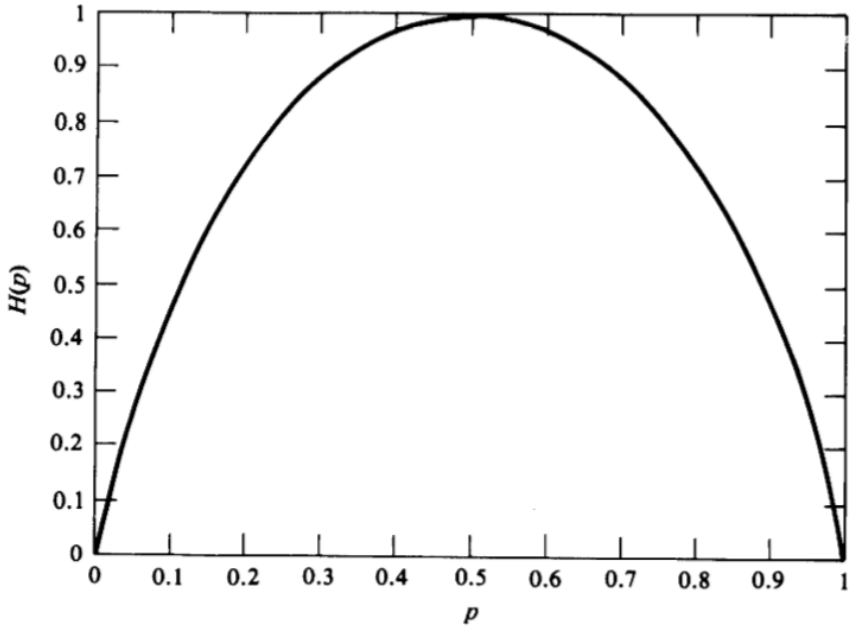

Fig. 9. $H(p)$ versus p.

We will consider also the conditional entropies of X_i given X_j.

$$H(X_i|X_j) = -\sum_{x_i,x_j} p(x_i,x_j)\log\frac{1}{p(x_i|x_j)}, \tag{19}$$

where $p(x_i,x_j)$ is the joint frequencies of two items and $p(x_i|x_j) = p(x_i,x_j)/p(x_j)$.
The average mutual information of two items is defined as:

$$I(X_i \wedge X_j) = \sum_{x_i,x_j} p(x_i,x_j)\log\frac{p(x_i,x_j)}{p(x_i)*p(x_j)} =$$

$$H(X_i) - H(X_i|X_j) = H(X_j) - H(X_j|X_i). \tag{20}$$

For formulating the suggested evaluation methods we use the following properties.

Property 1. In case of dichotomous data

$$0 \le H(X) \le 1,$$

where the left equality takes place if and only if $p_i = 0$ or $p_i = 1$, and the right equality takes place if and only if $p_i = 1/2$.

Property 2. The following inequalities take place

$$0 \le I(X_i \wedge X_j) \le \min[X(X_i), H(X_j)], \quad i,j = \overline{1,N},$$

where the left equality takes place if and only if X_i and X_j are independent.

Property 3. The conditional entropy does not exceed the non–conditional entropy.

$$H(X_i|X_j) \le H(X_i), \quad i,j = \overline{1,N},$$

where the equality takes place if and only if X_i and X_j are independent.

Based on these properties we formulate our test quality evaluation model.

Method 1. If value of $H(X_i)$ is close to 0, it means that we have a bad test item, which can be very easy or very difficult. If value of $H(X_i)$ is close to 1 we have a good test item.

Method 2. If value of $I(X_i \wedge X_j)$ is close to 0, it means that there is independency of test items X_i and X_j. In case of values close to $\min[H(X_i), H(X_j)]$, X_i and X_j items repeat each other.

Method 3. If value of conditional entropy $(H(X_i|X_j))$ is close to $H(X_i)$, then X_i and X_j are independent.

The error of measurements goes to 0 when the number of examinees grows. We can calculate the mentioned characteristics based on the test data.

Example. Let the experiment consist of 1000 examinee and 10 items.

Response data and calculated $H(p)$, CTT item difficulty and IRT item b parameter values are presented in Table 1.

Table 1:

	P_i for correct answers	$1-P_i$ for incorrect answers	$H(p)$ entropy	CTT item difficulty	IRT item b parameter
X_1	0.558	0.442	0.99	0.5580	0.34
X_2	0.408	0.592	0.98	0.4080	0.83
X_3	0.633	0.367	0.95	0.6330	0.72
X_4	0.850	0.150	0.61	0.8500	2.07
X_5	0.597	0.403	0.97	0.5970	0.64
X_6	0.446	0.554	0.99	0.4460	0.57
X_7	0.486	0.514	1	0.4860	0.06
X_8	0.754	0.264	0.81	0.7540	1.05
X_9	0.516	0.484	1	0.5160	0.10
X_{10}	0.532	0.468	1	0.5320	0.29

According to CTT test items have average difficulty (good items) if difficulty values are between 0.3 and 0.74. In our case $X4$ and $X8$ items are easy. The same conclusion can be drawn based on calculated $H(p)$ values. Also we can see that $H(p)$ value of $X4$ is smaller than the value of $X8$ item. It means that $X4$ item is easier than $X8$ item. If we compare IRT b parameters values we can see that $X4$ and $X8$ items differ with their values, too. The correlation between test items can be calculated in the scope of CTT. Results are presented in Table 2.

Table 2:

	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$	$X9$	$X10$
X_1	1.0000	0.0915	0.1244	0.0716	0.0898	0.0816	0.1926	0.1462	0.1212	0.0530
X_2	0.0915	1.0000	0.0411	0.0923	0.0059	0.0370	0.1495	0.0868	0.0630	0.0528
X_3	0.1244	0.0411	1.0000	0.0927	0.1104	0.0404	0.1509	0.2106	0.1261	0.1216
X_4	0.0716	0.0923	0.0927	1.0000	0.1173	0.0727	0.0835	0.2087	0.0863	0.1111
X_5	0.0898	0.0059	0.1104	0.1173	1.0000	0.0810	0.1381	0.1366	0.1099	0.0752
X_6	0.0816	0.0370	0.0404	0.0727	0.0810	1.0000	0.1258	0.0781	0.0196	-0.0253
X_7	0.1926	0.1495	0.1509	0.0835	0.1381	0.1258	1.0000	0.1745	0.1290	0.0740
X_8	0.1462	0.0868	0.2106	0.2087	0.1366	0.0781	0.1745	1.0000	0.1763	0.0925
X_9	0.1212	0.0630	0.1261	0.0863	0.1099	0.0196	0.1290	0.1763	1.0000	0.0701
X_{10}	0.0530	0.0528	0.1216	0.1111	0.0752	-0.0253	0.0740	0.0925	0.0701	1.0000

Correlation coefficient ranges from -1 +1. Coefficient value should be small or equal to 0.3. If coefficient value is close to +1, it means that test items repeat each other and one of those items should be removed from the test. In our case we have a negative correlation between $X6$ and $X10$ items. It means that those items are more independent. Also we have calculated mutual information function and conditional entropies between some items. Results are presented in Table 3 and Table 4. It is easy to see that mutual information function between $X6$ and $X10$ items is close to 0, which is an indicator of independency (method 2). The same conclusion can be drawn on $X6$ and $X10$ items based on method

3. As we can see condtional entropy of $X6$ and $X10$ items is maximum in case of $X6$ and $X10$ items, and conditional entropy is close to entropy of $X10$ item and it means that those items are not correlated.

Table 3:

$I(X_i \wedge X_j)$	X2	X4	X8	X10
X_1	0.006067595	0.1777555	0.01534219	0.002029427
X_6	0.007993387	0.003867708	0.004439025	0.0004612132

Table 4:

$H(X_i\|X_j)$	X2	X4	X8	X10
X_1	0.9693708	0.5652201	0.7895343	0.9950139
X_6	0.9708462	0.6059726	0.8004375	0.9965821

5. Conclusion

So, besides existing CTT and IRT models, we have developed a new approach based on Information Theory Measures. Its application is not very complex and is sufficiently informative.

References

[1] V. Avetisyan, "Investigation of knowledge control tests quality characteristics", *Proceedings of Engineering Academy of Armenia*, vol.11, no. 1, pp. 156-163, 2015.
[2] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory*, New York, NY: Holt, Rinehart and Winston. 1986.
[3] M. B. Chelishkova, *Theory and Practice of Pedagogical Tests Constructing*, Moscow: Logos, 2002.
[4] C.DeMars, *Item Response Theory*, Oxford University Press, 1 edition, 2010.
[5] R. K. Hambleton and R. W. Jones, Comparison of classical test theory and item response theory and their applications to test development , *Educational Measurement: Issues and Practice*, vol. 12, no.3, pp. 3847, 1993.
[6] C. Spearman, The proof and measurement of association between two things, *American Journal of Psychology*, vol. 15, pp. 72–101. 1904.
[7] M. R. Novick, The axioms and principal results of classical test theory, *Journal of mathematical psychology*, vol. 3, pp. 1 18. 1966.
[8] V. S. Kim, *Testing of Educational Achievements*, Ussuriysk: USPI Publishing, 2007.
[9] V. S. Avanesov, "The bases of the scientific organization of pedagogical control in the higher school", M, 1987.
[10] R. K. Hambelton, *Emergence of item response modeling in instrument development and data analysis*, *Medical Care*, vol. 38, pp. 60-65. 2000.
[11] R. M. Kaplan and D. P. Saccuzo , *Psychological testing: Principles, applications and issues*, Pacific Grove: Brooks Cole Pub. Company. 1997.

[12] Wim J. van der Linden and R. K. Hambleton, *Handbook of Modern Item Response Theory*, New York: Springer-Verlag. 1997.
[13] R. K. Hambleton, H. Swaminathan and H. J. Rogers, *Fundamentals of item response theory*, Newbury Park, CA: Sage. 1991.
[14] E. Muraki, RESGEN: Item response generator [computer program]. Princeton, NJ: Educational Testing Service. 1992.
[15] F. M. Lord, *The relation of the reliability of multiple-choice tests to the distribution of item difficulties*, Psychometrika, vol. 17, pp. 181-194. 1952.
[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 2nd Edition, 2006.

Submitted 02.08.2015, accepted 26.11.2015

Շեննոնի ինֆորմացիայի չափի մեծությունների վրա հիմնված թեստերի որակի գնահատման նոր մոտեցում

Մ. Հարությունյան և Վ. Ավետիսյան

Ամփոփում

Ներկայումս կա երկու հանրահայտ մոտեցում թեստերի տվյալների վիճակագրական վերլուծության համար՝ թեստերի դասական տեսությունը և ժամանակակից տեսությունը: Դրանցից յուրաքանչյուրն ունի իր առավելություններն ու թերությունները: Դասական տեսության մանրամասն նկարագրությունը ներկայացվել է Վ. Ավետիսյանի նախորդ հոդվածում [1]: Այս հոդվածում նկարագրվում են ժամանակակից տեսության մոդելները՝ մաթեմատիկական ապարատի բարդությունը ներկայացնելու համար: Այս հետազոտությունում մենք առաջարկում ենք թեստերի որակի գնահատման նոր մոդել՝ հիմնված ինֆորմացիայի չափի մեծությունների վրա, մասնավորապես, Շեննոնի էնտրոպիայի, պայմանական էնտրոպիայի և միջին փոխադարձ ինֆորմացիայի վրա: Մենք ցույց ենք տալիս, որ այս մոտեցումն ավելի պարզ է և ինֆորմատիվ:

Новый подход к оценке качества тестов на основе мер информации Шеннона

М. Арутюнян и В. Аветисян

Аннотация

В настоящее время есть два популярных статистических подхода для анализа тестовых данных: классическая теория и современная теория. Каждый из этих подходов имеет свои преимущества и недостатки. Подробное описание классической теории было приведено в предыдущей статье В. Аветисяна [1]. В данной статье приведено описание моделей современной теории, чтобы показать сложность используемого математического аппарата. В данном исследовании мы предлагаем новую модель оценки качества тестов, основанную на мерах информации, таких, как энтропия Шеннона, условная энтропия и средняя взаимная информация. Мы показываем, что этот подход является более простым и информативным.