SURVEY OF SOFTWARE FOR THE TEST QUALITY ANALYSIS

Varazdat Avetisyan

Abstract: *A test method of checking and evaluating the knowledge is one of the most reliable and promising ways to increase educational process efficiency. Anyway, the test method can be efficient only if the test system is provided with a qualified test and test items. The developed test theory has complex mathematical – statistical apparatus which makes its usage impracticable. There are a lot of software packages to contribute to the pilot testing results' analysis and quality features' evaluation of a test. In Armenia such software is missing, which will give advices and suggestions on quality features' improvement. To develop a new system of quality analysis of the tests in Armenian language researches in the field of similar systems have been carried out. In this survey the peculiarities, advantages and disadvantages of the most applicable modern software are discovered.*

Keywords: *Classical Test Theory (CTT), Item Response Theory (IRT), item difficulty, latent parameters*

ACM Classification Keywords: *G.3 Statistical software.*

Introduction

A test method of checking and evaluating the knowledge is one of the most reliable and promising ways to increase educational process efficiency.

Testing method has a number of advantages over the other ways of knowledge assessment: objectivity and independence, provision of the same conditions for all examinees, possibility to assess many students' abilities and analyze the results as well as to test the knowledge on the respective course material [Avanesov, 1989].

Testing method efficiency depends on not only the application of objective and reliable technology but also the quality of applied test items [Chelishkova, 2002; Avanesov, 1989]. Based on this fact, the problem of providing the testing system with reliable and valid tests becomes very important and modern.

The qualitative analysis of tests is based on the test system of correlated features. This system expresses the quality of test items and the entity of algorithms and formulas which calculate and evaluate certain features. Test theories are concerned with these issues. Nowadays two theories of tests are known: Classical Test Theory (CTT) and mathematical-Item Response Theory (IRT) [Kim, 2007].

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 (1986) [Crocker, 1986] in a modern way. In Russia one of the first introducers of this theory is V. Avanesov (1989) [Avanesov, 1989]. In the work by M. Chelishkova (2002) [Chelishkova, 2002] information about statistical methods of a test's quality assessment is presented.

By means of CTT application it is possible to calculate the reliability and criterion-related validity of the test, to evaluate the correspondence between test items and examinees' individual score, connection between test reliability and length, correlation between test items and so on [Kim, 2007].

IRT is foreseen to evaluate the examinees's latent parameteres as well as test items' parameteres [Avanesov, 2007]. In this theory the mathematical models are widely applied. IRT one parameter model is suggested by G. Rasch [Rasch, 1980]. The improved variants of IRT one parameter model are considered to be two and three parameter models suggested by Birnbaum [Birnbaum, 1968]. D. Andrich [Andrich, 2000] and B. Wright [Wright, 1979] 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. Besides, the parameters of the examinee and test items are assessed through the same scale and the measurements of implemented test scores are turned into line measurement. 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 IRTthe measurements are implemented based on the following models [Wim,1997]:

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 known.

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 two 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. The research aim is to discover the peculiarities and advantages of the similar systems. The research of multifunctional and widely-applied tests' qualitative analysis' systems is presented.

Analysis of the most applicable modern software

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 [Bigsteps, 1998], Facets [Facets, 1999], GENIRV [GENIRV, 1989], RESCEN [RESCEN, 1992]). 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 IRT programs with user-friendly interface are widely used.

CITAS: CITAS [CITAS, 2015] (Classical Item and Test Analysis Spreadsheet) is a straightforward Excel Workbook that provides basic analysis of testing results based on classical test theory.The results are such indicators as mean and SD of scores, reliability, SEM, item P values, item point-biserials, and distractor analysis. By means of CITAS it is possible to analyze the results of the test which consists of not more than 50 examinees and 50 dichotomous items. It is also available in OpenOffice.org format and is an ideal tool for pedagogists to organize the test analysis. CITAS is for free and available at www.assess.com website.

Iteman 4: Iteman 4 [ITEMAN 4, 2015] is Windows based software which enables to receive detailed analysis report of tests and test items based on classical test theory. The aim of the report is to use received indicators to assess the qualitative characteristics of the tests.

Testing results and the data of test items are uploaded as .txt or .csv format files and the result of the analysis is received in RTF (word) format as a separate file, which has the possibility of editing.

Main Features are:

- Results' analysis is made in Word format;
- Analysis is made in the form of graphs;
- Ensures analysis in the levels of tests and test items variants;
- Implements calculation of the various coefficients of reliability.

A number of organizations operating in the field of testing use this program to provide their customers with testing results' analysis report. The program is not free and is provided based on several licenses. Its minimum price is $ 495. It is available at www.assess.com website. In the analyzed results, there is no limitation to examinees' number. Maximum test items number is 10.000. It's free of charge and its demonstration variant is available as well. This variant can be used to make an analysis of the tests with not more than 50 examinees and 50 test items.

Xcalibre 4: Xcalibre4 [Xcalibre 4, 2015] is a Windows based software of making test results' analysis based on IRT theory. While analysing the tests through Xcalibre 4, 4 dichotomous and 5 polytomous IRT models are used. Detailed and summarized analysis is given. This kind of analysis includes graphics and tables. This program is used very easily, has a point-and click interface, there is no need to work with programming codes. Testing results and test items' data are uploaded as .txt or .csv format files. Analysis report is received in rich text file (RTF) format, which means that there is no need to develop a report based on analysis results. Graphics include the item response function (IRF), the item information function (IIF), the test information function (TIF), and conditional standard error of measurement (CSEM), and numerous frequency distributions and so on.

Supported IRT Models are:

- 3-parameter dichotomous model (3PL);
- 2-parameter dichotomous model(2PL);
- 1-parameter dichotomous model (1PL);
- Rasch dichotomous model, scaled to items rather than people;
- Rasch rating scale model (RRSM, or RSM);
- Rasch partial credit model (RPCM, or PCM);
- Generalized rating scale model (GRSM);
- Generalized partial credit model (GPCM);
- Samejima's Graded response model (SGRM, or GRM).

One of the feautures peculiarities is that some data of the analysis are received in CSV format as well. This format is widely used in the program working with electronic tables. Its enables to analyze 1500

test items. Xcalibre 4 program is developed and is periodically updated by the organization of **Assessment Systems Corporation**. The program is not free and there are a number of licensed variants. Its minimum price is $ 495. It is available at www.assess.com website and its demonstration variant is available as well. This variant can be used to make an analysis of the tests with not more than 50 examinees and 50 test tasks.

Winsteps: Winsteps is a Windows-based software [Winstep, 2014] by means of which, based on IRT theory, the analysis is made due to Rasch model (Rasch Analysis) [Rasch, 2015].

It is developed by Benjamin Wright and John Michael Linacre [Linacre, 2004] at the University of Chicago in the 1980s. It is applied to analyze teaching tests, public surveys and rating scales. Item's analysis includes dichotomous, multiple-choice (MCQ), Rating Scales (RSM), Partial Credit (PCM) scales each of which has up to 255 categories. The analysis includes the data tables and many charts. The analysis is presented according to different categories for both items and examinees. There are a number of comprehensive guides both in printed and electronic versions. Some of them are available for downloading. The program correlates with Excel, R, SAS, SPSS, STATA, Txt files, which enables to upload the test results in the form of the above mentioned files. The program provides the receiving of about 48 kinds of tables, files and charts. The module of chart presentation has its separate functionality. By means of the chart the different features of the tests, test items, examinees are received. It is always developed by the group of Winsteps. The last version is Winsteps 3.81.0 which is issued in February, 2014. By means of the program the analysis of 1000000 examinees and 30.000 items can be made. The program requires some fee, the price is 149$. There is also the demonstration variant of the program MiniStep [Ministep, 2015], by means of which it is possible to make the analysis of not more than 75 examinees and 25 items. The program is presented at http://www.winsteps.com official website, where a number of descriptions, guides publications on Rasch model's measurements [Winstep, 2014].

Facets: Like Winsteps, Facets [Facets, 2014] was also developed by Mike Linacre. It is foreseen for applications of more complex (Unidimensional) Rasch measurements. It is a powerful and flexible program and includes all the models and possibilities available in Winsteps and provides a many-facet Rasch model [MFRM, 2015] which is not found in Winsteps. Many-facet Rasch model is applied when it is necessary to analyze the results of the experiments where heterogenic different tests and tasks are applied. Facets program enables to present the results of different items flexibly enough. By means of Facets, in the result of analysis, it is possible to receive the results' files with a respective scale. These files are inserted in Winsteps very easily. Facets also requires some fee, the price is 149$. It enables to analyze the data of about 1.000.000 examinees. The free demonstration version is called Minifac [Minifac, 2015] through which it is possible to analyze the responses of up to 2000 examinees. In Facets, during the calculations of much more complex models much time is needed. It is advisable to

apply at least 1GB operative memory. The user guides and Minfac program are available at www.winsteps.com.

jMetrik: jMetrik is a free and open source computer program for psychometric analysis [jMetrik, 2015]. jMetrik is a pure Java application. It runs on Windows, Max OSX, and Linux operating systems. It features a user-friendly interface, integrated database, and a variety of statistical procedures and charts. The test results are inserted separately in .txt, .csv format files and their editing in the form of a table becomes possible. Based on this table the database is created. There is no limitation to the number of examinees and test items. It depends on the computer memory. The test is analyzed based on CTT and IRT theories. It provides a number of models like Rasch, 2PL, 3PL, PCM, RSM, GRM, GPCM as well as their combinations. A number of coefficients of test reliability, correlation, standard deviation of measurements are taken into account. The program provides charts concerning the tests, test items and examinees as well as the editing process of these charts. Analysis results are received in the program environment and can be downloaded in txt. and the charts in JPG, PNG formats. The program, sample files of test results, the guide are available at http://www.itemanalysis.com/ website. The book of Applied Measurement with jMetrik [Meyer, 2014] is available on a payable basis.

RUMM 2030: The RUMM 2030 [RUMM, 2015] is a Windows application which enables to make an analysis based on Rasch model. It is developed by David Andrich [Andrich, 2012], the author of RASCH - Andrich rating model as well as staff members of the laboratory in Australia, Perth. It is applied in case of the tests with both dichotomous and polytomous scales. It provides the receiving of necessary data and charts according to Rasch model. Particularly, this program is applicable in the educational process while studying the theory of measurements with Rasch model. There is a possibility to easily copy and insert the data taken from different electronic tables' formats /Excel or SPSS/. The disadvantage is the license price. It is 700$ for one academic year. It is necessary to participate in the courses for studying.

ACER ConQuest: ACER ConQuest [ACER, 2012] program combines the models of item response and latent regression. It is developed in Australian Council for Educational Research (ACER), by Margaret L. Wu**,** Ray J. Adams**,** M. R. Wilson. The program provides analysis and assessment by means of the following models:

- Rasch's Simple Logistic Model;
- Rating Scale Model;
- Partial Credit Model;
- Ordered Partition Model;
- Linear Logistic Test Model;
- Multifaceted Models;
- Generalized Unidimensional Models;

- Multidimensional Item Response Models;
- Latent Regression Models.

One of the important peculiarities is that it can assess by means of not only one-diamensional but also multi-diamensional IRT model [Briggs, 2003]. Input data may be uploaded immediately in SPSS format. The results may be received in Excel and SPSS formats. ConQuest has both graphical interface and consol interface. It provides the receiving of many colorful, information charts and maps. But because of the fact that here different measurement models and charts are applied, and there is no limitation to data amount, cost is much. ConQuest is run in Windows environment and costs $699. At http://www.acer.edu.au/conquest/ official website a short description of the program is available. The demonstration version is available for download.

IRTPRO: IRTPRO [IRTPRO, 2015] is a Windows application, which was developed by SSI (Scientific Software International). It is foreseen to make an IRT analysis. It provides an analysis through the following models:

- Parameter logistic (1PL);
- Two-parameter logistic (2PL);
- Three-parameter logistic (3PL);
- Graded;
- Generalized Partial Credit;
- Nominal;

In IRTPRO the data are inserted in the formats of .csv, .fixed, .txt, .xls. The inserted data are saved in the form of IRTPRO file (.ssig). In the environment of the program it is possible to receive different colorful charts. It has inclusive user guide, where IRT theory is described. The program requires some fee; some variants of license are available. The minimum price is $495. It is possible to download the demonstration variant of the program for 15 days duration. In this case the following limitations are found:

- The number of test items – 25;
- The number of examinees – 1000;
- The number of measurements- 3.

ConstructMap: ConstructMap program [ConstructMap, 2015] is developed in the centre of The Berkeley Evaluation and Assessment Research (BEAR). Two variants of the program are in use foreseen for Widows and MacOS operation systems. By means of the program it is possible to make an analysis due to both dichotomous assessment of Rasch model application and polytomous assessment of PartialCredit Model and Rating Scale Model's application. Test results may be inserted in the forms of txt and Excel files. There are a number of ready examples in the program packages. It enables to

receive different results of analysis, which include many IRT and CTT coefficients. It is possible to receive the features (TIF, ICC, etc.) of the test, test items and examinees in the form of charts. Analysis results may be saved in the form of .txt and the charts in the form of PNG or JPG files. The program, its description, user guide are available at the official website. ConstructMap is for free. Java environment needed for its installation.

Conclusion

The research shows that the main peculiarity of the qualitative examination of the tests are the list of supported IRT models, the number of examinees taking the tests and the number of items as well as tables, graphics of received data, formats of results' reports and so on.

Winsteps and Facets programs are distinguished due to the diversity and quantity, visual environment possibility to work with received graphics as well as existing multifaceted guides received in the result of analysis. The formats of received results are important as well. From this perspective Iteman 4, Xcalibre 4 programs are particularly distinguished. By means of these programs the test results are presented as full reports in rtf format. The report includes the tables with numbers as well as the graphics and information about the features. Such reports are much more available for the pedagogues. In case of other programs, the data are available in txt format and graphics- in png, jpg formats.

In the most part of the studied programs there is a limitation to the number of examinees and items of the test which is being analyzed. From this perspective jMetrik program is distinguished. In this program there is not such a limitation and based on the data in it a table is demonstrated in the visual environment. This table is kept in the database.

The diversity of supporting the formats of input and received files is an important feature as well. The files of the test results and items' answers are input in txt sometimes in Excel or CSV formats. In a number of programs (Winsteps, Facets, RUMM 2030) there is a possibility to correlate with R, SAS, SPSS, STATA formats' files.

During IRT analysis's implementation the diversity of models supported by a given program is important. Almost in all IRT programs 1PL, 2PL, 3PL models are applied. Acer ConQuest program is distinguished as it enables to analyze through multivariate IRT. Xcalibre 4, WinSteps, ACER programs are known with their supported models' diversity.

The studied programs are mainly based on Windows. The programs running in Linux, MacOS operating systems' environments are very rare.

The most of programs are applied to analyze not only test results. This fact makes the programs' functionality much more complex. Besides, the measured models become diverse.

The studied programs are mainly in English, require some fee, have different licensed packages. The demonstration variants of these programs are available as well. In the most part of such variants the limitations to the functional opportunities are found.

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;
- They mainly have multi-functionality and have appreciable values;

So, the issue of having such a system for the Armenian market comes forward. The new system requirements are 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.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[ACER, 2012] ACER ConQuest 3.0.1 computer program-http://www.acer.edu.au/conquest

[Andrich, 2000] Andrich D., SheridanB., Lyne A. & LuoG. RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models (Perth: Murdoch University), 2000.

[Andrich, 2012] Andrich D., Sheridan, B.S., & Luo, G. (2012). Rumm 2030: Rasch Unidimensional Measurement Models (software). RUMM Laboratory Perth, Western Australia.

[Avanesov, 1989] Avanesov V.S., The bases of the scientific organization of pedagogical control in the higher school (M., 1987)

[Avanesov, 2007] Avanesov V. S. Item Response Theory: The basic concepts and propositions. 2007. (Russian). (http://testolog.narod.ru/Theory67.html).

[Bigsteps, 1998] Bigsteps-DOS precursor to WINSTEPS. Final Version: 2.82, December 1998- Retrieved February 25, 2015, from http://www.winsteps.com/bigsteps.htm

[Birnbaum, 1968] 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, 1968. -Ch.17-20. -P.397-479.

[Briggs, 2003] Briggs D. C., & Wilson M. R. (2003). An Introduction to Multidimensional Measurement using Rasch Models. 4(1), 87-100.

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

[CITAS, 2015] CITAS - free item analysis with classical test theory, Assessment Systems Corporation- http://www.assess.com/xcart/product.php?productid=407&cat=25&page=1

[ConstructMap, 2015] ConstructMap software- http://bearcenter.berkeley.edu/software/constructmap

[Crocker, 1986] Crocker Linda, Algina James. Introduction to Classical and Modern Test Theory. –New-York: Harcourt Brace Jovanovich, 1986.

[Facets, 1999] Facets: DOS precursor to the current Windows-based Facets. Final Version: 3.22, October 1999. Retrieved February 25, 2015, from http://www.winsteps.com/facdos.htm

[Facets, 2014] John M. Linacre. A User's Guide to FACETS Rasch-Model Computer Programs http://www.winsteps.com/facetman/index.htm

[GENIRV, 1989] Baker F. B. (1989). GENIRV: A program to generate item response vectors (Unpublished manuscript). Madison, WI: University of Wisconsin, Laboratory of Experimental Design.

[IRTPRO, 2015] IRTPRO 2.1 for Windows by Li Cai, David Thissen & Stephen du Toit- http://www.ssicentral.com/irt/

[Iteman 4, 2015] Iteman 4 - Test and item analysis software with classical test theory, Assessment Systems Corporation- http://www.assess.com/xcart/product.php?productid=417&cat=25&page=1

[jMetrik, 2015] Metrik-computer program for psychometric analysis. Retrieved February, 2015, from http://www.jmetrik.com/index.php.

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

[Linacre, 2004] From Microscale to Winsteps: 20 years of Rasch Software development, Linacre J.M. … Rasch Measurement Transactions, 2004, 17:4 p.958- http://www.rasch.org/rmt/rmt174g.htm

[Linacre, 2015] Winsteps and Facets Comparison. In Winsteps and Facets Rasch Software. Retrieved February, 2015, from http://www.winsteps.com/winfac.htm.

[Meyer, 2014] J. Patrick Meyer. Applied Measurement with Metrik. Routledge – 2014 http://www.routledge.com/books/details/9780415531979/

[MFRM, 2015] John Michael Linacre. Brief Explanation of the theory behind Many-Facets Rasch Measurement (MFRM). Help for Facets Rasch Measurement Software: http://www.winsteps.com/facetman/theory.htm

[Minifac, 2015] MINIFAC- Evaluation, Student and Demonstration (Demo) Version of FACETS (http://www.winsteps.com/minifac.htm)

[Ministep, 2015] Ministep-Evaluation, Student and Demonstration (Demo) Version of WINSTEPS (http://www.winsteps.com/ministep.htm)

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

[Rasch, 2015] Rasch Analysis - http://www.rasch-analysis.com/index.htm

[RESCEN, 1992] Muraki E. (1992). RESGEN: Item response generator [computer program]. Princeton, NJ: Educational Testing Service.

[RUMM, 2015] RUMM for analyzing assessment and attitude questionnaire data -http://www.rummlab.com/

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

[Winstep, 2014] John M. Linacre- A User's Guide to W I N S T E P S® M I N I S T E P Rasch-Model Computer Programs- http://www.winsteps.com/winman/index.htm

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

[Xcalibre 4, 2015] Xcalibre 4 - Software for IRT analysis and calibration, Assessment Systems Corporation- http://www.assess.com/xcart/product.php?productid=415&cat=22&page=1

Authors' Information

Varazdat Avetisyan– *Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, PHD student, P.O. Box: P. Sevak 1, 0014 Yerevan, Armenia; e-mail: avetvarazdat@gmail.com*

Major Fields of Scientific Research: Test theory, e-learning, web programming, probability and statistics