Development of the Test Quality Evaluation System

Mariam Haroutunian
Institute for Informatics and Automation Problems, NAS RA
Yerevan, Armenia
e-mail: armar@ipia.sci.am

Varazdat Avetisyan
Institute for Informatics and Automation Problems, NAS RA
Yerevan, Armenia
e-mail: avetvarazdat@gmail.com

ABSTRACT
To provide qualified test or test items it is very important to analyze the pilot testing results and evaluate the test quality features. Typically the test analyzing process is based on the special test theories, for example, Classical Test Theory (CTT) and mathematical Item Response Theory (IRT) [1] are most widely used. Commonly available test theories have complex mathematical – statistical apparatus that makes their usage impractical for other specialists. The way out of the situation is to create a system that will evaluate the quality of the test. There are many such software packages mainly in English, but for the Armenian market there is not a similar system with Armenian language user interface. To develop such a system a research in the field of the similar systems has been carried out [2] and the advantages and disadvantages have been found out. In this paper the description of the new developed system of the tests quality analysis that has a number of advantages over other similar systems is provided.

Keywords
IRT, CTT, test's quality criteria, test reliability, test item difficulty, IRT a, b, c parameters, Test information function, Item response function, IRT programs, Java psychometrics API

1. INTRODUCTION
Recently, testing technologies have become more prevalent in education. A test method of checking and evaluating the knowledge is one of the most reliable and promising ways to increase educational process efficiency. Nowadays testing technologies are used for school graduation and university entrance examinations in Armenia. Also many Armenian universities have been implementing a practice of testing as one of the main tools for both intermediate and final evaluation of learning outcomes.

Testing method has a number of advantages over the other ways of knowledge assessment: higher objectivity, higher fairness, more complete coverage of the educational material, higher accuracy of estimation, higher economical efficiency, relatively little time spent on assessing procedures. [3] [4].

Testing method efficiency depends not only on the application of objective and reliable technology but also on the quality of applied test [5]. Based on this fact, the problem of providing theory for test quality evaluation becomes very important and modern. Test theories are important to the practice of educational and psychological measurement because they provide a framework for considering issues and addressing technical problems. One of the most important issues is the handling of measurement errors. A test theory can also provide a frame of reference for doing test design work or solving other practical problems. Nowadays two theories [1] of tests are widely used: Classical Test Theory (CTT) [4, 5] and mathematical Item Response Theory (IRT) [6].

CTT is a theory about test scores that introduces three concepts - test score, true score, and error score. The founder of CTT is considered to be British famous psychologist Charles Edward Spearman (1863-1945). R.Cattell and D.Wechsler were his students. A.Anastasi, J. P. Guilford, P.Vernon, C.Burt, A.Jensen. are considered to be his followers. Louis Guttman (1916-1987) has his great contribution in the development process of CTT. The classical theory of comprensive tests was first presented in H. Gulliksen's (1950.) work. The classical theory of tests is presented in L. Crocker J. Aligna's book [7] in a modern way. In Russia one of the first introducers of this theory is V. Avanesov [3]. In the work by M. Chelishkova [4] information about statistical methods of a test~~'s~~ quality assessment is presented.

CTT enables to estimate features of test task such as reliability, difficulty, dispersion test marks, criterion-related validity, correlation coefficients and distinguishing ability and so on based on statistical formulas [8].

Classical test models have the advantages of being based on relatively weak assumptions (i.e., they are easy to meet in real test data) and having a long track record. On the other hand, both person parameters (i.e., true scores) and item parameters (i.e., item difficulty and item discrimination) are dependent on the test and the examinee sample, respectively, and these dependencies can limit the utility of the person and item statistics in practical test development work and complicate any analyses.

IRT is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test [6]. Today, IRT is used commonly by the largest testing companies in the United States and Europe for design of tests, test assembly, test scaling anti calibration, construction and investigations of test item banks and other common procedures in the test development process. Within the general IRT framework, many models have been formulated and applied to real test data. IRT one parameter model is suggested by G. Rasch [9]. The improved variants of IRT one parameter model are considered to be two and three parameter models suggested by Birnbaum [10]. D. Andrich [11] and B. Wright [12] have greatly contributed to IRT theory development .

IRT primarily focuses on the item-level information in contrast to the CTT's primary focus on test-level information. IRT main advantage is that items' difficulty coefficients' assessment does not depend on the selection of a certain group of examinees taking the test. Some of the flexibility of IRT arises because the models link item responses to ability, and item statistics are reported on the same scale as ability. This is not the case in classical test theory. As a result the qualitative data are analyzed by means of quantitative methods. It is possible to decide the test item information function through IRT.

In IRT the measurements are implemented based on the following models [13]:

Unidimensional Dichotomous Models

- Normal Ogive Model

- One-Parameter Logistic Model (Rasch Model)
- Two-Parameter Logistic Model
- Three-Parameter Logistic Model
- 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.

IRT models are widely applied not only in the field of education but also psychology, medicine, sociology. As a result, computer programs of making analysis through the theory of IRT are widely used.

An awareness of the shortcomings of classical test theory and the potential benefits offered by item response theory has led some measurement practitioners to opt to work within an item response theory framework. The reason for this change of emphasis by the psychometric and measurement community from classical to item response models is as a consequence of the benefits obtained through the application of item response models to measurement problems. These benefits include:

- Item statistics that are independent of the groups from which they were estimated.
- Scores describing examinee proficiency that are not dependent on test difficulty.
- Test models that provide a basis for matching test items to ability levels.
- Test models that do not require strict parallel tests for assessing reliability.

Benefits obtainable through the application of classical test models to measurement problems include:

- Smaller sample sizes required for analyses (a particularly valuable advantage for field testing).
- Simpler mathematical analyses compared to item response theory.
- Model parameter estimation is conceptually straightforward.
- Analyses do not require strict goodness-of-fit studies to ensure a good fit of model to the test data.

Thus, for statistical analysis of tests it is necessary to apply some systems, software packages which will make some test results' analysis and qualitative features' assessment based on one or both test theories.

To develop a quality examination system of tests in the Armenian language and for the Armenian market some research has been done in the field of similar systems. In the research the peculiarities and advantages of the similar systems have been investigated [2]. A number of computer programs for simulating IRT data have been developed since the early 1970s. However, most of them were developed in the DOS environment (e.g., Bigsteps, Facets, GENIRV, RESCEN) [2]. As a result, these programs are limited today because of inherent problems in DOS: (1) slow performance speed (16-bit), (2) limited usable system resources, (3) incompatibility with recent 32-bit Windows-based OSs, and (4) not a user-friendly interface. Nowadays windows based on IRT programs with user-friendly interface are widely used (e.g., CITAS, Iteman 4, Xcalibre4, Winsteps, Facets, jMetrik, RUMM 2030, ACER ConQuest, IRTPRO, ConstructMap) [2]. Some of the disadvantages of the modern widely-known programs may be emphasized.

- The programs making the analysis through IRT are multifunctional and are applied to assess different measurements. To make an analysis connected with testing process it is necessary to find, take out and sort the test models of the program, which is not an easy task at all.
- Available systems are mainly in English. Very rarely they can be in Russian as well.
- They have complex mathematical apparatus, which is used not only for making test analysis. For pedagogues it is very difficult to comprehend the different features of the apparatus.
- The test analysis results are mainly received in the form of different tables, which are kept in txt formats. The graphics, in their turn, are received in the form of separate files, in jpg or png formats. So, in order to receive a report in the form of one file it is necessary to make edits in different files and receive a new report, which is more applicable for the pedagogue.
- There is no detailed description of the quality features, which are being assessed. There are no methodological instructions on quality features' change.

So, the issue of having such a system for the Armenian market came forward. The new system requirements were to:

- implement the test quality analysis based on CTT and IRT,
- have the peculiarities which are typical to similar systems,
- be in Armenian language,
- have very simple and available interface convenient for pedagogues,
- present results in the form of a report in one file,
- give the detailed description of assessed quality features,
- provide methodological instructions to change the value of this or that feature.

2. SOFTWARE ARCHITECTURE

The software uses both test theories for data analyzes, and consists of three main modules:

- User Interface,
- Module for analyzing test results based on mathematical statistical procedures,
- Final results report generating module.

The software provides possibilities to analyze the test based on two types of the matrixes [5]: dichotomous (two categories), such as right or wrong, yes or no, agree or disagree, and polytomous (more than two categories), such as a rating from an expert.

From the program user interface a user can choose preferable test theory or both of them, appropriate test quality characteristics, and also report format of results' analysis. The following characteristics can be calculated in the scope of the CTT [3] :

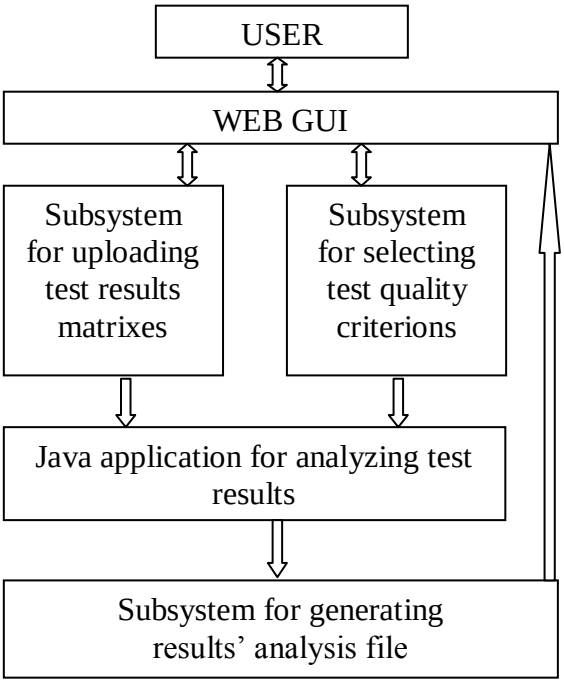
- A number of reliability coefficients,
- Test item difficulty,
- Discrimination ability,
- Dispersion test marks,
- Criterion-related validity,
- Standard error of measurement,
- Some correlation coefficients.

The following characteristics and graphics can be generated in the scope of the IRT [6]:

- Test item discrimination ("a" parameter),
- Test item difficulty ("b" parameter),

- IRT “c” parameter,
- Item options statistics,
- Conditional standard error of measurement (CSEM) function,
- Test information function (TIF),
- Item response function (IRF),
- Item information function (IIF).

The software architecture and use case flow is presented in the picture1.



Picture 1. Software architecture and use case flow

The user interface is web-based and provides the following feautures:

- User can upload the tests results matrices file in the txt, excel or csv formats,
- User can upload the test control file with the above format,
- User can choose appropriate quality criteria as well as other options that control the content of the output report.

To assist a user, the user manual is available in the web interface with detailed description about testing procedures, test quality criteria’s and measurement methods in Armenian language.

From the web interface a user can download example data files and templates as well. The example files are containing examinees and test items simple data for user reference. User can use these files to examine the software overall functionality and get familiar with the software available features and reports content.

The Java application for analyzing test results gets the test results as input data and calculates appropriate tests quality characteristics. It is completely separated from the graphical user interface and the test’s result’s database. It is based on the publicly available Java psychometrics API library [14], and can be downloaded from http://java.net/projects/psychometrics.

The library includes mathematical and statistical procedures that are not the part of the Apache commons math library [15].

It provides several classes for IRT parameter estimation, scale linking, and score equating. The estimation currently involves the joint maximum likelihood for the Rasch, the partial credit, and the rating scale models. The marginal maximum likelihood estimation procedures for the binary item response models (Rasch, 2PL, 3PL, 4PL) and the polytomous item response models (GPCM, PCM) are also available. Scale linking and score equating classes support a variety of item response models. Scale linking procedures are available in the library including the Stocking-Lord and Haebara procedures.

The library includes classes for classical test scaling methods, reliability estimation, item analysis, and differential item functioning (DIF). Examples of scaling methods include normalized scores and Kelley's regressed score. Reliability methods include Coefficient alpha Guttman's lambda, and other methods. There are classes to support the conditional standard error of measurement and decision consistency indices. Classes that support DIF include the Cochran-Mantel-Haenszel procedure and ETS DIF classification levels.

Reporting module generates report file in rich text format (rtf) based on the user input options and the program calculated characteristics. The report is final and well formatted with embedded tables and graphics. It contains sufficient information and prevents user to spend time on the further manual editing. The free Java chart library JFreeChart API [16] is used for graphics and charts generation.

It supports a wide range of chart types giving the report very professional look.

The report file consists of four sections:

- Specifications,
- Summary statistics,
- Item-by-item results,
- Recommendations.

The “specifications” section contains basic information about the analysis. Based on the test result’s file the section includes number of parameters such as number of examinees, total items and as well as parameters concerning IRT and CTT calibration such as item correlation, a, b, c parameter’s minimum and maximum values. The tables and graphics presenting the summary statistics of the test are in the “summary statistics” section. The test reliability indexes, minimum and maximum Score, Mean P value, SM values, Frequency Distribution for the P values, Rpbis, CSEM, Histogram of IRT a, b, c parameter’s, Test Information Function, Test Response Function are included in this section too.

Item-by-item results of the analysis are presented in the third section. IRT parameters table is presented for each item which includes a,b,c parameters values and standard errors (SE). Cronbach's alpha, P value and the point-biserial correlation values are presented in the Classical statistics table for each item. The options’ statisitics are also included in this section. Graphics that include the item response function (IRF), the item information function (IIF) and numerous frequency distributions are generated for each item. In addition, in each section for used parameters, tables, and graphics are explanations and tables of the valid values. The methodological instructions to change the value of this or that feature are provided in the “recommendations” section.

3. SOFTWARE IMPLEMENTATION

During the software implementation process the following main requirements are taken into account to select appropriate programming languages and technologies:

- System should have a user-friendly interface, and all the complexity of the mathematical and statistical apparatus should be hidden from a user.
- System should be platform-independent allowing a user to run it in any kind of platforms, such as Windows, Linux and MacOS.
- Ideally the user interface should be web-based allowing accessing it through the internet or a local network publicly.

The distributed software architecture is chosen to achieve such a functionality satisfying the original requirements. The first component of the system is the web-based user interface. This kind of implementation addresses the requirements being platform-independent and accessible through the internet or the local network. HTML, CSS, JQuery programming languages are used to design web based user interface.

The second part of the distributed system is the JAVA application, which gets executed from the server side PHP script. By the PHP script user submitted data and options are passed to the application as arguments.

The PHP script uses the JSON formatted data to transfer test criteria and other options from web interface to the JAVA application. The JSON format is an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs. It is used primarily to transmit data between a server and the web application, as an alternative to XML.

The JAVA application provides mathematical statistical library for calculating test quality characteristics. Core of this application is psychometrics API [14]. As an input data the application can read and parse txt, excel, and csv formats. After data importing the application internally stores it in the Java Apache Derby open source relational database. To generate the final reports the Apache POI API Java library [17] is used. The POI library contains classes and methods to work on all OLE2 Compound documents of MS-Office.

4. CONCLUSIONS

In addition to a number of common peculiarities with the similar systems, the test quality evaluation system developed by us has a number of advantages over them.

- A user interface is web-based and therefore publicly accessible through the local network or internet.
- System is platform-independent and as a result can run in different operation systems.
- User interface and reference documents are supporting Armenian language.
- The generated report is well formatted rtf document with nicely formatted tables and graphical charts.
- Methodological recommendations and detailed description of assessed quality features are included in the report.
- Because JAVA applications are platform-independent the server side program can be used as a standalone application in any platform.
- Due to the simplified user interface the system can be used by wide range of users.

REFERENCES

[1] Ronald K. Hambleton and Russell W. Jones, “Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development”, *Educational Measurement: Issues and Practice*, Volume 12, Issue 3, pages 38–47, 1993.

[2] Avetisyan V., “Survey of software for the test quality analysis”, *International Journal "Information Content and Processing"*, Volume 2, Number 1, pages 82-92, 2015.

[3] Avanesov V.S., *The bases of the scientific organization of pedagogical control in the higher school*, M., 1987.

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

[5] Kim V. S., *Testing of educational achievements*. Ussuriysk: USPI Publishing, 2007.

[6] DeMars Ch., *Item Response Theory (Understanding Statistics: Measurement)*, Oxford University Press; 1 edition, 2010.

[7] Crocker, L., & Algina, J. & Winston. *Introduction to classical and modern test theory*. New York: Holt, Rinehart, 1986

[8] Avetisyan V., "Investigation of knowledge control tests quality characteristics", *Proceedings of Engineering Academy of Armenia*, Volume 11-1, pages 156-163, 2015.

[9] Rasch G., *Probabilistic Models for Some Intelligence and Attainment Tests.-Copenhagen*, Danish Institute of Educational Research, 1960. (Expanded edition, Chicago, The University of Chicago Press, 1980).

[10] Birnbaum A. *Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability*// F.M. Lord and M.R.Novick. Statistical Theories of Mental Test Scores. Readinf Mass.: Addison-Wesly, Ch.17-20. -P.397-479. 1968.

[11] Andrich D., “Understanding resistance to the data-model relationship in Rasch’s paradigm: A reflection for the next generation”, *Journal of Applied Measurement*, 3, 325–359, 2002.

[12] Wright B.D. & Stone M.H. *Best Test Design*. -Chicago, MESA PRESS, -222 p. , 1979.

[13] Wim J. van der Linden, Ronald K. Hambleton. *Handbook of modern item response theory*. New York: Springer-Verlag. 1997.

[14] Psychometrics**:** Java library for psychometric methods (https://java.net/projects/)

[15] Commons Math: The Apache Commons Mathematics Library (http://commons.apache.org/proper/commons-math/)

[16] JFreeChart: A free Java chart library. (http://www.jfree.org/)

[17] Apache POI - the Java API for Microsoft Documents (https://poi.apache.org/)