Mathematical Problems of Computer Science 45, 35--43, 2016.

Analysis of Experiments of a New Approach for Test Quality Evaluation

Mariam E. Haroutunian, Varazdat K. Avetisyan

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

Abstract

In the previous paper [1] we suggested a new model of test quality evaluation based on Information measures such as Shannon entropy and average mutual information. To establish the practical bounds of these measures and the required number of examinees, some experiments were conducted. In this paper the analysis of these experiments are provided.

Keywords: Test quality, Shannon entropy, Average mutual information, Classical test theory, Item response theory.

1. Introduction

Test developers are basically concerned about the quality of test items and how examinees respond to it when constructing tests. Test theories and related models provide a frame of reference for doing test design work or solving other practical problems. A good test model might specify the precise relationships among test items and ability scores, so that careful design work can be done to produce desired test score distribution and errors of the size that can be allowed. A good test theory or model can also handle errors of measurements by helping understand the role that measurement errors play in estimating examinee’s ability and correlations between variables and true scores or ability scores. There are two currently popular statistical frameworks to address test data analysis and test quality evaluation: Classical Test Theory (CTT) [2] and Item Response Theory (IRT) [3]. CTT is a theory about test scores that introduces three concepts - test score, true score and error score. In the CTT, the notion of ability is expressed by the true score, which is defined as "the expected value of observed performance on the test of interest." An examinee's ability is defined only in terms of a particular test. When the test is "hard," the examinee will appear to have low ability; when the test is "easy," the examinee will appear to have higher ability. CTT was the dominant statistical approach for testing data until Lord and Novick (1968) placed it in the context with several other statistical theories of mental test scores, notably IRT. IRT is a model-based measurement statistical theory in which the performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and the relationship between the examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC). Each of these approaches has its advantages and disadvantages [4]. For example, in CTT item parameters are dependent on the examinee sample from which they are obtained, but in IRT these parameters are examinee group independent. But on the other hand, in case of CTT smaller examinee sample sizes are required for analysis and the methods are simpler compared to IRT. Besides the existing CTT and IRT models, we have developed a new approach [1] based on Information measures such as Shannon entropy and average mutual information.

The main idea of the new approach is the following. Suppose that the test consists of N items, each item can be considered as a binary random variable (RV) X_1, X_2, .., X_N with probabilities p for correct answers and $1-p$, for incorrect answers:

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

We consider Shannon entropy of RV X_i

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

and the average mutual information of two items:

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

Our test quality evaluation model consists of the following methods:

Method 1. If the 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 the value of $H(X_i)$ is close to 1 we have a good test item.

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

Method 3. If the value of conditional entropy $\left(H(X_j \mid X_i)\right)$ is close to $H(X_i)$, then Xi and Xj
are independent.

However, several questions remain open.

1. How precisely our model evaluates the quality of test items and how comparable is it to CTT and IRT estimation methods?
2. Which are the permissible limits of $H(X_i)$ and $I(X_i \wedge X_j)$?

3. Which is the sufficient number of the examinee samples for precise evaluation?

The answers to these questions can be found experimentally.

2. Description of Experiments

The results of school final exams of Armenian Language and Literature held in 2008 were selected for testing. The results were provided in encrypted form by the Center for assessment and testing. Four test-results are chosen to be analyzed. Each test consists of 80 items, and the number of schoolchildren who participated in the examination process is 2000. The names of the first 50 X_i items are $A_1, A_2, \ldots A_{50}$ and the names of the last 30 X_i items are $B_1, B_2, \ldots B_{30}$. For analysis test quality evaluation system developed by us was used in [5].

For each item of four tests the $H(X_i)$, CTT difficulty index [2] and IRT b parameter [3] values have been calculated, the comparability of the mentioned parameters observed and the permissible limits of $H(X_i)$defined. Difficulty is defined in both CTT and IRT.

In CTT **the difficulty index P** is the proportion of examinees who answer the item correctly. For multiple-choice, true/false, and other items that are scored as right (1 point) or wrong (0 points), item difficulty is the proportion of examinees who answered the item correctly. It ranges from 0 to 1. Item difficulty for a polytomous item (an item scored in more than two ordinal categories) is simply the item mean or average item score. It ranges between the minimum and the maximum possible item scores.

In IRT **the difficulty index b (IRT b parameter)** is on the same metric as the proficiencies or traits. This metric is arbitrary, but often it is anchored so that the proficiency distribution in a designated group has a mean of 0 and standard deviation of 1. The item difficulty identifies the proficiency at which about 50% of the examinees are expected to answer the item correctly.

To observe the dependency of $H(X_i)$ and $I(X_i \wedge X_j)$ values on the examinee sample size and define the enough number of examinee samples five experiments are carried out for each test. The analysis was conducted by choosing the same test at random based on 500, 300, 200, 100, 50 participants' results.

For each test $I(X_i \wedge X_j)$ and CTT correlation coefficient $R(X_i, X_j)$ between X_i and X_j items [2] was calculated, their compatibility was observed and the permissible limits of $I(X_i \wedge X_j)$ were defined. 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 that items should be removed from the test. The negative correlation means there is an independency of test items.

3. The Analysis of Results.

Based on the first experiment results comparison of $H(X_i)$, CTT difficulty index P and IRT b parameter values of Test1 are shown in Figure 1.

According to CTT test items for which difficulty values are between 0.3 and 0.74 interval are good items (not easy and not very difficult - 34 items), and based on the analyzed data we can see that $H(X_i)$ values of Test 1 for these items are between 0.82 and 1.0. For easy test items, difficulty values are between 0.75 and 0.9 (29 items), $H(X_i)$ values are between 0.48 and 0.81. For very easy test items difficulty values are between 0.9 and 1.0 (14 items), $H(X_i)$ values are between 0.12 and 0.47. Approximately the same results were obtained for the tests 2, 3 and 4.

As we can see in case of $H(X_i)$'s large values close to 1 IRT b parameter gets large values.

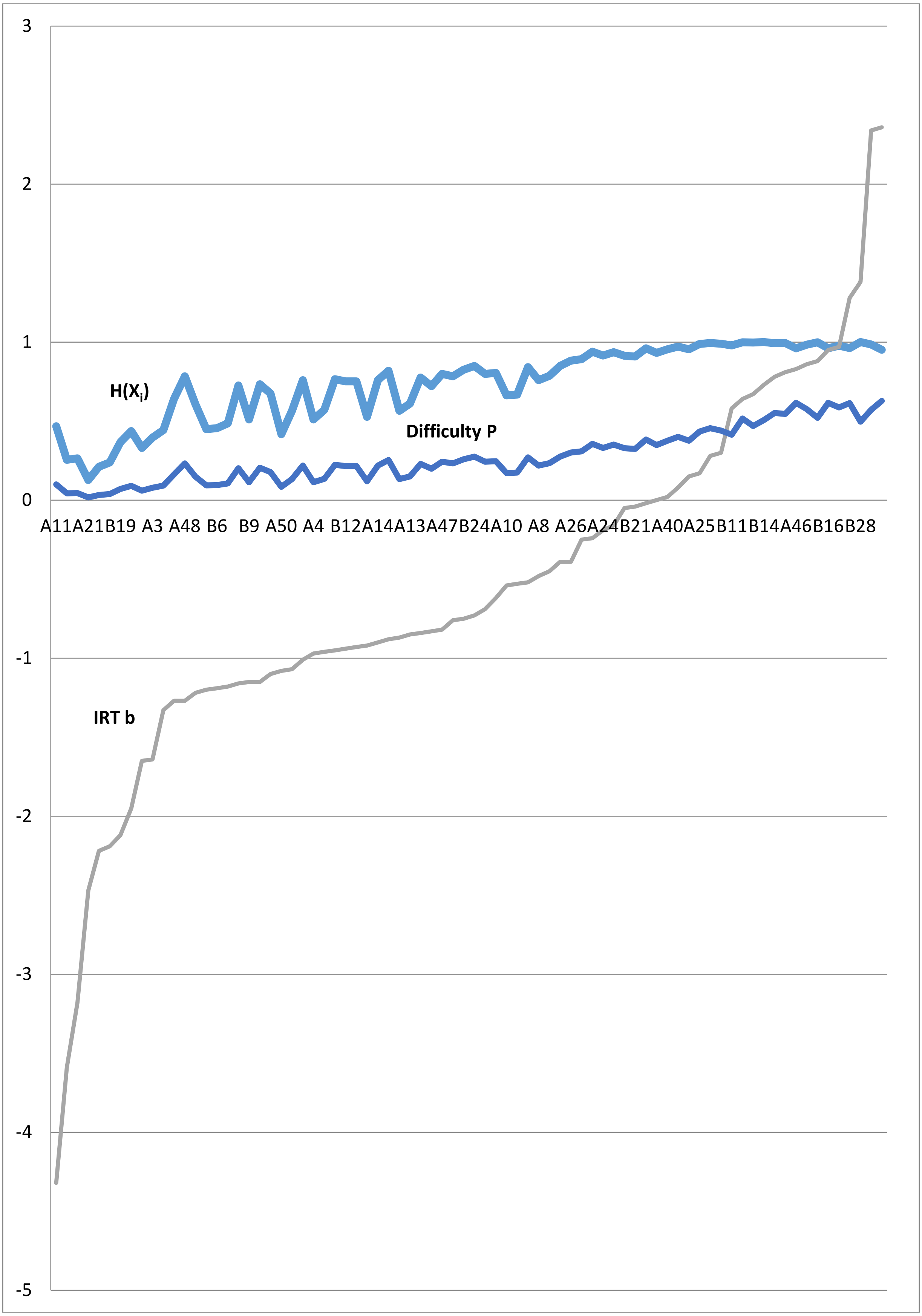

Fig. 1. $H(X_i)$, CTT difficulty parameters P and IRT b parameters of Test1.

To analyze the dependency of $H(X_i)$ values on the number of examinee sample we draw $H(X_i)$ graphics of each test based on the results of five experiments. The graphics are shown in Figure 2.

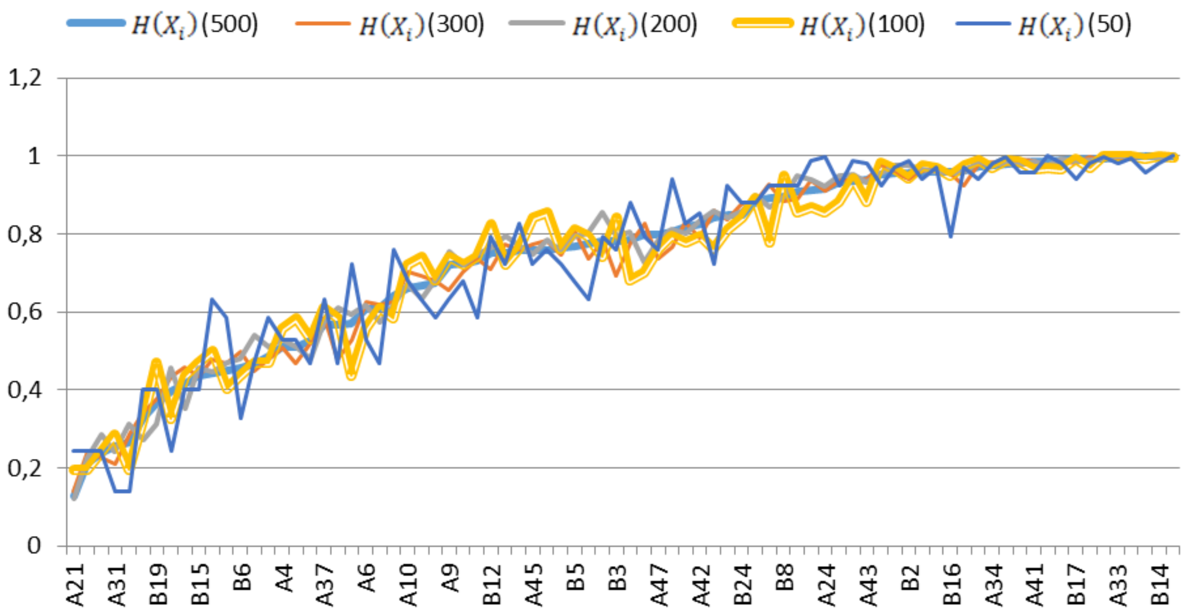

Fig. 2. $H(X_i)$ graphics for five experiments of Test1.

The maximum differences of $H(X_i)$ values are presented in Table1.

Table 1.

Maximum difference of $H(X_i)$ **values**	500	300	200	100	50
500	-	0.089	0.07	0.135	0.2
300	0.089	-	0.13	0.15	0.19
200	0.07	0.13	-	0.16	
100	0.135	0.15	0.16	-	0.3
50	0.2	0.19	0.23	0.3	-

While decreasing the examinee sample size until 100, it is obvious that the differences of $H(X_i)$ values are small and the maximum difference is 0.15. But when examinee sample size is decreased more than 100, the difference is close to 0.3, and in case of values equal to 50 the difference is close to 0.3.

With the same principle for each test the mutual information $I(X_i \wedge X_j)$ was calculated and the dependency of $I(X_i \wedge X_j)$ values on the number of examinee sample was observed. The graphics are shown in Figure 3.

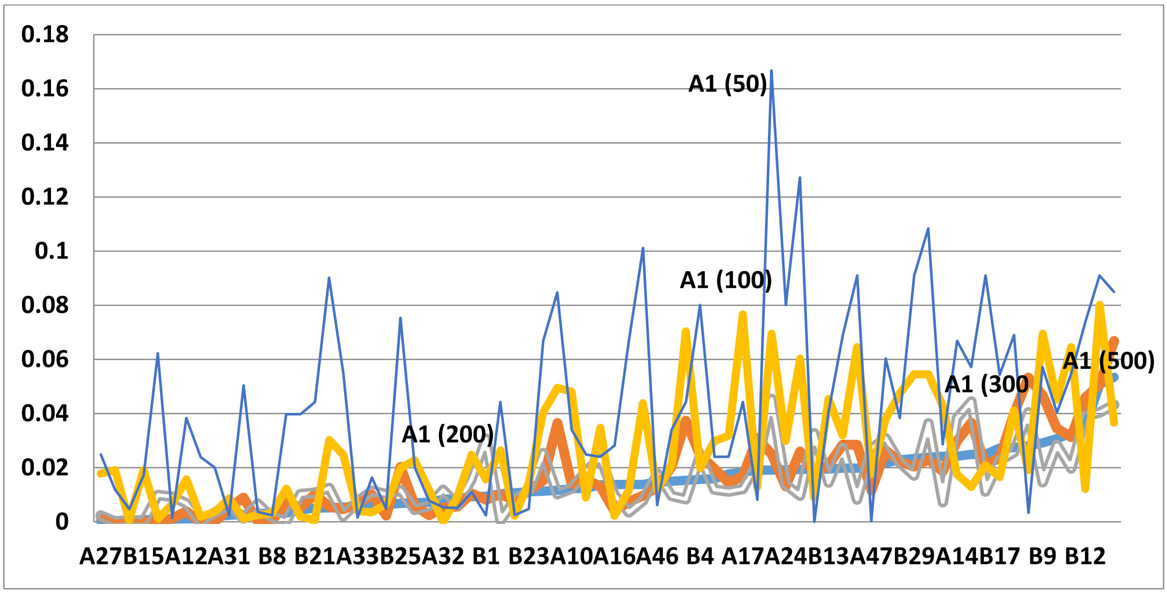

Fig. 3. $I\left(X_i \wedge X_j\right)$ for five experiments of Test1 A1 item.

The average mutual information and correlation between test items also have been analyzed. The graphics based on some items' data are presented in Figure 4 and Figure 5.

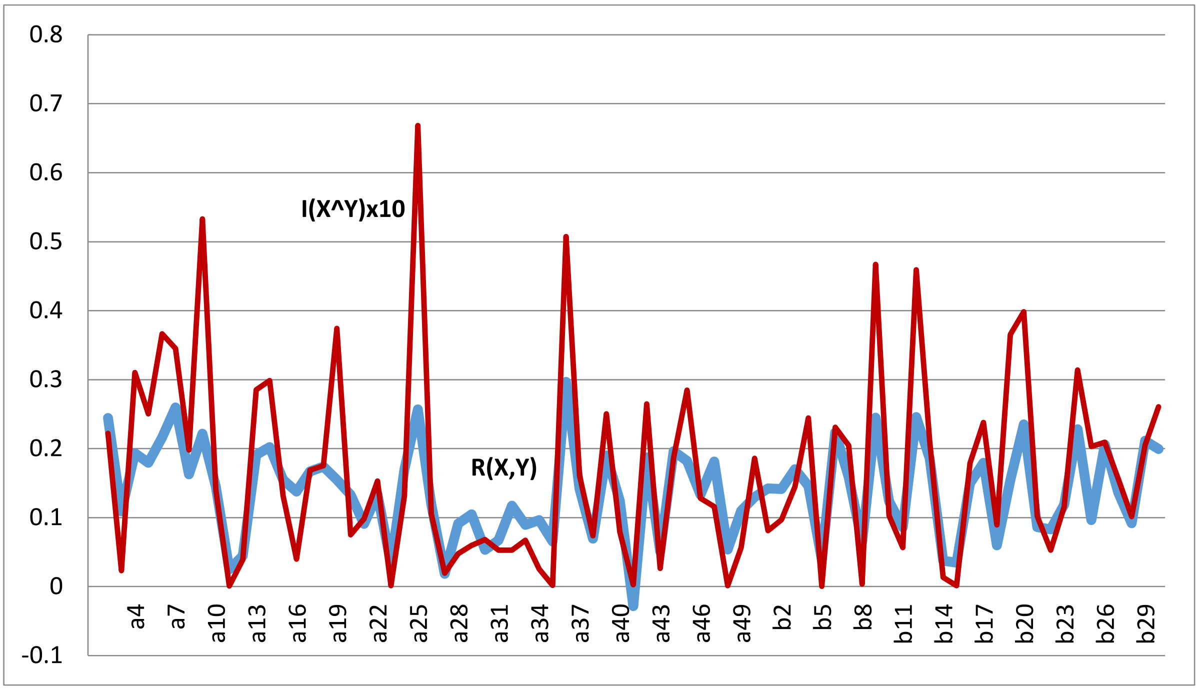

Fig. 4. Correlation $\left(R\left(X_i, X_j\right)\right)$ and average mutual information $\left(I\left(X_i \wedge X_j\right)\right)$ between A1 item and other 80 items.

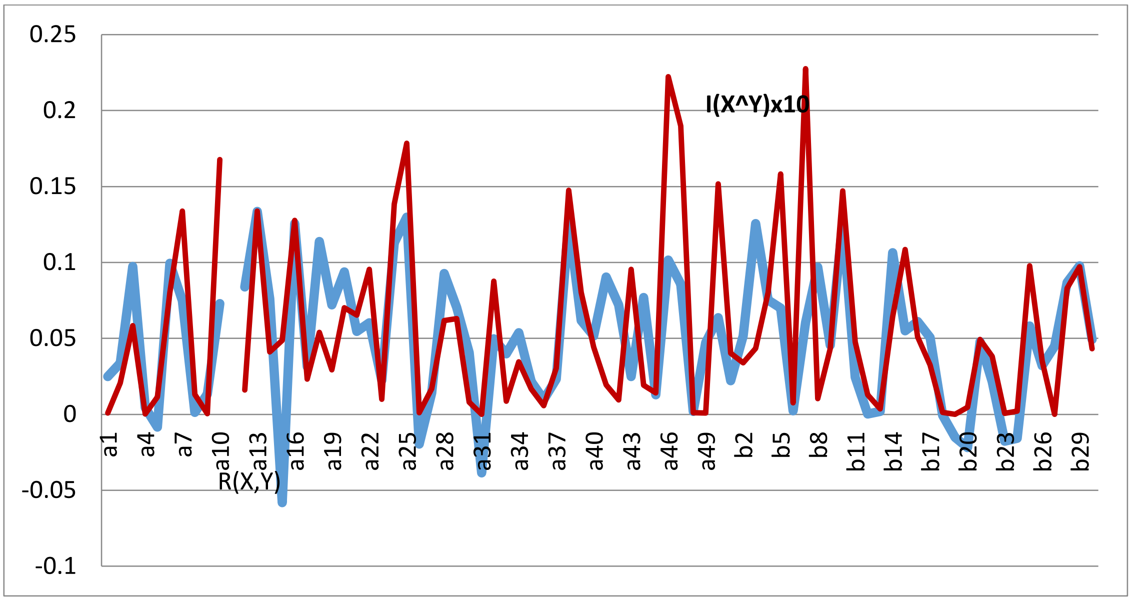

Fig. 5. Correlation $\left(R(X_i , X_j)\right)$ and average mutual information $\left(I(X_i \wedge X_j)\right)$ between A11 item and other 80 items.

Table 2.

A11	A5	A8	A15	A26	A31	B18	B19	B20	B23
R(X,Y)	-0.0083	-0.0186	- 0.0581	-0.0195	-0.0385	-0.0005	-0.0151	- 0.0219	-0.0177
I(X∧Y)	0.0011	0.0013	0.0049	0.0001	0.00001	0.0001	0.000004	0.0004	0.00006

For item A11 the negative correlation with other items and average mutual information values are presented in Table 2. When we compare items' correlations and average mutual information graphics, it is easy to see that the results are comparable. For example, from the correlation matrix and graphics shown in Figure 5 we can see that A11 test item has negative correlation with other test items, for these items the values of mutual information are presented in Table2. If correlation values are negative, average mutual information values are small enough, but from test we should remove those items, so the smallest permissible limit value of average mutual information should be 0.005.

4. Conclusion

In this research while analyzing the data, the following has been determined.

1. The methods suggested by us are correctly defining the quality of test items and the results are comparable with CTT and IRT estimation methods.
2. Simpler mathematical analysis is needed compared to IRT.
3. $I\left(X_i \wedge X_j\right)$ describes the dependence of test items which does not have an equivalent in IRT.
4. For test good items $H(X_i)$ values should be between 0.8 and 1.0, for fairly good items (easy) values are between 0.45 and 0.8, and for bad test items $H(X_i)$ values are between 0 and 0.45.

5. For $I\left(X_i \ \wedge \ X_j \ \right)$ preferable are the values smaller than 0.05 and greater than 0.005

$$0.005 \leq I\left(X_i \ \wedge \ X_j \ \right) \leq 0.05$$

6. Smaller sample sizes are required in comparison with CTT and IRT. The sample size should be more than 100. In CTT the sample size is between 200 and 500, and in IRT it depends on the IRT model, but samples over 500 are needed.

References

[1] M. Haroutunian and V. Avetisyan, "New approach for test quality evaluation based on Shannon Information measures", *Transactions of IPIA of NAS RA, Mathematical Problems of Computer Science*, vol. 44, pp. 7-21, 2015.

[2] M. B. Chelishkova, *Theory and practice of pedagogical tests constructing*, Moscow: Logos, 2002.

[3] C. DeMars, *Item Response Theory*. Oxford University Press; 1 edition, 2010.

[4] K. Hambleton and W. Jones, "Comparison of classical test theory and item response theory and their applications to test development", *Educational Measurement: Issues andPractice*, vol. 12, no. 3, pp. 38-47, 1993.

[5] M. Haroutunian and V. Avetisyan, "Development of the test quality evaluation system", *Proceedings of the International Conference on Computer Science and Information Technologies (CSIT 2015)*, Yerevan, Armenia, September 28-October 2, pp. 372--375, 2015.

Submitted 10.10.2015, accepted 20.01.2016

Թեստի որակի գնահատման նոր մոտեցման փորձարկումների վերլուծություն

Մ. Հարությունյան,Վ. Ավետիսյան

Ամփոփում

Նախորդ հոդվածում [1] հեղինակների կողմից առաջարկվել է թեստերի որակի գնահատման նոր մոդել՝ հիմնված Շենոնի էնտրոպիայի և միջին փոխադարձ ինֆորմացիայի վրա: Այս մեծությունների սահմանային արժեքները և թեստավորման մասնակիցների բավարար քանակը որոշելու համար կատարվել են փորձարկումներ: Հոդվածում ներկայացված է այդ փորձարկումների վերլուծությունը, որից հետևում է, որ հաշվարկները ավելի պարզ են IRT-ի համեմատությամբ, թեստավորման արդյունքների վերլուծության համար պահանջվում է մասնակիցների ավելի քիչ քանակ, քան՝ CTT-ում և IRT-ում:

Анализ экспериментов нового подхода для оценки качества теста

М. Арутюнян, В. Аветисян

Аннотация

В предыдущей статье [1] авторами была предложена новая модель оценки качества теста на основе энтропии Шэннона и средней взаимной информации. Для того, чтобы установить практические пределы этих величин и необходимое количество экзаменуемых, были проведены эксперименты. В данной статье представлен анализ этих экспериментов, из которого следует, что расчеты более просты по сравнению с IRT, для анализа результатов тестирований требуется меньшее количество экзаменуемых, чем в CTT и IRT.