Id: 735248

Participatory Educational Research (PER)
Vol. 8(2), pp.147-162, April 2021
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.21.34.8.2

# Development of a Web-Based Environment for Test and Item Statistics

Gözde Tekalmaz
*Master's Degree: Kocaeli, Turkey, Orcid: 0000-0002-2158-5138*

Fatih Kezer*
*Department of Measurement and Evaluation, Kocaeli University, Kocaeli, Turkey; Orcid: 0000-0001-9640-3004*

**Article history**
**Received:**
12.05.2020

**Received in revised form:**
18.11.2020

**Accepted:**
03.12.2020

**Key words:**
Item statistics,
test statistics,
web-based environment

The aim of the research is to develop an online environment where teachers can access test and item statistics and a comprehensible report regarding the tests they have used, to ensure the storage of their reports in their own online environment as well as to have an idea on the utility of the subject matter environment. The research is planned as a design and development research. Research data was collected in two stages. The data which was needed to compare were obtained simulatively through the TAP program. Also, dta was collected from a group of 15 teachers to collect evidence for the usability of the application. Test and item statistics were calculated for the collected data and compared with the results of the programs making similar calculations. Descriptive analysis was used in the interpretation of the usability questionnaire and of teacher opinions. As a result of the research, the item and test statistics calculated for all data sets created by sampling different situations were found to be similar to those of reliable equivalent package programs. With the opinions received from the teachers, it has been concluded that the environment is useful and practical. While one of the important differences of the online environment formed by this study is its keeping a record of the past; another is its accessibility through the internet without the need for an additional program and thirdly, its usability on a mobile platform.

## Introduction

Measurement and evaluation is one of the indispensable parts of education. One of the important stages of educational or training processes is to see to what extent the desired behaviors are gained or whether they are gained. It is through the measurement and evaluation processes to be carried out along with the education that the level of student behavior, the points of deficiencies, if any, and even whether undesirable behaviors occur can be determined (Kutlu, 2003). Knowing the degree of success or failure helps to plan more realistic and substantial training activities for the educational programs to be implemented in

* Correspondency: : fatih.kezer@kocaeli.edu.tr

future (Turgut & Baykul, 2010).

The constructivist approach adopted with the renewed education programs enables the individual to create a unique world view and perspective during the learning period by establishing a link between the new information acquired and the information he / she formerly had (Brooks, 2002). In classical classrooms where the behavioral approach is adopted, measurement and evaluation are generally carried out with open-ended questions, true-false tests, and multiple-choice tests (Turan-Oluk & Ekmekçi, 2017). Multiple choice tests as a measurement tool are frequently preferred in selection and placement exams and in classroom measurements. In cases of multiple choice tests, it is difficult to have detailed information about a student by merely looking at his or her answers. In this case, a question arises: How can the extent to which the information given by the teacher in the classroom is really acquired can be assessed? The answer to this question is the measurement and evaluation processes that use valid and reliable measurement tools. Quality measurement tools are needed to determine the realization level of the achievements expected in the lessons. It can be said that the process of developing a measurement tool is a systematic whole consisting of many stages. These stages can be listed as informing the students beforehand about the type and level of the tool to be applied, creating an item pool, selecting the items to be directed to the students with the tool according to a table of the markers, arranging the measurement tool, applying it on the students and analyzing the items (Kubiszyn & Borich, 1996; Özçelik, 2014). In order to have general information about the tests applied, besides the item analysis, some statistical information about the test is also needed. While statistics such as reliability coefficient, standard deviation, arithmetic mean, mode, and median give information about a test in general or about students' level of success, statistics such as item difficulty index and item discrimination index provide information about each item that exists in a given test. When all these points are aptly taken into consideration with rigor and determination, they can significantly improve the quality of measurement and evaluation processes in education.

The professional competence of teachers is an important issue that directly affects the quality of education. According to the General Competencies Guide for Teaching Profession published by Republic of Turkey Ministry of National Education, one of the most important competencies that teachers should have is pertinent tomeasurement and evaluation. Nevertheless, an important part of the studies showed that teachers' in-class measurement and evaluation knowledge and skills were not sufficient, and that they were below the desired and required level (Bıçak & Çakan, 2004; Daniel & King, 1998; Güven, 2001). Kubiszyn and Borich (1996) listed the subjects of measurement and evaluation regarding which teachers should have knowledge and skills as follows:

- Which test types are suitable for the measurements planned for different purposes,
- To be able to correctly determine what the objectives of the course are,
- Knowledge and skills on how written examination tests should be improved,
- How to ensure reliability and validity of tests, basic test statistics,
- How test scores should be used,
- How student achievements or grades should be communicated to families in an effective and useful way, using an effective communication technique.

In a research conducted on teachers working in primary and secondary education (Daniel & King, 1998), it was found out that they did not have sufficient or comprehensive knowledge on measurement and evaluation. It was also observed that they did not have simple statistical information used in measurement and evaluation. Besides, it was revealed by various studies that teachers experienced problems regarding the measurement and evaluation applications in their programs and many of them considered themselves inadequate in this regard compared to other fields, stating their need for training in measurement and evaluation (Gözütok Akgün & Karacaoğlu 2005; Yapıcı & Demirdelen 2007; Mertler, 2009; Yaşar et al., 2005). Research on teachers' measurement and evaluation competencies (Birgin & Gürbüz, 2008; Gelbal & Kelecioğlu, 2007) revealed that teachers mostly prefer traditional measurement methods, and they need training in the use and preparation of measurement techniques. The insufficiency of teachers in the field of measurement and evaluation and their need for improvement in the use of technology are also emphasized in different studies in the related literature (Bıçak & Çakan, 2004; Güven, 2001). Therefore, considering the content of teachers' measurement and evaluation courses they get during their undergraduate education and their perceptions of technology proficiency, the use of programs such as SPSS, Jmetrik, TAP, R and ITEMAN, which perform all these statistical calculations on the computer, seems to be necessary but not very practical. Programs such as SPSS, Jmetrik and Iteman are package programs that the users can install and use on their computer, which can do the aforementioned statistical calculations and some others and deliver them to the users in the form of a report. The absence of the said package programs in the Turkish language may also be a negative situation for the Turkish users. Together with this, the process that needs to be followed by entering the data (such as student answers, answer key) before the analysis can be quite complicated for teachers in Turkey. When the reports given based on the results of all these analyzes are examined, it is seen that a remarkably high level of competence on measurement and evaluation is needed to know which value corresponds to what kind of situation. Also, while it is not possible to store private retrospective data in Iteman, SPSS and TAP programs, data only up to a certain quota in the Jmetrik package program can be stored for the users. In addition, package programs can mostly be used on computers and their mobile applications have not yet become widespread.

The fact that teachers cannot have better systems to handle measurement as an important component of education is a hindrance to a higher quality education. Hence, an application that can analyse the overall test and all questions individually, providing an easy-to-use, detailed and an understandable report, and that can store the information of all the classes of a teacher for him or her to track their data in their system and present it back when necessary would be handy to meet the needs of teaching staff in Turkey.

When the test and item analyses are correctly evaluated and assessed based on the scores obtained from a test used in-class measurements, they can provide essential information about that test. In light of the data obtained, the quality of the next measurement process can be improved gradually. It can be thought that it is not very practical to calculate the test and item statistics manually by the teacher. Moreover, considering that all the statistical calculations are done in some way (using package programs or using a calculator), it is possible to reach a clear and understandable interpretation only if the person has a required level of competence. On the other hand, it is not possible for the teaching staff to reach an interpreted report and to follow-up class development after any other calculation.

In this study, an alternative system was attempted to be created regarding the test and item analysis needed by teachers in classroom measurements. Thus the research, aimed to develop

an online environment where teachers can access test and item statistics and a comprehesible report after the tests they have applied, to store and track their reports in their own online environment and to determine the usability of the subject matter environment/application.

## Method

Aiming to obtain a more meaningful report by calculating the test and item statistics that need to be run after the measurement activities, this study is a design and development research. Design and development research is a method/means for developing concrete and feasible solutions or products for real-life problems. Richey and Klein (2014) define design and development research as the process of developing products such as tools, teaching materials and as the development process of models used in the development of these products. Regarding those two objectives, design and development research is handled under two categories: Type 1 and Type 2. While the development of products, that is, the discovery of the main product, is the target of the research process in Type 1, the development of models including the stages and processes to enhance these products is the main target of the research in Type 2. Since a design and development study focuses on development, it is usually recommended especially in project studies. It is a research model that is expected to be labor intensive by its nature and it usually involves or may involve a product that may have a potential to obtain a patent. This research has been developed in three stages: (1) In order to be able to demonstrate all aspects of the research, first of all, expert opinion (in order to determine to need) has been taken, and the related literature review has been done, (2) the product development process has been completed and subsequently the application was carried out with simulative data using the resulting product and (3) finally the overall effect of the product, its contribution, strengths and aspects to be developed were determined and revealed.

### *Data*

Research data was collected in two stages. The data needed to compare the data obtained from the developed environment with SPSS and TAP programs was obtained simulatively through the TAP program. Also, data was collected from a group of 15 teachers to collect evidence of the usability of the application.

#### *Simulative Data*

In order to test the reliability of the application at the stage of comparison to the equivalent program outputs, nine different simulative data groups were created through the TAP program in order to generate the outputs of the web based application. While generating simulative data, it was assumed that 3 data groups of 30, 100 and 500 people were given tests with an average difficulty level of 0.40, 0.60 and 0.90 each respectively. Thusly, a total of three difficulty levels were obtained for each data group, allowing 9 different situations to be examined.

#### *Usability Survey*

A questionnaire developed by IBM (Computer System Usability Questionnaire - CSUQ) and adapted to Turkish by Erdinç and Lewis (2013) was used to determine the usability of the developed web-based application.

Usability is a concept that examines the interaction of individuals with their environment, features, limitations, and expresses the use for system design by looking at the data obtained (Alexander, 2003). Coşkunserçe and Dursun (2008) expressed usability as “performing jobs quickly and easily using a product”. Usability Professionals Association (UPA), on the other hand, denotes usability as a "criterion for usability and ease of use of software, hardware or any product" (UPA, 2008). Unlike these definitions, Jacob Nielsen (2003) signified usability as "a quality feature that evaluates how to create a simple user interface". It is observed that the International Organization for Standardization (ISO) has created the ISO 9241 standard regarding usability (ISO, 2008). Based on the ISO 9241 standard, various research on usability have been conducted. According to these studies, there are five criteria that determine usability and, in the literature, and these criteria are described as “5E” namely effectiveness, efficiency, engaging, being error tolerant and easy to learn (Alexander, 2003). Groups ranging from five to fifteen are sufficient for studies that need usability testing to be performed (Nielsen, 2000). When determining the users to take part in these studies, first of all, it is necessary to determine what is targeted with the website and for whom (Lazar, 2001). The usability questionnaire of the research includes questions under 4 dimensions: the system's usefulness, data quality, interface quality and overall satisfaction. Participants were asked to think aloud while answering the questions during the use of the application. Click counts and access times for each user during the use of application were recorded (with representative test data provided to them). This process includes the analysis by the user of the data given to him at the beginning, his accessing the report, saving the report he has reached and viewing the report he has saved.

Of the fifteen teachers who contributed to the research, six teachers teach at secondary school level and nine teach at high school level. Whilst the branches of the teachers who teach at the secondary school level are distributed as two mathematics teachers, three English teachers and one visual arts teacher; the branches of teachers working at high school level are distributed as three mathematics teachers, a chemistry teacher, a physics teacher, a history teacher, a Turkish language and literature teacher, and two English teachers. The application has been revised (changing illustrative and result expressions, etc.) in accordance with the feedback from the teachers and, in its final form, it has been applied to the same group of teachers again and their final opinions have been taken. The group of 15 teachers on whom the usability questionnaire was applied, was determined by the appropriate sampling method. The sampling method where the researcher starts from the most accessible responders until he reaches a group he needs is called the appropriate sampling method (Büyüköztürk et al., 2015).

### *Data Analysis*

Test and item statistics were calculated and compared for the collected data. While making these calculations, both SPSS package program and TAP package program were used and to test the developed online environment was tested. At this point, the aim was to have an idea about the reliability of the application. The reason for using the SPSS package program with the data obtained was to obtain statistics to be compared with the results of the developed environment. Since item analysis was not possible with the SPSS program, the TAP program previously used in obtaining simulative data was resorted to in item analysis. The data obtained in the simulative way were examined to check if they yield the same numerically accurate values or not. In the developed online environment, while making all these analyses, the manual that explains the formulas used by the SPSS program and the statistical formulas in the related literature are used. Descriptive analysis was referred to in

the interpretation of the usability questionnaire and teacher opinions.

## *Improving the Online Environment*

### *1. Design of the Online Environment*

It is thought that such an online environment is needed considering the difficulties teachers meet in classroom applications and the problems experienced by the teachers during the analysis of the items and the tests, which are stated in the related literature, too. Firstly, detailed literature research was carried out and the scope of the need was determined. The statistics to be calculated after reviewing the literature were determined as follows;

• Arithmetic mean, mode, median, range, standard deviation, variance, test mean difficulty, test reliability coefficient, standard error of the test, relative variability coefficient, skewness coefficient, kurtosis coefficient, Z scores, T scores,

• Item difficulty index, item discrimination index, item variance, item standard deviation, item reliability coefficient.

In addition, the design of the report to be submitted online was structured with the expert opinion regarding the content and the way of presenting the information. At the same time, it was mutually concluded that it was important to keep the data obtained retrospectively so that the teachers could follow the development of their classes. In this context, the environment was planned to develop a system where each user could log in with his own e-mail address, and it was designed in such a way that he could store his exams retrospectively in his profile and access this information at a later time at his own discretion. The application, which was designed to be user-friendly, is going to have a user-friendly guide with the same clarity, as work is being carried out to that end.

### *2. Development of the Online Environment*

Technologies such as HTML, CSS, Javascript, Bootstrap, Firebase have been used in the development of the environment. The online environment has been preferred due to such reasons that it does not require installation for different platforms, it can be used easily on devices such as mobile phones, tablets and computers, and it is easy to access. This online environment is written in Javascript programming language (Figure 1) and Google Firebase is used for database management and server operations.

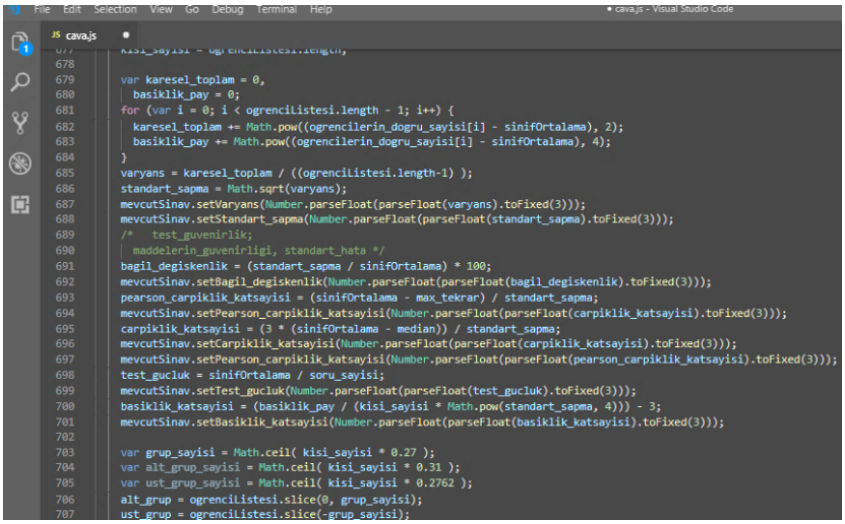


**Figure 1.** Javascript code screen

Users must be members of the system in order to use the application, perform analysis and access a report as an output. After being a member, a verification mail is sent to the e-mail addresses of the users, and after the verification process is completed, the system is opened for the use of members. Once registered, users can use the online environment simply by logging in by virtue of their membership for their subsequent works.

After logging in, users who can easily upload their data into the application (Figure 2) can obtain the analysis result as a report by entering their information such as the answer key, school name, class name, course name and date, without any further action.

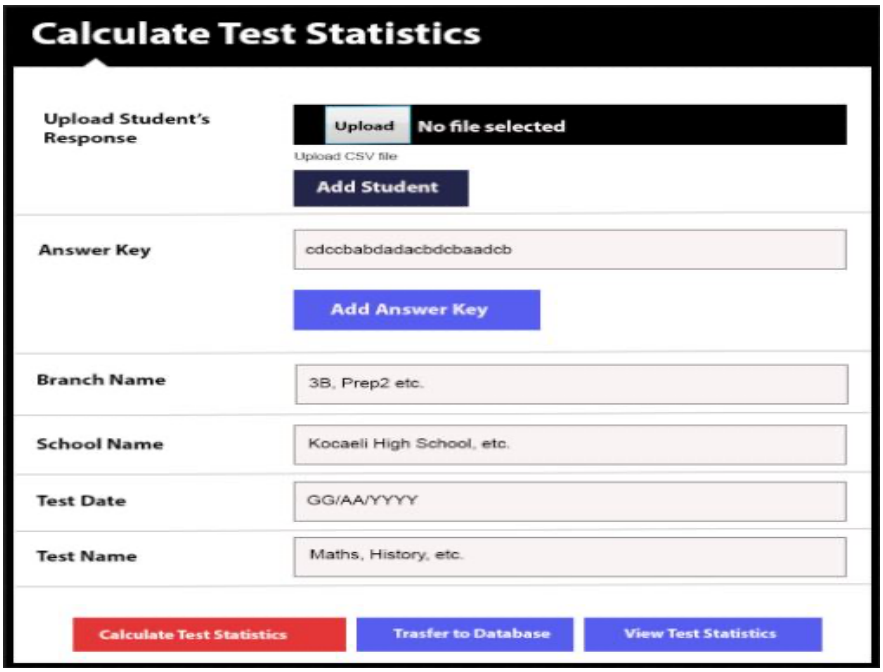


**Figure 2.** User exam diagnostic screen

The application, which presents the reports, tests and item statistics to the user in detail, also provides tips and information for him or her for the interpretation of the analyses. Also, the users transferring their exam data to the system can easily re-examine the reports stored in the

system when they want to go back and access to them at a later time.

Besides, a guide (as a reading file) has been created for the user to use the application easily. In this way, the user has comfortable access to the application practices. Furthermore, a separate section has been prepared for users who want to access more detailed information about test and item analysis than information given in the report and for the users who are curious to know how statistical calculations are made. In this section, the formulas of all statistical calculations and what they mean are explained extensively.

## Findings

### *1. Findings regarding the comparison of the data obtained from the online environment with the data obtained from other package programs*

At this stage, findings regarding the validity and reliability of the online environment (TOE) were shared. Data were produced in a simulative way through the TAP program, representing groups of 30, 100, and 500 students. While producing the data, 0.40, 0.60 and 0.90 average difficulties were chosen for each group. The results obtained for these groups were compared with the results obtained with SPSS and TAP package programs. The data obtained from both applications (TOE, SPSS) were grouped as test statistics and item statistics and compared on different tables. Results for the calculated test statistics are given in Table 1.

**Table 1.** Sample test statistics related to different data sets

| | N=30, $P_{mean}$=0.40 | | N=100, $P_{mean}$=0.60 | | N=500, $P_{mean}$=0.90 | |
|---|---|---|---|---|---|---|
| | SPSS | TOE | SPSS | TOE | SPSS | TOE |
| Minimum | 0.000 | 0.000 | 2.000 | 2.000 | 16.000 | 16.000 |
| Maximum | 22.000 | 22.000 | 28.000 | 28.000 | 30.000 | 30.000 |
| Mode | 9.000 | 9.000 | 15.000 | 15.000 | 29.000 | 29.000 |
| Median | 9.000 | 9.000 | 15.000 | 15.000 | 28.000 | 28.000 |
| Range | 22.000 | 22.000 | 24.000 | 24.000 | 14.000 | 14.000 |
| Mean | 9.400 | 9.400 | 15.400 | 15.400 | 27.228 | 27.228 |
| Std. Deviation | 5.834 | 5.834 | 5.963 | 5.963 | 2.408 | 2.408 |
| Variance | 34.041 | 34.041 | 35.556 | 35.556 | 5.800 | 5.800 |
| Coefficient of Variation | - | 62.069 | - | 38.720 | - | 8.845 |
| Skewness | 0.749 | 0.749 | -0.056 | -0.056 | -1.128 | -1.128 |
| Kurtosis | -0.182 | -0.182 | -0.707 | -0.707 | 1.324 | 1.324 |
| Test Difficulty Coeff. | 0.313 | 0.313 | 0.513 | 0.513 | 0.908 | 0.908 |
| KR-20 Reliability Coeff. | 0.863 | 0.863 | 0.835 | 0.835 | 0.603 | 0.603 |
| KR-21 Reliability Coeff. | 0.838 | 0.838 | 0.816 | 0.816 | 0.586 | 0.586 |

As can be seen in Table 1, the test statistics obtained from the online environment developed and the test statistics obtained from the SPSS package program are exactly the same. Comparisons were made for nine different data sets, and three are presented as an example. As a result of all comparisons, it was seen that the environment developed mostly gave the same results with the SPSS package program. The differences that occur in the decimal part vary depending on how many digits after the commas are used and reflected into the report. Using the same data sets, not only test statistics but also item statistics were calculated. Calculated item statistics are compared with those of TAP program and tabulated. Three different values are configured for the item discrimination in the developed online environment. These are sub-group-upper group item discrimination, biserial correlation and point bi-serial correlation values. However, since a statistical report based on point bi-serial correlation was presented in the TAP program, the only item discrimination value to be

reported in this study was the point bi-serial correlation. Three examples are presented for comparisons pertaining to nine different situations.

**Table 2.** Item statistics calculated in n = 30, $p_{ort}$ = 0.40 data set

| | Item Dificulty Index | | Item Discrimination Index | | | Item Dificulty Index | | Item Discrimination Index | |
|---|---|---|---|---|---|---|---|---|---|
| Item | TAP | TOE | TAP | TOE | Item | TAP | TOE | TAP | TOE |
| **M1** | 0.330 | 0.333 | 0.440 | 0.440 | **M16** | 0.200 | 0.200 | 0.370 | 0.370 |
| **M2** | 0.070 | 0.067 | 0.240 | 0.230 | **M17** | 0.170 | 0.167 | 0.530 | 0.520 |
| **M3** | 0.300 | 0.300 | 0.640 | 0.630 | **M18** | 0.670 | 0.667 | 0.410 | 0.400 |
| **M4** | 0.100 | 0.100 | 0.270 | 0.260 | **M19** | 0.330 | 0.333 | 0.170 | 0.170 |
| **M5** | 0.430 | 0.433 | 0.630 | 0.620 | **M20** | 0.070 | 0.067 | 0.030 | 0.030 |
| **M6** | 0.200 | 0.200 | 0.420 | 0.410 | **M21** | 0.270 | 0.267 | 0.630 | 0.620 |
| **M7** | 0.470 | 0.467 | 0.520 | 0.510 | **M22** | 0.630 | 0.633 | 0.080 | 0.080 |
| **M8** | 0.130 | 0.133 | 0.400 | 0.390 | **M23** | 0.370 | 0.367 | 0.380 | 0.370 |
| **M9** | 0.130 | 0.133 | 0.690 | 0.680 | **M24** | 0.300 | 0.300 | 0.470 | 0.470 |
| **M10** | 0.530 | 0.533 | 0.580 | 0.570 | **M25** | 0.130 | 0.133 | 0.230 | 0.230 |
| **M11** | 0.230 | 0.233 | 0.510 | 0.500 | **M26** | 0.470 | 0.467 | 0.270 | 0.270 |
| **M12** | 0.170 | 0.167 | 0.510 | 0.510 | **M27** | 0.270 | 0.267 | 0.430 | 0.420 |
| **M13** | 0.470 | 0.467 | 0.670 | 0.660 | **M28** | 0.530 | 0.533 | 0.390 | 0.380 |
| **M14** | 0.470 | 0.467 | 0.470 | 0.460 | **M29** | 0.300 | 0.300 | 0.630 | 0.620 |
| **M15** | 0.430 | 0.433 | 0.490 | 0.480 | **M30** | 0.230 | 0.233 | 0.590 | 0.580 |

Item reliability and item variance statistics are not presented to the user by the TAP program. Therefore, item difficulty and item discrimination indices are tabulated. When Table 2 is analyzed, it can be seen that the values of item difficulty and discrimination indices are the same in some items, while they differ in some others. When item difficulty index is considered, it is seen that the difficulty index of 6 items is exactly the same and 24 items are quite close to one another. This differentiation varies within a range of 0.000 - 0.010. These differences vary depending on how many digits are used after commas. Although the same formulas are used at this point, differentiation can be observed in the statistics obtained due to the differentiation in mathematical rounding operations.

**Table 3.** Item statistics calculated in n = 100, $p_{ort}$ = 0.60 data set

| | Item Dificulty Index | | Item Discrimination Index | | | Item Dificulty Index | | Item Discrimination Index | |
|---|---|---|---|---|---|---|---|---|---|
| Item | TAP | TOE | TAP | TOE | Item | TAP | TOE | TAP | TOE |
| **M1** | 0.500 | 0.500 | 0.400 | 0.400 | **M16** | 0.410 | 0.410 | 0.330 | 0.330 |
| **M2** | 0.490 | 0.490 | 0.490 | 0.480 | **M17** | 0.650 | 0.650 | 0.420 | 0.420 |
| **M3** | 0.600 | 0.600 | 0.360 | 0.360 | **M18** | 0.720 | 0.720 | 0.360 | 0.360 |
| **M4** | 0.290 | 0.290 | 0.370 | 0.370 | **M19** | 0.330 | 0.330 | 0.310 | 0.310 |
| **M5** | 0.410 | 0.410 | 0.500 | 0.500 | **M20** | 0.680 | 0.680 | 0.490 | 0.480 |
| **M6** | 0.410 | 0.410 | 0.390 | 0.380 | **M21** | 0.790 | 0.790 | 0.250 | 0.240 |
| **M7** | 0.680 | 0.680 | 0.340 | 0.340 | **M22** | 0.280 | 0.280 | 0.350 | 0.350 |
| **M8** | 0.600 | 0.600 | 0.440 | 0.440 | **M23** | 0.420 | 0.420 | 0.510 | 0.500 |
| **M9** | 0.470 | 0.470 | 0.450 | 0.440 | **M24** | 0.660 | 0.660 | 0.400 | 0.400 |
| **M10** | 0.740 | 0.740 | 0.290 | 0.290 | **M25** | 0.660 | 0.660 | 0.360 | 0.360 |
| **M11** | 0.530 | 0.530 | 0.390 | 0.390 | **M26** | 0.390 | 0.390 | 0.570 | 0.570 |
| **M12** | 0.360 | 0.360 | 0.370 | 0.370 | **M27** | 0.550 | 0.550 | 0.480 | 0.480 |
| **M13** | 0.320 | 0.320 | 0.430 | 0.420 | **M28** | 0.320 | 0.320 | 0.400 | 0.400 |
| **M14** | 0.570 | 0.570 | 0.330 | 0.330 | **M29** | 0.550 | 0.550 | 0.540 | 0.540 |
| **M15** | 0.610 | 0.610 | 0.420 | 0.410 | **M30** | 0.410 | 0.410 | 0.600 | 0.600 |

When Table 3 is analyzed, it is observed that the difference between item difficulty and discrimination indices varies between 0.000-0.010 in the data set of 100 people, wherein average difficulty is 0.60. Comparisons of data produced for 500 people with average difficulty of 0.90 are presented in Table 4.

**Table 4**. Item statistics calculated in n = 500, $p_{ort}$ = 0.90 data set

| | Item Dificulty Index | | Item Discrimination Index | | | Item Dificulty Index | | Item Discrimination Index | |
|---|---|---|---|---|---|---|---|---|---|
| Item | TAP | TOE | TAP | TOE | Item | TAP | TOE | TAP | TOE |
| **M1** | 0.830 | 0.828 | 0.320 | 0.320 | **M16** | 0.870 | 0.872 | 0.280 | 0.280 |
| **M2** | 0.960 | 0.956 | 0.200 | 0.200 | **M17** | 0.930 | 0.930 | 0.240 | 0.240 |
| **M3** | 0.870 | 0.870 | 0.270 | 0.270 | **M18** | 0.940 | 0.942 | 0.250 | 0.250 |
| **M4** | 0.810 | 0.808 | 0.370 | 0.370 | **M19** | 0.980 | 0.976 | 0.170 | 0.170 |
| **M5** | 0.950 | 0.950 | 0.260 | 0.260 | **M20** | 0.940 | 0.936 | 0.180 | 0.180 |
| **M6** | 0.860 | 0.856 | 0.360 | 0.360 | **M21** | 0.900 | 0.900 | 0.290 | 0.290 |
| **M7** | 0.810 | 0.810 | 0.370 | 0.370 | **M22** | 0.970 | 0.974 | 0.200 | 0.200 |
| **M8** | 0.890 | 0.886 | 0.390 | 0.390 | **M23** | 0.960 | 0.958 | 0.210 | 0.210 |
| **M9** | 0.920 | 0.920 | 0.230 | 0.230 | **M24** | 0.980 | 0.982 | 0.150 | 0.150 |
| **M10** | 0.900 | 0.898 | 0.370 | 0.370 | **M25** | 0.970 | 0.968 | 0.210 | 0.210 |
| **M11** | 0.950 | 0.950 | 0.210 | 0.210 | **M26** | 0.890 | 0.888 | 0.310 | 0.310 |
| **M12** | 0.850 | 0.854 | 0.380 | 0.380 | **M27** | 0.850 | 0.852 | 0.340 | 0.340 |
| **M13** | 0.980 | 0.982 | 0.060 | 0.060 | **M28** | 0.950 | 0.948 | 0.230 | 0.230 |
| **M14** | 0.920 | 0.924 | 0.270 | 0.270 | **M29** | 0.980 | 0.978 | 0.150 | 0.150 |
| **M15** | 0.790 | 0.790 | 0.440 | 0.440 | **M30** | 0.840 | 0.842 | 0.380 | 0.380 |

When the item statistics related to all data sets are examined individually, it is seen that the values are exactly the same or they have quite negligible differences compared to those of the TAP program. It can be said that these slight differences depend on the amount of the decimal portion of the other variables used in calculating item discrimination, item variance, item difficulty and item reliability values and on whether the rounding operation is performed or not.

### *2. Findings regarding the usability of the improved online environment*

While making the usability analysis, the records of the clicks and access times of each of the users were kept during the usage period of the application (with the representative test data provided to them). This process includes the analysis of the data initially given to user, accessing, saving and re-viewing the report he or she recorded. Recording click counts and access times were carried out to demonstrate the effectiveness and easy learnability of the application. Access times of the users are given in Table 5.

Table 5 displays how many times each user clicked and how much total time he or she spent during his or her experience with the application. When the data is analyzed, it is seen that the average number of clicks of the users is 27.80 and the total time spent in the application is 1 min. 35 seconds on average. Considering the time spent, effort exerted, the detail and function of the report reached as a result of the analysis, it can be said that the online environment is quite functional.

**Table 5.** Number of clicks and access times of users

| Users ID | Clicks | Access Times (min:sec) |
|---|---|---|
| $U_1$ | 22 | 01:22 |
| $U_2$ | 27 | 01:37 |
| $U_3$ | 19 | 01:20 |
| $U_4$ | 25 | 01:35 |
| $U_5$ | 24 | 01:32 |
| $U_6$ | 32 | 01:40 |
| $U_7$ | 35 | 01:45 |
| $U_8$ | 28 | 01:40 |
| $U_9$ | 26 | 01:35 |
| $U_{10}$ | 31 | 01:32 |
| $U_{11}$ | 27 | 01:35 |
| $U_{12}$ | 27 | 01:37 |
| $U_{13}$ | 29 | 01:40 |
| $U_{14}$ | 28 | 01:32 |
| $U_{15}$ | 37 | 01:45 |

The scores given to each item by the teachers who were surveyed are given in Table 6 below. At this point, it was seen that teachers generally had dissatisfaction about the interface. As it will be mentioned later in the opinions section pertaining to the online environment, users generally gave negative opinions about the interface due to its color, and about the fact that when it was clicked to perform an operation, it was not clear whether or not that operation occurred. In line with this feedback, changes were made in the environment and these 15 teachers were asked to use the environment again. After the changes, it was observed that there were positive changes in the opinions of the teachers. The results of the teachers' survey scores are presented in Table 6.

Table 6. The scores given by the users to the application through the usability survey

| Users ID | First Survey Scores | Last Survey Scores* |
|---|---|---|
| 1 | 12 | 12 |
| 2 | 14 | 12 |
| 3 | 18 | 13 |
| 4 | 15 | 13 |
| 5 | 18 | 13 |
| 6 | 16 | 14 |
| 7 | 16 | 12 |
| 8 | 15 | 13 |
| 9 | 15 | 13 |
| 10 | 12 | 12 |
| 11 | 15 | 12 |
| 12 | 14 | 12 |
| 13 | 16 | 13 |
| 14 | 15 | 13 |
| 15 | 18 | 13 |

* Less score is better

When Table 6 is examined, it can be seen that upon the changes made on the application no change was observed in the opinions of two of the teachers, and the opinions of 13 of them changed positively.

The feedback of the users about the online environment was recorded. In line with this feedback, corrections were made both on the application and on the format of the report output provided by the application. Sample statements regarding the opinions of the users are presented below.

U1 displays his positive opinion about the report output given by the application as follows,

U1- "… It is really nice to have value ranges under each value in the report, all being explained to us…".

U7 emphasized that in some stages, the application was insufficient due to the fact that it was hard to find out whether a given operation performed by the user was really accomplished or not and the application was corrected after this feedback. This criticism about the interface of the application was reported as follows,

U7- *"... After adding the answer key, I clicked on it, nothing happened, I was not sure if it was added or not, I could not understand what happened ..."*

One of the criticisms about the online environment is that the report output provides only a result of the test statistics. At this point, U5 stated that he wanted to see information about students in the report;

U5- *"... Actually, if we can see not only the exam statistics but also the numbers of the correct and wrong answers of the student, it might be better…".*

One of the criticisms about the interface is related to whether the system transfers data to the database works or not. It is parallel to the problem stated by U7. U3 expressed this inconvenience as follows,

U3- *"... After calculating the exam statistics, I wanted to transfer them to the database, I clicked, I clicked again, but nothing happened. But later, I realized that I already saved it in the database when I first clicked to do so, but unfortunately I couldn't understand it because the system did not inform me to that end in any way… "*

U12 expressed their opinions on uploading the student information collectively in the CSV file as follows.

U12- *"... Uploading the students with a CSV file frightened me a bit at first, as I wasn't sure if I could do it but I managed to do it eventually because it was explained nicely on the site…"*

U9, on the other hand, gave an opinion on the timing of the confirmation mail sent to the people during membership application.

*U9- "... When I got my membership, the confirmation e-mail came a little late, I waited for a long while, so the delay of things like this is not very nice, obviously, but it was due to the internet maybe, I don't know if it was so. But if it is a system related problem, I think it should*

*be corrected ...”*

## Conclusion and propositions

Measurement and evaluation have always been included in teacher training programs as one of the teacher competencies to be developed. In order to evaluate an educational program and its process, learning deficiencies, the development of the student in the process, whether the goal is reached at the end, a control mechanism is needed. Although traditional and complementary evaluation approaches are used together in making classroom measurements, multiple choice test has recently become one of the most popular types of tests worldwide due to the number of students and to the need for central exams. In our country, ÖSYM (MSPC /The Measuring, Selection and Placement Center), and MEB /MoNE please state the full form here prior to the acronym make use of this type of tests in their central exams and they are still used in class measurements at secondary school, high school and university level. The quality of the evaluation with multiple choice items is without doubt related to the quality of the items used. Using this test type, like other test types, imposes the teacher the responsibility to have the information and equipment suitable for the test type. Preparing distinguishing items that are appropriate for a given purpose and that are suitable for a given class in terms of the level of difficulty require knowledgeability and calculation on teachers’ part. In addition, interpreting about the group with the scores obtained from the test is one of the parts of the evaluation. Although test and item statistics are included in the measurement and evaluation course given in teacher training programs, it is seen that teachers cannot use and interpret them effectively. In many studies conducted on the measurement and evaluation competencies of teachers, these deficiencies stand out. One of the common emphases of the research is that, as stated by Anılan, Anagün, Atalay and Kılıç (2016), Duban and Küçükyılmaz (2008), teachers consider measurement and evaluation activities important. However, research show that they experience problems related to the measurement and evaluation process and teachers’ self-perception of the knowledge and skills they possess are not adequate (Anıl & Acar, 2008; Adıyaman, 2005; Anılan, Anagün, Atalay & Kılıç, 2016; Bal, 2009; Çakan, 2004; Çoruhlu, Nas & Çepni, 2008 Duban & Küçükyılmaz, 2008; Gelbal & Kelecioğlu, 2007; Evin-Gencel & Özbaşı, 2013; Gömleksiz & Bulut, 2007; Güneş, Dilek, Hoplan, Çelikoğlu & Demir, 2010; Güven, 2008; Kilmen, Akın Köstereliođlu & Köstereliođlu, 2007; Özenç, 2013; Yanpar, 1992; Yapıcı & Demirdelen, 2007).

Considering that it is not possible to give up multiple choice tests easily, it is seen that there is a need for environments / applications that will facilitate the work of teachers. Hence, this research aimed to develop an online environment intending to address that need. It is concluded that the online environment developed within the scope of the research is above all quite practical and it saves time. This convenience is an important advantage for teachers with heavy workload having difficulties in creating time for test and item statistics. The users have obviously reached the report with the analysis results quite quickly since all the statistics in the report presented to the user in the developed online environment depending only on the process of uploading their files to the web environment along with entering the answer key. At the end of the research, the item and test statistics calculated in all data sets created for sampling different situations were quite similar to or the same as those of reliable and equivalent package programs. This shows that the mathematical infrastructure created in the background is working correctly.

The report to be obtained from the developed environment was designed to guide the teachers. For this reason, comments and remarks were provided on it according to definitions

and common critical threshold values stated in the literature. With this aspect, the online application facilitates the work of teachers who have incomplete and incorrect information in interpreting the test and item statistics. One of the important differences of the online environment created by this research is its keeping a record of the past, accessing it with an internet connection without the need for an additional program and it is also usable on mobile platforms. Users can access their historical records (and reports) and review them from any platform having internet access. It is thought that the ability of users to keep records in their own profile individually for classes and exams will have a facilitating and positive effect on the follow-up of student success and exam parameters in the long term. The fact that the environment is in the Turkish language has facilitated the use of teachers. As a result of the usability questionnaire, it was concluded that the online environment has high availability/usability based on the opinions of the teachers. Further work has been carried out to develop it in order to support the English language and to use the data directly in the studies, the options of histogram, line, circle graphs and normal distribution curves are considered to be provided in the outputs, as well. Considering their intricacy and significance, the quality of measurement and evaluation and thus of education will increase by proceeding systematically and decisively, and by taking the right steps in places where precautions should be taken.

**Acknowledgement**

This study was based on Gözde Tekalmaz's master thesis titled "Developing an Online Environment for the Calculation of Test and Item Statistics" (2019), at Kocaeli University Institute of Social Sciences.

**References**

Adıyaman, Y. (2005). *İlköğretim 4, 6 ve 8. sınıflarında Türkçe dersine giren öğretmenlerin ölçme değerlendirme düzeyleri [The measurement and assessment levels of teachers who take Turkish lessons in primary education 4, 6 and 8 grades]*. Unpublished doctoral dissertation. Afyon Kocatepe University, Afyon.

Alexander, D. (2003). *Usability and human factors. Proceedings of Web Workshop series at Monash University*. 12 Mart 2019 http://www.its.monash. edu.au/staff/ web/ slideshows/ usability-humanfactors/.

Anıl, D., & Acar, M. (2008). Elementary school theachers' vıews on issues they experıence through measurement and evaluatıon processes. *Yüzüncü Yıl University, Journal of Education*, 5(2), 44-61.

Anılan H., Anagün Ş. S., Atalay N., & Kılıç Z., (2016). Classroom teachers vıews about measurement and evaluatıon approaches based on learnıg process. Journal of Research in Education and Teaching, 5(22), 200-221.

Bal, A. P. (2009). Evaluation of assessment and evaluation approaches in primary school fifth grade mathematics teaching in line with the opinions of teachers and students. Unpublished doctoral dissertation. Çukurova University, Adana.

Bıçak, B. & Çakan, M. (2004). *Lise öğretmenlerinin sınıf içi ölçme ve değerlendirme uygulamalarına dönük görüşleri[*The opinions of high school teachers on in-class assessment and evaluation practices*]*. Ministry of Education, Symposium of Reconstruction in Secondary Education, Ankara.

Birgin, O. & Gürbüz, R. (2008). Sınıf öğretmeni adaylarının ölçme ve değerlendirme konusundaki bilgi düzeylerinin incelenmesi [Examination of the knowledge level of elementary teacher candidates on measurement and assessment.]. *Selçuk University Journal of the Institute of Social Sciences*, 20, 163-179.

Brooks, V. (2002). *Assesment in Secondary Schools*. Buckingham: Open University Pres. Bloom, B. S. (1956). Taxonomy of Educational Objectives, The Classification of Educational Goals. Handbook I. Cognative Domain. New York, David McKay Company Inc.

Büyüköztürk,Ş., Çokluk, Ö. & Köklü, N. (2015). *Sosyal bilimler için istatistik [Statistics for Social Sciences]*. Ankara, Pegem Akademi.

Coşkunserçe, O. & Dursun, Ö. Ö. (2008). Methods used in the evaluation of websites. 8th International Educational Technology Conference Proceedings, 12 Mart 2019 http://ietc2008.home.anadolu.edu.tr/ ietc2008/172.doc.

Çakan M. (2004). Comparison of Elementary And Secondary School Teachers in Terms of Their Assessment Practices And Perceptions Toward Their Qualification Levels. *Ankara University, Journal of Faculty of Educational Sciences*, 37(2) 99-114.

Çoruhlu, T.Ş., Nas, S.E., & Çepni, S. (2009). Problems facıng scıence and technology teachers usıng alternatıve assesment tecnıcs: Trabzon sample. *Yüzüncü Yıl University, Journal of Education*, 6(1), 122-141.

Daniel, L. G. & King, D. (1998). A knowledge and use of testing and measurement literacy of elementary and secondary teachers. *Journal of Educational Research*, 91(6), 331-344.

Duban, N., & Küçükyılmaz E. A. (2008). Primary education pre-service teachers' opinions regarding to the use of alternative measurement-evaluation methods and techniques in practice schools. *Elementary Education Online*, 7(3), 769-784.

Erdinç O., & Lewis J. R., (2013) Psychometric Evaluation of the T-CSUQ: The Turkish Version of the Computer System Usability Questionnaire. *International Journal of Human-Computer Interaction*, 29, 319-326.

Evin-Gencel. İ., & Özbaşı D. (2013). Investigating prospective teachers' perceived levels of competence towards measurement and evaluation. *Elementary Education Online*, 12(1), 190-201.

Gelbal, S. & Kelecioğlu, H. (2007). Teachers' proficıency perceptıons of about the measurement and evaluatıon technıques and the problems they confront. *Hacettepe University Journal of Educaiton*, 33, 135-145.

Gömleksiz, M. N., & Bulut, D. (2007). An evaluation of the effectiveness of the new mathematics curriculum in practice. *Educational Sciences:Theory and Practice*, 7(1), 41-49.

Gözütok, F. D., Akgün, Ö. E. & Karacaoğlu, Ö. C. (2005). Evaluation of primary education programs in terms of teacher competencies. Symposium on Evaluation of New Primary Education Programs (Kayseri). 17-40. Ankara: Sim Matbaası.

Güneş, T., Dilek, N. Ş., Hoplan, M., Çelikoğlu, M., & Demir, E. S. (11-13 Kasım 2010). Teachers' opinions on alternative assessment and their applications. *International Conference on New Trends in Educationand Their Implications*, 925-935.

Güven, S. (2008). The classrom teachers' views concernıng the applıcatıon of the new primary school programmes. *Journal of National Education*, 177, 224-236.

Güven, S. (2001). *Sınıf öğretmenlerinin ölçme ve değerlendirmede kullandıkları yöntem ve tekniklerin belirlenmesi* [Determining the methods and techniques used by classroom teachers in measurement and evaluation] 10th National Educational Scieences Congress, Abant İzzet Baysal University, Bolu.

Kilmen, S., Akın Köstereliođlu, M., & Köstereliođlu, İ. (2007). Teacher candidates' perceptions of competence in children for assessment and evaluation tools and approaches. *AİBU Journal of Faculty of Education*, 7(1), 129-140.

Kutlu, Ö. (2003). Cumhuriyetin 80. yılında: Ölçme ve değerlendirme[In the 80th year of the republic: measurement and evaluation], *Journal of National Education*,160.

Lazar, J. (2001). *User-centered web development*. Boston: Jones and Bartlett Publishers.

Mertler, C.A. Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. Improving Schools, 12(2), 101-113. https://doi.org/10.1177/1365480209105575

Nielsen, J. (2000). *Why you only need to test with 5 users*. Jakob Nielsen's Alertbox web sitesinden 9 Mart 2019 tarihinde http://www.useit.com/ alertbox/20000319.html adresinden erişildi.

Nielsen, J. (2003). *Usability 101: introduction to usability*. Jakob Nielsen's Alertbox web sitesinden 9 Mart 2019 tarihinde http://www.useit.com/ alertbox/20030825.html adresinden erişildi.

Özenç, M. (2013). Determination of levels of primary school teachers' alternative assessment and evaluation knowledge. Dicle University Journal of *Ziya Gökalp Education Faculty*, 21, 157-178.

Özçelik, D.A. (2014). *Eğitim programları ve öğretim[Educaitonal Programs and Teaching]*. (3. Baskı). Ankara: Pegem A Yayıncılık.

Richey, R. C., & Klein, J. D. (2014). *Research on design and development*. In J. M. Spector, M. D. Merrill, J. van Merrienboer, M. P. Driscoll (Eds.), Handbook of Research for Educational Communications and Technology (pp. 141-150). NewYork :Springer.

Turan-Oluk, N. & Ekmekci, G. (2017). Comparıson of alternatıve assessment technıques wıth tradıtıonal technıques ın terms of measurement of student success. *JRES*, 4(2), 172-199.

Turgut, F., & Baykul, Y.(2010). *Eğitimde ölçme ve değerlendirme [Measurement and Evaluation in Education]*. Ankara :Pegem Akademi.

UPA. (2008). *UPA*, 9 Mart 2019 http://www.upassoc.org/usability_resources/ about_usability/ index.html.

Yanpar, T. (1992). *Ankara ilkokullarındaki ikinci devre öğretmenlerinin öğretmenlik mesleği ve konu alanlarıyla ilgili eğitim ihtiyaçları[*Training needs of second cycle teachers in Ankara primary schools regarding the teaching profession and subject areas*]*. Unpublished doctoral dissertation. Hacettepe University Educational Science Instute, Ankara.

Yapıcı, M. & Demirdelen C., (2007). Teachers' views with regard to the primary 4th grade social sciences curriculum. *Elementary Education Online*, 6(2), 204-212.

Yaşar, Ş., Gültekin, M., Türkan , B., Yıldız, N., & Girmen, P. (2005). Determination of classroom teachers' readiness levels and educational needs regarding the implementation of new primary education programs (Eskişehir Sample).*Symposium on Evaluation of New Education Programs*, Kayseri.