Participatory Educational Research (PER)
Special Issue 2015-II, pp., 110-116; 5-7 November, 2015
Available online at http://www.partedres.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.15.spi.2.13

Computer Aided Analysis of Multiple Choice Test Results

Ertuğrul ERGÜN* and Ali AYDIN
Distance Education Vocational School, Afyon Kocatepe University, Afyonkarahisar, Turkey

Abstract
One of the most widely used assessment technique in educational institutions are the multiple-choice tests. Several analyses have to be made in order to determine the validity and reliability of these multiple-choice tests and items in the test. In order to make some comments about multiple choice tests, test's average, test's reliability, mean difficulty, standard deviance, measures of central tendency, measures of central distribution should be computed. And also to make some comments about multiple choice tests' items, Item Difficulty index, Item Discrimination Index, item variance and standard deviance, item reliability index should be computed. These computations are time-consuming and hard to do by hand. Also even if data may be entered in a spreadsheet, formulas can be hard for a teacher to form in the software. To make comments about the produced values is also a hard point for educators. As a result, teachers in educational systems don't/can't do evaluations about the assessments they applied. In this study, a software has been developed for the statistical evaluation of multiple-choice tests' results. With this software, test and item analysis of the multiple-choice exam can be done and also statistical results can be presented to the user by colorized graphics. Examinees' scores, frequency table and analyses about the test (range, mean, median, Kr20, test's mean difficulty, standard deviance, variance, coefficient of variation, and coefficient of skewness), every item's Item Difficulty index, Item Discrimination Index, item variance and standard deviance, item reliability index, Point-Biserial Correlations are the main outputs of the software. Also distracters in choices can be seen easily in the graphics section. Also there is an info box in the developed software. The info box shows several information about the computed properties and their values. This box can be helpful for users who have limited information about these statistics.

Keywords: test analysis, item analysis. assessment and evaluation, multiple choice tests

Introduction

Measurements and assessments play an important role in the evaluation of the education. Evaluations made during teaching process can provide feedback, which can increase teaching efficiency by determining learning deficiencies and failing points in the process. At the end of the teaching process, assessments can be used to learn, if there has been a change in knowledge of students. At the end of the teaching process, evaluation can be used to judge whether the program or instruction has met its intended learning outcomes.

Evaluation of learning and teaching in a curriculum with examinations is important for

* ertue@aku.edu.tr

education. It can assess the effect of a teaching program and the levels of knowledge absorbed by students.

Testing and evaluation done by teachers in classrooms can provide feedback to the teachers related to the mastery level of the students on a skill that has been touched in the classroom, and also observe the problems that arise in the teaching sessions. With that, a teacher can find out the level of improvement of a student in a classroom on whether the student is in the "very satisfactory", "moderate", "poor" or "no improvement whatsoever" category. From the evaluation done, the teachers can also determine active students who need enrichment and also the weaker students who need enrichment activity. The teacher will also make a decision on whether to change the strategy of teaching so that it is more suitable with the students' needs or repeat same strategies or not (Hamzah and Abdullah, 2011).

One of the most widely used assessment technique in educational institutions is the multiple-choice tests. In Turkey these tests are widely used in schools and also Student Selection and Placement Center (OSYM) and Ministry of National Education use these tests in nationwide exams.

Kuran and Kanatlı (2009) stated that over 80.8% of teachers use multiple choice tests in assessment. Other techniques that used were: short answer questions (66.7%), true-false statements (64.3%), essays (60%), matching method (50.6%).

Çelikkaya et al (2010) found that the most used assessment technique used by social sciences teachers' are multiple choice tests. Almost 100% of the teacher's used multiple choice tests in the assessment process. 71.1% of teachers had no problems regarding this kind of technique, but also 19.2% of teachers expressed that item (question) preparation is hard and time consuming.

Xu and Liu (2009) stated that the teachers' knowledge in assessment and evaluation is not a static process but rather a complex, dynamic, and ongoing activity.

Swanson et al (2005) stated that multiple choice questions are globally the most utilized application among different types of students learning achievements and progress.

Çakan (2004) found that most of the teachers perceived themselves as unqualified in terms of measurement and evaluation applications. On the other hand, compare to secondary school teachers, elementary school teachers perceived themselves more qualified. Although most of elementary school teachers use multiple choice items most frequently, secondary school teachers prefer using essay tests most often than any other item type (Çakan, 2004).

In a research to determine teachers' perceived levels of efficacy towards measurement and evaluation, it has been stated that levels of perceived efficacy of prospective teachers on measurement and evaluation were appeared to be low. (Yaman & Karamustafaoğlu, 2011).

Test and Item Analysis

The process of testing usually begins with the preparation stage, followed by the implementation (test administration) and ends with the answer script inspection. Through this testing process, a teacher can understand whether or not his/her students have mastered the skills learnt (Hamzah and Abdullah, 2011).

Item analysis is the process by which test items are examined critically. Its purpose is to identify and reduce the sources of error in measurement (Osterlind, 2002).

In order to assure the validity and reliability of an examination, items in an examination should be subject to thorough investigation with some psychometric methods (Yang et al. ,2011).

In the standardized and objective evaluation of student performances, the item analysis is a process in which both students' answers and test questions are examined in order to assess the quality and quantity of the items and the test as a whole (Siri & Freddano, 2011).

Several analyses have to be made in order to determine the validity and reliability of these multiple-choice tests and items in the test. Anastasi (1997) stated that the validity of a test concerns what the test measures and how well it does so. And, Osterlind (2002) stated that test validation is the process of gathering evidence for a specific interpretation of the scores yielded by a given test.

Teachers should routinely perform item analysis so that they may gauge the quality of items and discard those which are unacceptable, repair those which can be improved, and retain those which meet criteria of merit.

The items that constitute a test can have different characteristics. The answering ratio of these items, the group in which they are answered correctly at a higher rate, and their difficulty and discrimination level can all be identified through evaluations performed at an item-level (Tomak and Bek, 2015).

In order to make some comments about multiple choice tests, test's average, test's reliability, mean difficulty, standard deviance, measures of central tendency, measures of central distribution should be computed. And also to make some comments about multiple choice tests' items, Item Difficulty index, Item Discrimination Index, item variance and standard deviance, item reliability index should be computed.

Sometimes it is useful to compare subgroups of the examinee population to determine how an item is performing. For this analysis, the population is often divided into two groups, a high-achieving group and a low achieving group. Typically, the groups are examinees whose total score on a test comprise the top 27 percent of all examinees, and those whose scores place them in the bottom 27 percent of the examinees. The figure 27 percent is chosen because it is used in some computational algorithms for determining internal reliability indices and Kelly (1939) demonstrated that this number will provide a stable index of differences between high and low ability groups. For this analysis, the principal focus is on determining how well the item is functioning for the extremes of the ability range (Osterlind, 2002).

These computations are time-consuming and hard to do by hand. Also even if data may be entered in a spreadsheet, formulas can be hard for a teacher to form in the software. To make comments about the produced values is also a hard point for educators. As a result, teachers in educational systems don't/can't do evaluations about the assessment systems.

Yang et al. (2011) used Rasch model to get valuable information related to test reliability, item difficulty and examinee ability in an examination in anesthesiology for medical students. They found that the test reliability was an unsatisfactory 0.63, which means that the test results were not so reliable and also they stated that the examination was relatively easy for

most of the students. To improve the test reliability, it was advised to increase item numbers and to enhance the discrimination of the test, item difficulty should be adjusted to promote usefulness of the exam.

Siri and Freddano (2011) investigated the effect of the analysis of multiple choice questions designed by the teachers on the quality of the tests. After the administration of the test they computed facility index and the selectivity index to analyze the items. They stated that item analyses should be utilized to improve already existing tests instead of developing new items to avoid wastage in time.

Tomak and Bek (2015) compared the classical and the latent class models used in item analysis, as well as their efficacy in the evaluation of the examinations of the medical faculty. They obtained similar results by classical and latent methods. They stated that classical theory is easy to understand and to apply, while Item Response Theory is, on the contrary, sometimes rather difficult to understand and to implement

Yurdugül and Batenburg (2006) applied Graphical Item Analysis to the SSPE-SE (Student Selection and Placement Examination for Secondary Education) in Turkey. They found a linear relation between difficulty values of test items in GIA and other traditional item analysis techniques.

Software Development

In this study, a software has been developed for the statistical evaluation of multiple-choice tests' results. This software is developed in C#, one of the programming languages which is used quite a lot in recent years. In this software, multiple choice exams which were previously applied and results had been saved to computer, can be analyzed.

With this software, test and item analysis of the multiple-choice exam can be done separately, and also statistical results can be presented to the user by colorized graphics. In addition, user can produce and save reports of analyses to evaluate later (Aydın, 2013).

Details of Software

In the main window of the software there is five tabs (Figure 1 – red zone) (Giriş, Ayarlar, Test Analizi, Madde Analizi Grup, Madde Analizi Tüm – Input, Settings, Test Analysis, Item Analysis Group, Item Analysis All).

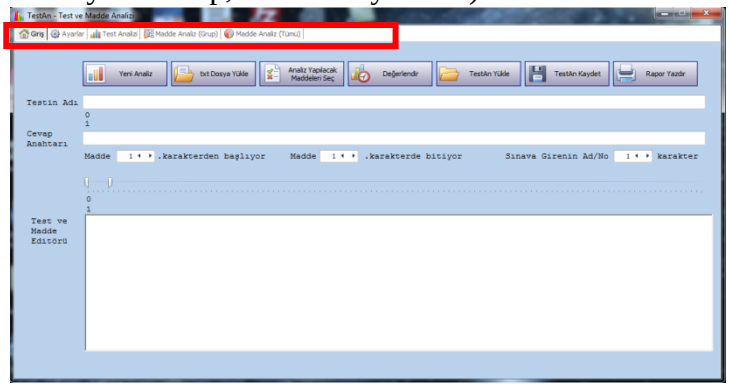

Figure 1. Main window of software

In analyzing process exam data must be entered to the software. In order to easily enter data, data must in a text file which consists of rows which must include student number and student's response to the questions. Then place where answers start and end and examinee name's place must be determined (Figure 2). The key for the exam must be entered in the "Cevap anahtarı" box in this window. Afterwards items which will be examined in the process can be chosen in "Ayarlar-settings" tab or all of the items can be used. "Değerlendir-Calculate" button must be clicked to finish analyze process.

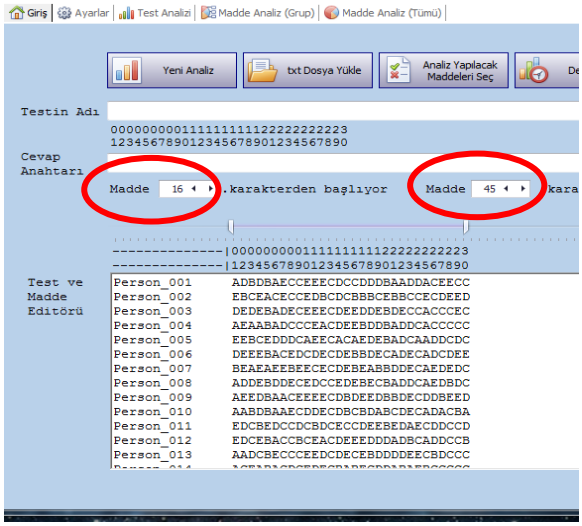

Figure 2. Input stage of test data

Examinees' scores (a in Figure 3), frequency table (b in Figure 3), and analyses about the test (c in Figure 3) (range, mean, median, Kr20, test's mean difficulty, standard deviance, variance, coefficient of variation, and coefficient of skewness) can be seen in "test analizi – test analysis" tab.

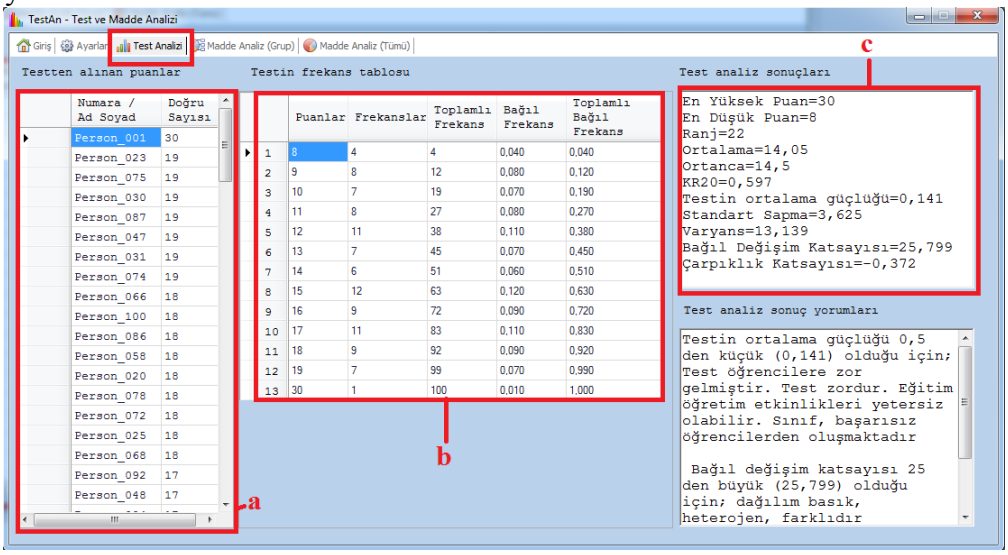

Figure 3. Test analizi- Test Analysis window

Every item's Item Difficulty index (1), Item Discrimination Index (2), item variance (3) and standard deviance, item reliability index (4), Point-Biserial Correlations (5) can be seen in "Madde Analizi Grup – Item Analysis Group" tab (Figure 4). And also distracters can also be seen easily in the graphics section of this tab. In the "a" section of Figure 4, the green zone shows the correct answer. Frequencies of the answers of top and bottom groups can also be

seen in this section. Computations in this window are made using the data of the 27% of the students at the top and the 27% at the bottom according to their total score.

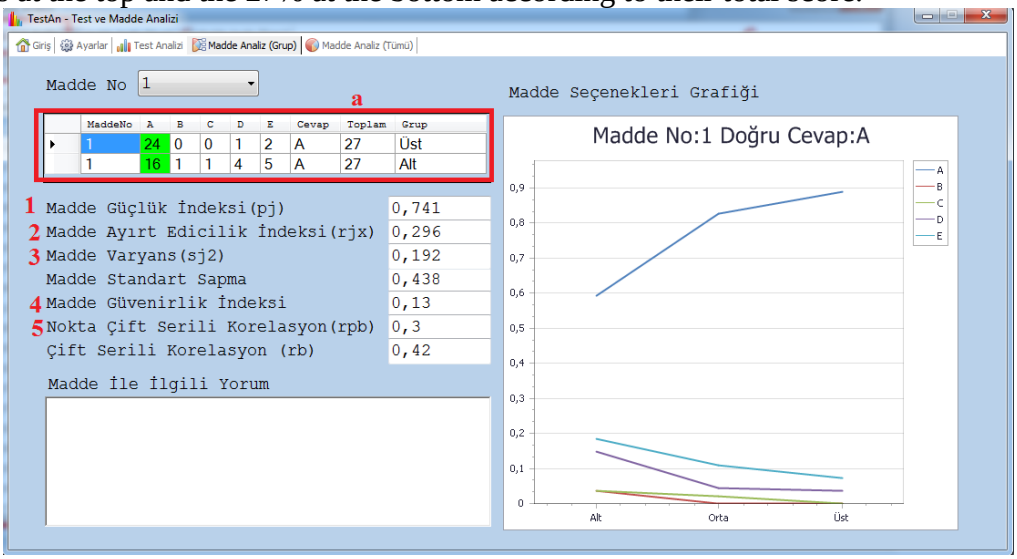

Figure 4. Madde analizi (Grup) - Item Analysis (group) window

In "Madde Analizi Tüm – Item Analysis All" tab, the computations which are made by using all of the examinees' data can be seen (Figure 5).

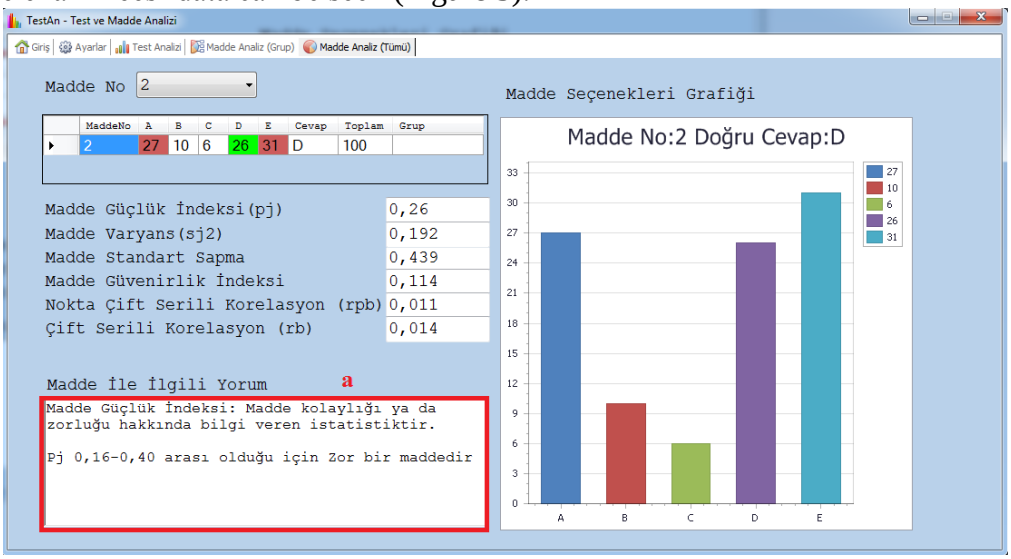

Figure 5. Madde analizi (Tüm) - Item Analysis (All) window

Also there is an info box in this developed software (a section in Figure 5). The info box shows several information about the computed properties of the item and their values. This box can be helpful for users who have limited information about these statistics.

A web site has designed for the software (www.testanalizi.com). Software can be downloaded from this website and used freely.

Conclusions

With this software, educators can easily produce statistical information and detailed item analysis about the multiple choice tests' they used. With this information they can easily

see the accuracy of the assessment and evaluation processes. Also this analysis process shows what must be changed in test as a whole or items in particular.

Software can also be used while teaching test analysis and item analysis in assessment and evaluation courses in universities.

Several developments are being planned for software. Especially generating detailed reports and graphics, generating random test data for analyzing and input problem for exams which has more than one answering group are the main processes that are worked on.

References

Anastasi, A. (1997). *Psychological Testing (7h Ed.).* New York: Macmillan Publishing.

Aydın, A. *Çoktan seçmeli ölçme sonuçlarının bilgisayar yardımıyla analizi.* (Unpublished master's thesis). Afyon Kocatepe University, Afyonkarahisar.

Çakan, M. (2004). Öğretmenlerin Ölçme-Değerlendirme Uygulamaları ve Yeterlik Düzeyleri: İlk ve Ortaöğretim. *Journal of Faculty of Educational Sciences, 37*(2), 99-114.

Çelikkaya, K., Karakuş, U., & Öztürk Demirbaş, Ç. (2010). Sosyal Bilgiler Öğretmenlerinin Ölçme- Değerlendirme Araçlarını Kullanma Düzeyleri ve Karsılastıkları Sorunlar. *Ahi Evran Üniversitesi Eğitim Fakültesi Dergisi, 11*(1), 57-76.

Hamzah, M., & Abdullah, S. (2011). Test Item Analysis: An Educator Professionalism Approach. *US-China Education Review*(3), 207-322.

Kuran, K. (2009). Alternatif Ölçme Değerlendirme Teknikleri Konusunda SInıf Öğretmenlerinin Görüşlerinin Değerlendirilmesi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 6*(12), 209-234.

Osterlind, S. (2002). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats.* New York: Kluwer Academic Publishers.

Siri, A., & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia - Social and Behavioral Sciences, 29*, 188-197.

Swanson, D., Holtzman, K., Clauser, B., & Sawhill, A. (2005). Psychometric characteristics and response times for one-best-answer questions in relation to number and sources of options. *Acad Med*(80), 93-96.

Tomak, L., & Bek, Y. (2015). Item analysis and evaluation in the examinations in the faculty of medicine at Ondokuz Mayis University. *Nigerian Journal of Clinical Practice, 18*(3). doi:10.4103/1119-3077.151720

Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: a narrative inquiry of a chinese college EFL. *TESOL Quarterly, 43*(3), 493-513.

Yaman, S., & Karamustafaoğlu, S. (2011). Investigating prospective teachers' perceived levels of efficacy towards measurement and evaluation. *Journal of Faculty of Educational Sciences, 44*(2), 53-72.

Yang, S., Tsou, M., Chen, E., Chan, K., & Chang, K. (2011). Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model. *Journal of the Chinese Medical Association*(74).

Yurdugül, H., & Van Batenburg, T. (2006). Item Difficulty From Graphical Item Analysis. *Eurasian Journal of Educational Research*(24), 209-218.