Eurasian Journal of Educational Research 69 (2017) 135-158

Eurasian Journal of Educational Research
www.ejer.com.tr

Inventory of Motive of Preference for Conventional Paper-and-Pencil Tests: A Study of Validity and Reliability*

Mehmet Taha ESER[1] Nuri DOGAN[2]

ARTICLE INFORMATION

Article History:

Received: 08 December 2016

Received in revised form: 18 March 2017

Accepted: 04 May 2017

DOI:http://dx.doi.org/10.14689/ejer.2017.69.8

Keywords

Choice of examination type
content validity
exploratory factor analysis
confirmatory factor analysis

ABSTRACT

Purpose: The objective of this study is to develop the Inventory of Motive of Preference for Conventional Paper-And-Pencil Tests and to evaluate students' motives for preferring written tests, short-answer tests, true/false tests or multiple-choice tests. This will add a measurement tool to the literature with valid and reliable results to help determine why students prefer certain exam types and their level of preference. **Research Methods:** In this study, a screening research design was employed during the data collection and the analysis phases. **Findings:** Cronbach's alpha coefficients were calculated for reliability and it was concluded that the inventory was reliable. First, the exploratory factor analysis was conducted; this was followed by a second confirmatory factor analysis and finally a content validity study to determine the construct validity. A total of 14 items, including 11 items according to the results of the exploratory factor analysis, 1 item based on expert opinion and 2 items according to the results of the confirmatory factor analysis were removed from the survey form of the inventory, resulting in a final form containing 20 items. It was observed that the content validity values of each item in every subtest were sufficient. **Implications for Research and Practice**: The study results showed that this inventory was an appropriate instrument for evaluating high school students' preference for paper-and-pencil tests. An inventory developed under the scope of this study may be used to determine the factors predicting the examination type preference levels of students by using different samples. These results may be used when deciding the actions to be taken.

* The present study is based primarily on the master thesis of Mehmet Taha ESER at Hacettepe University, supervised by Nuri DOĞAN, entitled "Factors Affecting Students' Preferences for the Type of Exam."This study was partly presented at the 3rd International Eurasian Educational Research Congress in Muğla, 31 May – 03 June, 2016.

[1] Corresponding Author: Department of Education, Union of Municipalities of Turkey, Ankara, Turkey, tahaeser@gmail.com

[2] Hacettepe University, Ankara, Turkey, nuridogan2004@gmail.com

Introduction

Evaluation is a process of judging based on the comparison of results obtained from a measurement process of criteria (Turgut, 1997). Evaluation usually takes place when the learning process ends, and it is carried out independently from teaching (Gülbahar and Büyüköztürk, 2008). However, methods for evaluating students should be helpful in providing information and feedback on what is learned by students at what level, what they face during the learning process, and how they prepare for exams (Gülbahar and Büyüköztürk, 2008; Birenbaum, 1997;Struyven, Dochy and Janssens, 2005). In the Turkish educational system, usually the post-examination choices of students are considered and discussed. The grade points of students from a large-scale examination are used to allow them to choose their university and department. These large-scale examinations consist of multiple-choice tests, yet the examination type choices of students are never taken into consideration. Students are compressed into a single model and only given multiple-choice test items.

In most traditional methods, student achievement is typically evaluated using mainly written exams, short answer tests, true/false tests and multiple-choice tests (Turgut, 1988; Atılgan, Kan and Dogan, 2009; Gelbal and Kelecioglu, 2007). The classroom and out-of-classroom behaviours of students are followed by using conventional paper-and-pencil tests. Their performance is examined and students are evaluated invarious aspects of the subject. As teachers are used to it, they prefer the traditional paper-and-pencil tests as a measurement tool (Gelbal and Kelecioglu, 2007).

Considering the qualities of the exam types, we see that exams have different advantages and disadvantages. The most significant advantage for multiple-choice, true/false and short-answer tests is that they are quick and easy to score. Written tests offer students an opportunity to demonstrate their knowledge, skills and abilities in a variety of ways. Multiple-choice tests take time and skill to construct; true/false tests encourage guessing; short-answer tests encourage students to memorize terms and details; and written tests require extensive time to grade. Some of these advantages work in the students' favour and some have a positive effect on the validity and reliability of the measurement results (Zoller, 1994). While some researchers and implementers have theoretically mentioned the positive effects of the exam types, there is relatively little research regarding the advantages and disadvantages of the exam types from the eyes of the students (Zoller and Ben-Chaim, 1998; Zoler and Ben-Chaim, 1990).

The initial studies focused on the type of examination chosen by students and whether these choices varied based on gender (Grandt, 1987; Zoller and Ben-Chaim, 1990). The majority of studies since 1994 used the Assessment Preference Inventory developed by Birenbaum (1994, 1997, 2007). Studies after this date mainly reviewed the relations between the learning-related features of students and their assessment preferences. These studies placed emphasis on learning-related qualities, such as assessment preference choices, learning strategies, motivation strategies, learning approaches, study strategies and academic achievement. The findings revealed that there are strong relations between the assessment preference choices of students and their learning-related qualities and emphasized the importance of considering their assessment preferences during the education process (Birenbaum 1997, 2003, 2007;

Biggs, 2003; Struyven, Dochy and Janssens, 2005; Wilson and Fowler, 2005; Birenbaum and Rosenau 2006; Watering, Gijbels, Dochy and Rijt, 2008).

There are various studies on assessments in the literature, particularly for teachers (Cavanagh, 2006; Cooney, Sanchez & Ice, 2001; Kyriakides, 1997; Miller, 2004; Motsoeneng, 2005; Saxe, Franke, Gearhart, Howard & Crockett, 1997; Sherin & Drake, 2009; Uchiyama, 2004, 2005); however, the number of studies on students, particularly in higher education, is limited (Ben-Chaim & Zoller, 1997; Birenbaum & Feldman, 1998; Struyven et. al., 2005 and Zeidner, 1987). These studies indicate that the assessment preferences may vary based on the education, departments and gender (Beller and Gafni, 2000; Ben Chaim & Zoller, 1997; Birenbaum & Feldman, 1998; Birenbaum, 1997; Brown & Hirschfeld, 2007; Bryant, 2001; Struyven et al., 2005; Watering et al., 2008; Zoller & Ben-Chaim, 1990). In this sense, the determination of assessment preference of students studying at the education faculties may be considered as an important factor to reflect their viewpoints on education, and to increase the quality of teaching and provide effectiveness in the program.

When we examined the relevant literature in Turkey, we found very few studies which attempted to determine the examination types of students (Gülbahar and Büyüköztürk, 2008; Bal, 2012; Bal, 2012). It was considered necessary to contribute to the field by developing "The Inventory of Motive of Preference for the Conventional Paper-and-Pencil Tests" (IMP-PAPT) as there was scant research to determine the reason for students' preference of an examination type.

Bal (2012) conducted research on the measurement and assessment preferences of prospective classroom teachers in mathematics.The study used the Assessment Preference Scale (APS) tool for the data collection which was developed by Birenbaum (1994) for university students and adapted forthe Turkish culture by Gülbahar and Büyüköztürk (2008). The Assessment Preference Scale used in the study includes mixed types of questions and intends to determine the level of preference of the assessment types in an integrated way, and not to determine the specific assessment type against certain conditions. However, *IMP-PAPT* developed within the scope of thisstudy, doesnot include mixed types of questions and this inventory provides detailed information on the type of assessment preferred under certain conditions. This study is a scale development study, rather than a scale adaptation study. Scale adaptation studies are more limited in terms of time, budget, and in making an international assessment in a cultural sense. They are also limited in researchers' knowledge ofscale development and any literature that has a strong validity and reliability value in relation to the relevant measurement results in the literature (Hambleton and Patsula, 1999). Taking into account the factors mentioned above, a scale development study on the subject has been carried out.

Purpose of the Study

The objective of this study is to develop *IMP-PAPT* for evaluating the motives of students to prefer written tests, short-answer tests, true/false tests and multiple-choice tests. This will add to the literature a measurement tool with valid and reliable measurement results to help determine the motives of students to prefer written tests, short-answer tests, true/false tests and multiple-choice tests and the level of of

preference for these exams. On the other hand, this study will provide teachers with information on the factors affecting the students' preference of examination types and the way these factors affect the examination preference level. Depending on the results, teachers may increase their efforts to develop measurement toold according to the certain qualities of students when they draft examinations to measure the student achievement. It is believed that the factors the teachers pay attention to in the test development process will reflect positively on students, thereby minimizing thenegative effects of tests on students.

In this study, we want to explore which assessment formats are preferred and how students perceive rather conventional assessment formats. Furthermore, we want to investigate the role of perceptions of assessment in the learning process. It is thought that having information about students' preferences for evaluation types will help students become knowledgeable about test anxiety and trait anxiety, as well as identify student learning strategies and learning styles. At the same time, the scale developed within the scope of this study can be used in studies where the factors affecting students' preferences regarding the types of evaluation are to be determined.

Method

Research Design

This study used the screening model. The studies on the screening model by Cohen, Manion and Morrison (2007) indicate that this is an ideal research methodfor studies on variables requiring a wide sample, such as preference and attitude.

Research Sample

The population of the study consisted of the 9th and 12th grade students studying in the central districts of the Bartin province. The exploratory factor analysis (EFA) was used in a study group of 100 student volunteers. The confirmatory factor analysis (CFA) was conducted on the data collected from 783 student volunteers consisting of 485 girls from various high schools (Bartin Davut Firincioglu Anatolian High School, Köksal Toptan Anatolian High School, Bartin Science High School, Bartin Religious Vocational High School) who studied in the Bartin province and completed and agreed to the research application. The 12th year students study in different fields, which are classified as numerical, verbal, equal weight and language. The size of the study group was considered sufficient for both types of analysis (Klein, 1994; Byrne, 1998). The Davis technique was used in the content validity study; and in this context, meant that opinions were received from 12 experts in the field of assessment and educational evaluation who are competent in the related field.

Many studies, which were inspired by Gardner's AMTB, were conducted in the field. Some of them focused on instrumental and integrative orientations for learning. In the Chinese EFL context, Xiong, 2010 investigated motivational differences among middle school students and observed that they had both instrumental and integrative motivation for learning English. In the Iranian EFL context, studies examined learners' motivational orientations and reported high instrumental motivation among foreign language learners (Hashemi and Hadavi, 2014; Vaezi, 2008). In the Turkish context,

some studies supported that finding (Bektas-Cetinkaya, 2012; Koseoglu, 2013; Ozturk and Gurbuz, 2013). All studies indicated the dominance of instrumental motivation among EFL students.

Research Instrument and Procedure

IMP-PAPT drafted by Eser (2011) was created to reveal the motives of preference on the examination types, such as written, short answer, true/false and multiple-choice and to measure the level of preference of these examinations by students. The survey form of the inventory consisted of 34 items. In this study, both exploratory and confirmatory factor analyses were used. CFA and EFA are, in fact, two stages of a whole process and cannot be effectively separated. If the researcher can use these two methods together, the research will achieve a deeper degree of understanding. Anderson and Gerbing suggested that during the procedure of proposing a theory, it is better to establish a model by EFA and verify the model or modify the model by CFA (Anderson & Gerbing, 1990). EFA provides concepts of the hypothesis and calculating tools, which are an important basis and guarantee for the establishment theory in CFA. It is uncertain if anyone in EFA or CFA is omitted in factor analysis (Hu ve Li, 2015).The final form of *IMP-PAPT* consisted of 20 items. Fourteen items were removed from the initial scale, i.e., 1 item by the expert opinion view method, 11 by exploratory factor analysis, and 2 items by confirmatory factor analysis. When writing the items, the motives of preference of student were considered to be the qualities of examinations that were found to be important with respect to validity, reliability and usefulness. Students were asked to state their preference level on the examination types of written, short answer, true/false and multiple-choice. In the process of preparing the inventory, views and feedback were taken from three PhD students and one associate professor, all of whom are experts in the field of measurement tools.

The scoring of the inventory was based on the following: For me, the responses given to the items are not correct=1, partly correct=2 and totally correct=3. When scoring the items, separate scoring was made for each examination type. Points given for each item indicate the level of preference of individuals while the total points indicate the preference level of the concerned examination by individuals. The examination preference levels of individuals indicate a value between one and three, as they were obtained by taking averages. The values closer to three indicate a higher preference level and show that generally a high point is obtained from the motives of preference for the concerned examination. The points of individuals closer to one indicate lower preference level and show that generally a low point is obtained from the motives of preference for the concerned examination.

Results

Results of Exploratory Factor Analysis

The exploratory factor analysis was applied to the items on each subtest to determine the number of dimensions of the subtests in the inventory. As a result of the analysis, the factor loads for the written examination subtest were found to be between 0,32 and 0,69;those for the short answer examination were between 0,32 and 0,68;those for the true/false subtest were between 0,42 and 0,64; and those for the multiple-choice subtest were between 0,31 and 0,66. According to Tabachnick and Fidell (2001), the

factor load value of each item should be 0,32 or higher. Therefore, the factor load lower limit was accepted at 0,32 when deciding the items to remain in the scale. The KMO values for subtests were between 0,71 and 0,75. It was decided that the data number was sufficient for the factor analysis according to the KMO values results. In addition, the Bartlett test results for all tests were found to be significant at a level of 0,01. This result was considered to be proof that the factor analysis could be applied to the data.

When we look at the eigenvalues of the written examination subtest, seven factors were found with eigenvalues higher than one. The variance disclosed by the first factor (eigenvalue 5,806) was found to be 26,392% while the variance disclosed by the second factor (eigenvalue 2,233) was 10,151%. The factors consisting of all components on the written examination subtest were found to explain 65,303% of the total variance. When we look at the eigenvalues of the short-answer examination subtest, eight factors were found with eigenvalues higher than one. The variance disclosed by the first factor (eigenvalue 5,133) was found to be 23,332% while the variance disclosed by the second factor (eigenvalue 1,815) was 8,249%. The factors consisting of all components on the short-answer examination subtest explained 67,231% of the total variance. When we look at the eigenvalues of the true/false examination subtest, eight factors had eigenvalues higher than one. The variance disclosed by the first factor (eigenvalue 5,338) was found to be 24.265%, while the variance disclosed by the second factor (eigenvalue 1,713) was 7,784%. The factors consisting of all components on the true/false examination subtest explained 66,763% of the total variance. When we look at the eigenvalues of the multiple-choice examination subtest, six factors were found with eigenvalue higher than one. The variance disclosed by the first factor (eigenvalue 5,377) was found to be 24.439%, while the variance disclosed by the second factor (eigenvalue 1,839) was 8,359%. The factors consisting of all components on the short-answer examination subtest explained 57,924% of the total variance.

The factor loads and scree plots on the four subtests were examined and a majority of the items in each subtest was collected under a single dimension (Appendix 1, Appendix 2, Appendix 3, Appendix 4). Depending on the factor analysis results, items that are not included in the first dimension and do not have sufficient factor load to be included in any dimension or those that have high or similar factor load in multiple dimensions were removed from the subtests. After evaluating this, it was deemed appropriate to remove 11 items from the test for all subtests (items 13, 15, 16, 17, 23, 26, 27, 30, 31, 32, 34). Experts agreed on the fact that the fourth item was not suitable for the inventory, and, as a result, the fourth item was removed from all subtests regardless of its statistical values.

In conclusion, it was determined that each subtest was one-dimensional and the practice was continued with 22 items taking into consideration the factor loads, eigenvalues, disclosed variance values and scree plots. An inventory was prepared for the motives of preference using four subtests: written examination, short-answer test, true/false test and multiple-choice test. Subsequently, the correlation values between the corrected test points (obtained by subtracting the correlated item from the total point) and item points were checked in order to determine Cronbach's alpha's internal consistency reliability and item discriminating power.

The Pearson correlation of the test and item points for the written examination scale varied between 0,217 and 0,606; the short-answer test scale varied between 0,217 and

0,598; the true/false test scale varied between 0,215 and 0,532; and the multiple-choice test scale varied between 0,236 and 0,571 (Table 1). Since we paid attention to keep the same items for the four subtest types, each item with a test-item correlation of less than 0,20 for any subtest was removed regardless of the test-item correlation level in the subtests (Ebel, 1979, Field, 2009).

Table 1

Item-Test Correlations and Cronbach's Alpha Values for The Written, Short-Answer, True/False and Multiple-Choice Tests.

Item	Written examination	Short answer test	True/false test	Multiple-choice test
1	.392	.456	.471	.315
2	.232	.344	.366	.335
3	.284	.297	.385	.245
4	.554	.447	.532	.523
5	.594	.510	.527	.524
6	.498	.481	.494	.513
7	.606	.598	.497	.518
8	.401	.282	.466	.385
9	.480	.408	.465	.417
10	.395	.434	.417	.571
11	.485	.426	.367	.384
12	.217	.261	.242	.239
13	.418	.217	.233	.253
14	.380	.218	.225	.276
15	.452	.398	.413	.431
16	.393	.438	.426	.488
17	.466	.458	.387	.486
18	.343	.408	.482	.487
19	.577	.512	.506	.470
20	.228	.266	.215	.236
21	.501	.433	.426	.422
22	.560	.398	.396	.479
Cronbach's Alpha	.856	.831	.838	.838

When we looked at the Cronbach's alpha internal consistency coefficients for the points from four subtests on 22 items, we found that these coefficients varied between 0,831 and 0,856. These values are high and the measurement results are sufficiently

reliable. At the same time, the reliability values of the subtest scores are similar and very close to each other with respect to homogeneity.

The factor loads given in Table 1 relate only to the EFA results. Since the EFA was conducted with100 students, and the sample is small, the factor load wastaken as the lower limit of 0.20.

In Table 1, the averages of the item discrimination indices are shown. The mean of the item discrimination indices is 0.39 for the written test, 0.36 for the short answer test, 0.37 for the true/false test, and 0.34 for the multiple-choice test. The subscales are sufficiently distinguished as the average discrimination values for the subtests are over 0.30.

Results of the Confirmatory Factor Analysis

Each subtest was applied to 783 individuals for the confirmatory factor analysis that was planned to test the construct validity of the subtests. The confirmatory factor analyses included the testing of single dimensionality of the subtests as a model. As the second and fourth items caused autocorrelation in some items during the confirmatory factor analysis, these items were removed from the subtests. The confirmatory factor analyses were done after removing the two items. Table 2 includes the model concordance indicators obtained after the confirmatory factor analysis.

Table 2

Model Concordance Indicators According to the Confirmatory Factor Analysis on the Subtests of the Inventory of Motive of Preference for Examinations

Subtest	Chi-square/ 2	GFI/AGFI	NFI	NNFI	CFI	RMSEA	RMR	SRMR
Written	4,81	0,96 / 0,94	0,99	0,99	0,99	0,079	0,032	0,065
Short-Answer	4,52	0,96 /0,96	0,99	0,99	0,99	0,067	0,028	0,057
True/false	4,01	0,97 / 0,96	0,99	0,99	0,99	0,062	0,027	0,052
Multiple-choice	4,01	0,97 / 0,96	0,99	0,99	0,99	0,062	0,028	0,052

Looking at the confirmatory factor analysis result in Table 2, we can state that there is sufficient evidence on the one dimensionality of each subtest. The chi-square statistics in the literature show a lack of index fit (Stapleton, 1997). Therefore, a small chi-square value indicates that the model is fit for the observed structure and vice versa. That is, a big chi-square value indicates that the model does not sufficiently explain the structure. However, as the chi-square statistic is a sum statistics, it will be as high as the number of variants. Therefore, the use of chi-square/degree of freedom might be recommended (Dogan and Basokcu, 2010). Having a chi-square/degree of freedom lower than five indicates that the model fits and a value lower than three indicates that the model has a very good fit (Byrene, 1998). Having chi-square/degree

of freedom values between three and five in the study indicates that the one-dimensional models created for the subtests are fit for the observed structures.

A goodness of fit index is usually a measurement of the variance and covariance amount disclosed by the model. The coefficient of determination calculated in the multiple regression can be interpreted as R^2. The closer the value of the goodness of fit index, the better the fit of the model for the data (Dogan and Basokcu, 2010). For the goodness of fit indices, the values between 0,90-0,95 indicate an acceptable fit; values above 0,95 indicate a high fit (Dickey, 1996; Stapleton, 1997; Byrne, 1998). The values in Table 2 show that the fit indices other than RMR and SRMR are larger than 0,95. The GFI/AGFI, NFI, NNFI and CFI values indicated that the measurement tool had a high fit. Particularly, having the index value of Root Mean Square Error of Approximation (RMSEA) between 0,08-0,05 shows that the model is acceptable, and a value lower than 0,05 shows that the model is good. Particularly, a good fit is indicated by an index value of the Root Mean Square Error of Approximation (RMSEA) closer to 0,00 (Du Toit and Du Toit, 2001). In our study, the RMSEA values lower than 0,08 indicate an acceptable fit. A good fit is also indicated by the fact the RMR and SRMR values are ≤ 0.08, as these two values are indicators of lack of fit (Jöreskog and Sörbom, 1993). A high fit is proven by the fact that the RMR value, which is an indicator of lack of fit, is between 0,027 and 0,032 for each subtest, while the SRMR values are observed to be lower than 0,08 by varying between 0,052 and 0,065. Considering and interpreting all values together provides a verification of the one dimensionality structure of the subtests. The path graph of the confirmatory factor analysis for the subtests is given in the appendices (Appendix5, Appendix 6, Appendix 7, Appendix 8).

Results of Content Validity

For each item in the subtest composing the assessment tool, opinions were received from 12experts in the field of assessment and evaluation in education. In the determination of content validity related to items, the Davis technique (1992) was used. Considering the requirement that a minimum of three experts use the Davis Technique, this number was met as we received opinions from seven experts in terms of content validity. The surveys related to content validity were conducted with the remaining items after the items having a negative effect on content validity were excluded from the test. Using the Davis technique each item related to the subtests were evaluated as 1=*not relevant*, 2=*somewhat relevant*, 3=*quite relevant*, 4=*highly relevant*. When determining the content validity index for each item, the number of experts choosing the option (3) or (4) was divided by the total number of experts to obtain content validity index and 0,80 was determined as the standard value for CVI's (Davis, 1992).

The content validity indexes of the items forming the assessment tool varied between 0,86 and 1 for written examinations, short answer tests, true/false test and multiple-choice tests. Considering that the limit value for the Davis technique is 0,80, the content validity values of each item in every subtest was sufficient.

Discussion and Conclusion

In this study, a scale was developed to determine the levels of high school students regarding their motives of preference for paper-and-pencil tests. The relevant

literature was reviewed to develop the draft scale and then the scale was applied to the high school students. Cronbach's alpha coefficients were calculated for reliability and it was concluded that the inventory was reliable. First the exploratory factor analysis and then the confirmatory factor analysis were conducted to determine the structure validity. A total of 14 items were removed from the survey, including 11 items according to the results of the exploratory factor analysis, 1 item by expert opinion and 2 items according to the results of the confirmatory factor analysis, leaving 20 items in the final form.

The Assessment Preference Scale,developed by Birenbaum (1994) for university students and adapted forthe Turkish culture by Gülbahar and Büyüköztürk (2008) contains similar objectives to the inventory developed in thepresent study and this scale was used in a majority of similar studies (Gülbahar and Büyüköztürk, 2008; Bal, 2012; Birenbaum, 1994; Birenbaum, 1996; Birenbaum, 1997). Further studies may be recommended to examine the criteria validity studyofthe level of relations between the inventory developed in the present study and the Assessment Preference Scale.

The subtests of the inventory developed by the study consist of four traditional examinations: written, short-answer, true/false and multiple-choice test. Future studies may include different types of traditional examinations and the research may revise the scale or develop an inventory of motives of preference for the examination type created by the complementary measurement approach. The inventory developed under the scope of this study may be used to determine the factors predicting the examination type preference levels of students by using different samples. These results may be used when deciding the actions to be done and tools to be used in the assessment process by determining the examination type preferences of the students.

The Assessment Preference Scale used in the study includes mixed types of questions and intends to determine the level of preference of the assessment types in an integrated way, rather than determine aspecific assessment type against certain conditions. However, *IMP-PAPT*developed within the scope of the study does not include mixed type of questions and this inventory provides detailed information on the type of assessment preferred undercertain conditions. As mentioned earlier, this study is a scale development study. Therefore, in order to avoid the difficulties such as limited time, low budget, a language and culture adapted from a different language and culture, a detailed plan was made prior tothe study. As a result, it will be useful for the researchers to make a detailed plan before the scale development studies are carried out.

References

Anderson, J. C. & Gerbing, D.W. (1990). Structural equation modeling in practice: A review and recommended two-step procedure. *Psychology Bulletin*, V103:411-423.

Atilgan, H., Kan, A., Dogan, N. (2009). *Egitimde Olcme ve Degerlendirme*. Edit. H. Atilgan (3. Basim): Ani Yayincilik.

Bal, A. P. (2012). Sinif Ogretmenligi Ogretmen Adaylarinin Matematik Dersine İliskin Olcme-Degerlendirme Tercihleri. *Turk Egitim Bilimleri Dergisi*, 10(3), 459-479.

Bal, A. P. (2012). Ogrencilerin Matematik Dersine İliskin Degerlendirme Tercihleri. *Selcuk Universitesi Sosyal Bilimler Enstitusu Dergisi*, 27, 59-72.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open ended) account for gender differences in mathematics achievement? Sex Roles, 42, 1-21.

Ben-Chaim, D., & Zoller, U. (1997). "Examination-type preferences of secondary school students and their teachers in the science disciplines", *Instructional Science*, 25(5), 347–367.

Birenbaum, M. (1994). Toward Adaptive Assessment - The Student's Angle. *Studies in Educational Evaluation*, 20, 239-255.

Birenbaum, M. & Feldman, R. A. (1998). Relationships between learning patterns and attitudes towards two assessment formats, *Educational Research*, 40(1), 90–97.

Birenbaum, M. ve Gutvirzt, Y. (1995). On The Relationship Between Assessment Preferences, Cognitive Style, Motivation and Learning Strategies. Paper presented at the 11th conference of the Israeli Research Association. Jerusalem. The Hebrew University.

Birenanbaum, M. (1996). Alternatives in Assessment of Achievements, Learning Process and Prior Knowledge. USA: Kluwer Academic Publishers.

Birenbaum, M. (1997). Assessment Preferences and Their Relationship to Learning Strategies and Orientations. *Higher Education*. 33, 71-84.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment, in Segers, M., Dochy, F. and Cascallar, E. (eds.), Optimizing New Methods of Assessment: In Search of Qualities and Standards. Dordrecht, The Netherlands: Kluwer, pp. 13-36.

Birenbaum, M. ve Rosenau, S. (2006). Assessment preferences, learning orientations, and learning strategies of pre-service and in-service teachers. *Journal of Education for Teaching*, 32(2), 213-225.

Birenbaum, M. (2007). Assessment and instruction preferences and their relationship

Brown, G. T. L. & Hirschfeld, G. H. F. (2007). Students' conceptions of assessment and mathematics achievement: Evidence for the power of self-regulation. *Australian Journal of Educational and Developmental Psychology*, 7, 63-74.

Byrne, B. M. (1998). Structural equation modeling with lisrel, prelis and simplis: Basic concepts, aplications and programming. Mahwah, NJ: Lawrence Erlbaum Associates.

Cavanagh, M. (2006). Mathematics teachers and working mathematicaly: Responses to curriculum change. 2016. Retrieved from http://www.merga.net.au/publications/counter.php?pub=pub_conf&id=289

Cohen, L., Manion, L. & Morrison, K. (2007) Research Methods in Education. 6th edn. London: Routledge.

Cooney, T. J.,Sanchez, W. B. & Ice, N. F. (2001). Interpreting teachers' movement toward reform in mathematics. *The Mathematics Educator*, 11(1), 10-14.

Davis, L. L. (1992). Instrument review: Getting the most from your panel of experts. *Applied Nursing Research*, 5, 194–197.

Dickey, D (1996), Testing The Fit of Our Models of Psychological Dynamics Using Confirmatory Methods: An Introductory Primer. (Advances in Social Science Methodology, 4 icinde. Editor: Bruce Thompson). London: JAI press Ltd.

Du Toit, M. ve Du Toit, S. (2001). Interactive Lisrel: User's guide. Lincolnwood: Scientific Software International Inc.

Dogan, N. ve Basokcu T. O. (2010). "İstatistik Tutum Olcegi icin Uygulanan Faktor Analizi ve Asamali Kumeleme Analizi Sonuclarinin Karsilastirilmasi", *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, C. 1, S. 2, s. 65-71.

Ebel, R. L. (1979). Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall.

Eser, M. T. (2011*).Ogrencilerin sinav turu tercih nedenlerini etkileyen bazi faktorlerin incelenmesi.* Yayimlanmamis yuksek lisans tezi, Hacettepe Universitesi Sosyal Bilimler Enstitusu.

Everitt, B. S. (2002). The Cambridge dictionary of statistics (2nd ed). Cambridge; New York: Cambridge University Press.

Field, A. (2009). Discovering statistics using SPSS (Third Ed.). London: SAGE.

Gelbal, S., & Kelecioglu, H. (2007). Teachers' proficiency perceptions of about the measurement and evaluation techniques and the problems they confront. *Hacettepe University Journal of Education*, 33, 135-145.

Grandt, J. (1987). Characteristics of Examinees Who Leave Questions Unanswered on The GRE General Test Rights-Only Scoring. ETS Research Report 87- 83, Princton, NJ: Educational Testing Service.

Gulbahar, Y. ve Buyukozturk, S. (2008). "Degerlendirme Tercihlerin Olceginin Turkceye Uyarlanmasi", Hacettepe Universitesi Egitim Fakultesi, 35, 148-161.

Hambleton, R.K. ve Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, 1(1), 1-30.

Ho, R. (2006). Handbook of univariate and multivariate data analysis and interpretation with SPSS. Florida: Chapman ve Hall/CRC.

Hu, Z., Li, J. (2015). The integration of efa and cfa: One method evaluating the construct validity. *Global Journal of Human-Social Science*, 15 (6).

Joreskog, K. G., & Sorbom, D. (1993). LISREL 8: structural equation modeling with the simplis command language. Lincolnwood: Scientific Software International, Inc.

Kline, P. (1994). An easy guide to factor analysis. New York, NY: Routledge.

Kyriakides, L. (1997). Primary teacher's perceptions of policy for curriculum reform in mathematics. *Educational Research and Evaluation*, 3(3), 214-242.

Miller, T. (2004). Assessment in practicegrade 9 academic and applied mathematics. (Master Thesis). Queen's University, Kingston, Ontario, Canada.

Motsoeneng, K. G. (2005). The attitude of teacher and parents and learners involved in primary and intermediate schools in the Thabo Mofutsanyane district regarding assessment reform in education. Master Thesis, Bloemfontein University, Mofutsanyane Thabo.

Oren, S. F., Ormanci, U., Evrekli, E. (2014). Ogretmen Adaylarinin Tercih Ettikleri Alternatif Olcme-Degerlendirme Yaklasimlari İle Bu Yaklasimlara İliskin Oz-yeterlilikleri. *Egitim ve Bilim Dergisi*, 39, 173.

Saxe, G. B., Franke, M. L., Gearhart, M., Howard, S. ve Crockett, M. (1997). Teachers' shifting assessment practices in the context of educational reform in mathematics. CSE Technical Report 471, CRESST University of California, Los Angeles.

Sherin, M. G., & Drake, C. (2009). Curriculum strategy framework: investigating patterns in teachers' use of a reform-based elementary mathematics curriculum. *Journal of Curriculum Studies*, 41(4), 467-500.

Stapleton, C. D. (1997). Basic concepts and procedures of confirmatory factor analysis. *Educational Research Association, Reports-Evaluative* (142), Speeches / Meeting Papers (150)

Struyven, K., Dochy, F. & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: a review. *Assessment & Evaluation in Higher Education*, 30 (4), 325- 341.

Tabachnick, B.G., ve Fidell, L.S. (2001). Using Multivariate Statistics. (4th ed), Allyn& Bacon, Boston,

Turgut, M.F. (1997). *Egitimde olcme ve degerlendirme metodlari*. Ankara: Gul Yayinevi.

Turgut, M. F. (1988). *Egitimde Olcme ve Degerlendirme Metotlari*. Ankara: Saydam Matbaacilik, Altinci Baski.

Uchiyama, M. K. (2004). Teachers use of formative assessment in middle school reform based mathematics classrooms. PhD Dissertation, University of Colorado Boulder, Colorado.

Uchiyama, M. K. (2005). Teachers' use of formative assessment. Annual meeting of the american educational research association, Colorado State University, www.aera.net adresinden 13 Nisan 2005 tarihinde alinmistir.

Watering, G. V., Gijbels, D., Dochy, F. ve Rijt, J. V. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *High Education*, 56, 645-658.

Wilson, K., & Fowler, J. (2005). Assessing the impact of learning environments on students' approaches to learning: Comparing conventional and action learning designs. Assessment and Evaluation in Higher Education, 30(1), 87–101.

Worthington, R.,& Whittaker, T. (2006). Scale development research: A content analysis and recommendations for best practices. *Counseling Psychologist*, 34, 806-838. doi:10.1177/0011000006288127

Zoller, U. (1994). The examination where the student asks the questions. *School Science and Mathematics* 94(7): 347–349

Zoller, U. & Ben-Chaim, D. (1988). Interaction between examination type, anxiety state, and academic achievement in college science: an action-oriented research. *Journal of Research in Science Teaching 26(2): 65-77.*

Zoller, U. & Ben-Chaim, D. (1990). Gender differences in examination-type preferences, test anxiety, and academic achievements in college science education-A case study. *Science Education* 74(6): 597–608.

Geleneksel Kağıt-Kalem Testleri İçin Tercih Nedenleri Envanteri: Geçerlik ve Güvenirlik Çalışması

Atıf:

Eser, M. T. & Dogan, N. (2017). Inventory of motive of preference for conventional paper-and-pencil tests: A study of validity and reliability. *Eurasian Journal of Educational Research, 69, 135-158. http://dx.doi.org/10.14689/ejer.2017.69.8*

Özet

Problem Durumu: Bireylerin birbirlerinden farklı olmadığı fikri daha çok 20.yüzyıl inanışıdır. Bu fikir büyük olasılıkla Batı dünyasında gelişen "demokrasi" fikrine bağlıdır. Bu inanışa göre, en basit tanımlama ile insanlar birbirlerine eşit ise birbirlerinin aynısı olmalıdırlar. Ancak, yapılan araştırmalar sonucunda, her bireyin farklı karakter özellikleri, farklı zeka seviyeleri ve fiziksel yapıları ile oldukça özel bir donanıma sahip olduğu ortaya çıkmıştır. Bu yaklaşıma göre öğretmenlerin kendi sınıflarında daha başarılı sonuçlar almaları için öğrencilerinin karakterlerini, karakterlerini etkileyen etkenleri, öğrencilerin öğrenme modellerini ve öğrenme modellerini etkileyen etkenleri çok iyi bilmeleri ve göz önünde bulundurmaları gerekir.

Öğretim ve değerlendirme süreçlerinin daha da yakınlaştığı ve etkileşim içerisinde bulunduğu modern eğitim sistemlerinde, öğrencilerin değerlendirme süreci üzerindeki algıları ve değerlendirme yöntemleri seçimlerinin eğitim süreci ve öğrenimi boyunca dikkate alınması gerekir. Öğrencilerin başarıları belirlenirken uygulanan geleneksel kağıt kalem testleri; yazılı sınavlar, kısa cevaplı testler, doğru yanlış testleri, çoktan seçmeli testler, performans görevleri, portfolyo vb.'dir. Öğrencilerin bu geleneksel kağıt kalem testleri konusunda görüşlerini almak, öğretmenlere öğrenci başarısını belirlemede geri besleme ve öğrencilerin öğrenme süreçleri konusunda bilgi edinilmesi gerekmektedir. Bu çalışma öğrencilerin değerlendirme süreçleri üzerindeki algılarının önemini ve değerlendirme yöntemlerinin seçimlerini göz önüne alarak gerçekleştirilmiştir.

Araştırmanın Amacı: Araştırmanın amacı, öğrencilerin yazılı, kısa cevaplı, doğru-yanlış ve çoktan seçmeli testleri tercih etme nedenlerini değerlendirmeye ilişkin ''*Geleneksel Kağıt Kalem Testleri İçin Tercih Nedenleri Envanteri*'' geliştirerek, literatüre öğrencilerin bu sınav türlerini tercih etme nedenleri ile bu sınavları tercih düzeylerini tespit etmeye yardımcı olacak ölçme sonuçlarının geçerliği ve güvenirliği sağlanmış bir ölçme aracı kazandırılacağı düşünülmektedir. Elde edilen sonuçlara bağlı olarak öğretmenler öğrenci başarısını ölçmek amacıyla sınav hazırlarken öğrencilerin belirli özelliklerine göre ölçme aracı geliştirme çabasını arttırabilirler. Öğretmenlerin test geliştirme sürecinde dikkat edeceği faktörler öğrencilere olumlu bir şekilde yansıyacağı, testlerin öğrenciler üzerinde oluşturduğu olumsuz etkilerin en aza indirileceği düşünülmektedir.

Araştırmanın Yöntemi: 100 lise öğrencisinin oluşturduğu bir örneklemden elde edilen envanter ile ilgili veri setine ilişkin faktör analizi sonuçlarına göre; alt ölçekler için elde edilen faktör yükleri 0,32 ile 0,69 arasında değişmektedir. Alt ölçekler için KMO

değerleri 0,71 ile 0,75 arasında bulunmuştur. KMO değeri sonuçlarına göre veri sayısının faktör analizi için yeterli sayıda olduğuna karar verilmiştir. Tüm alt ölçekler için Bartlett testi sonuçları 0,01 düzeyinde manidar bulunmuştur. Bu sonuç, veri setinin faktör analizine uygun olduğunun bir işaretidir. Dört alt ölçeğe ilişkin faktör yükleri ve yamaç- birikinti grafikleri incelenmiş ve birinci boyutta yer almayan, herhangi bir boyutta yer alması için faktör yükü yetersiz olan veya birden fazla boyutta faktör yükü yüksek olan 11 maddenin envanterden çıkartılması uygun görülmüştür. Uzmanlar 4. maddenin envanter için uygun olmadığını bildirmişler ve 4. madde envanterden çıkartılmıştır. Sonuç olarak her bir ölçeğin tek boyutlu olduğuna karar verilmiş ve uygulamaya 22 madde ile devam edilmiştir. Her bir alt ölçek için iç tutarlığı görmek açısından Cronbach Alfa iç tutarlık katsayıları incelenmiş ve iç tutarlık katsayılarının 0,831 ile 0,856 arasında değiştiği gözlemlenmiştir. Bu değerler ölçeklerin kabul edilebilir güvenirliklere sahip olduğunu göstermektedir.

783 kişiye yapılan ikinci uygulama sonucuna doğrulayıcı faktör analizi uygulanmış; 2. ve 4. maddelerin diğer maddelerle otokorelasyona girdiği gözlemlenmiş ve bu maddelerin atılması uygun görülmüştür.

Doğrulayıcı faktör analizine ilişkin sonuçlar için X^2/sd' nin 5'ten küçük olması modelin uyum iyiliğine sahip olduğunun göstergesidir (Byrne, 1998). RMR değerlerinin 0,05' ten küçük olması mükemmel uyuma, SRMR değerlerinin 0,05 ile 0,08 arasında olması ise iyi uyuma işarettir. GFI/AGFI, NFI, NNFI, CFI değerleri ölçme aracının yüksek uyum verdiğini gösteren değerler almıştır. RMSEA değerlerinin 0,10' dan küçük olması kabul edilebilir bir uyumun göstergesidir. Bütün değerler bir arada ele alınıp yorumlanacak olursa; alt testlerin tek boyutluluk yapısına ilişkin doğrulamanın yeterince güvenilir biçimde sağlandığı söylenebilir (X^2/sd: 4,01-6,54; GFI: 0,96-0,97; AGFI: 0,94-0,96; NFI: 0,99; NNFI: 0,99; RMSEA: 0,062-0,084; RMR: 0,027-0,032; SRMR: 0,052-0,065). Araştırma kapsamında son olarak, kapsam geçerliği çalışması yürütülmüştür. Kapsam geçerliği anlamında ölçme aracını meydana getiren her bir alt testi oluşturan maddeler için, konu alanında yeterli donanım ve bilgiye sahip, çalışmanın öneminin farkında olan 12 eğitimde ölçme ve değerlendirme uzmanının görüşleri alınmıştır. Maddelere ilişkin kapsam geçerlik oranları belirlenirken Davis tekniği kullanılmıştır. Geliştirilen ölçme aracını meydana getiren maddelere ilişkin kapsam geçerlik indekslerinin yazılı sınav, kısa cevaplı test, doğru-yanlış testi ve çoktan seçmeli test için 0,86 ile 1 arasında değiştiği gözlemlenmiştir. Davis tekniği için sınır değerin 0,80 olduğu göz önünde bulundurulduğunda, maddelerin her bir alt testteki kapsam geçerlik değerlerinin yeterli düzeyde olduğu söylenebilir.

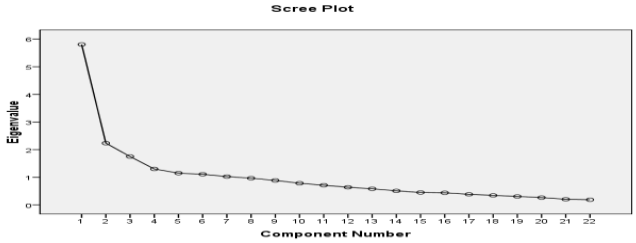
Araştırmanın Bulguları: Bu çalışma sonucunda, öğrencilerin geleneksel kağıt kalem testleri konusunda tercihlerinin belirlenmesine yönelik olan GKKT-TNE geliştirilmiştir. Envanter, 2 bölümden meydana gelmektedir. Envanterin ilk bölümünde demografik bilgilerin yer aldığı 4 madde, ikinci bölümünde ise 3'lü derecelendirilmiş 20 madde yer almaktadır.

Araştırmanın Sonuç ve Önerileri: Araştırma sonuçları, geliştirilen ölçeğin, lise öğrencilerinin kağıt ve kalem testlerine ilişkin tercih sebeplerini değerlendirmek için uygun bir araç olduğu görülmektedir. Bu çalışma kapsamında geliştirilen envanter, öğrencilerin ilgili sınavlara ilişkin sınav türü tercih seviyelerini farklı örnekler

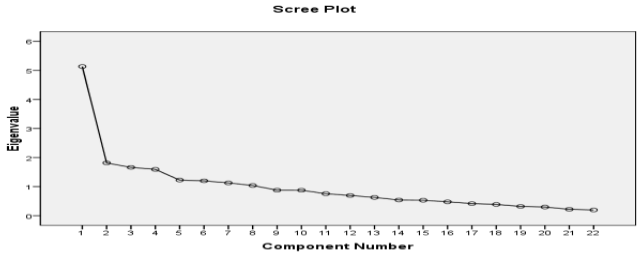
kullanarak tahmin eden faktörleri belirlemek için kullanılabilir. Bu sonuçlar, öğrencilerin sınav türü tercihlerini belirleyerek değerlendirme sürecinde gerçekleştirilecek eylemleri ve araçları belirlerken kullanılabilir.

Anahtar Kelimeler: Sınav türü tercihi, kapsam geçerliği, açımlayıcı faktör analizi, doğrulayıcı faktör analizi.

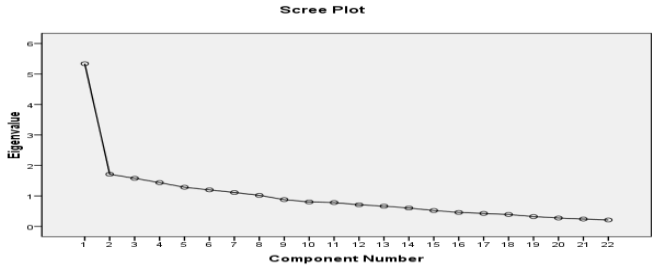
Appendix1**.** Scree Plot of the Written Examination Subtest



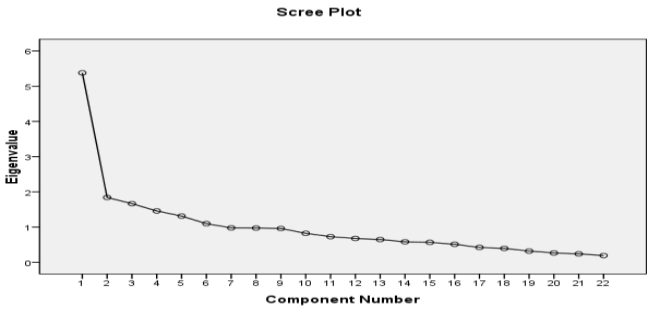
Appendix 2. Scree Plot of the Short Answer Examination Subtest



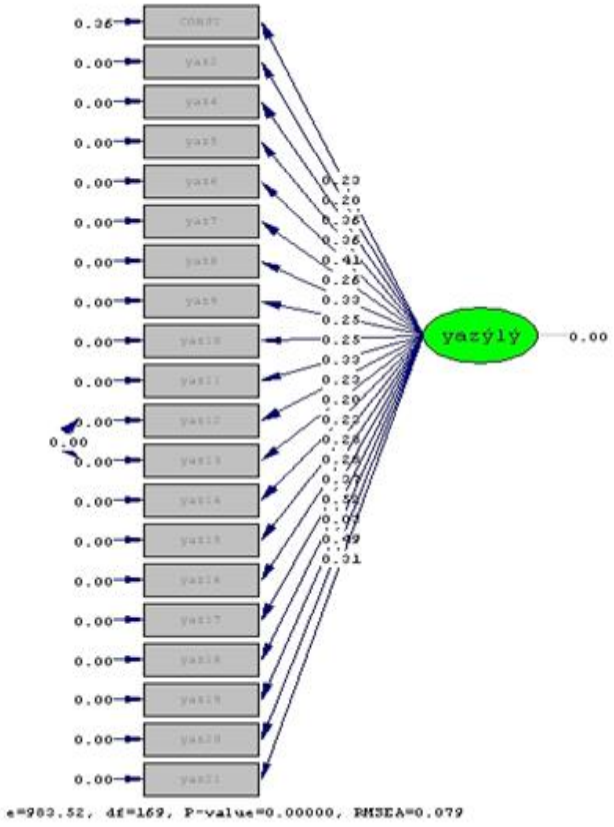
Appendix 3. Scree Plot of the True/false Examination Subtest



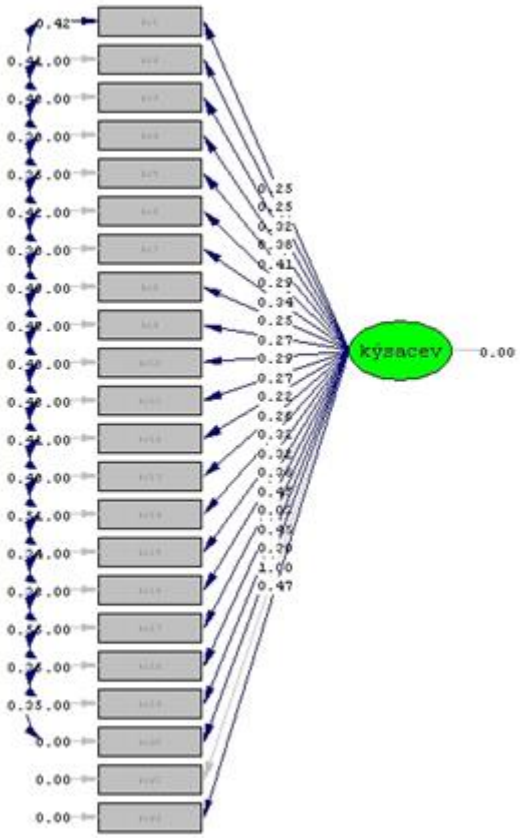
Appendix 4. Scree Plot of the Multiple-choice Examination Subtest



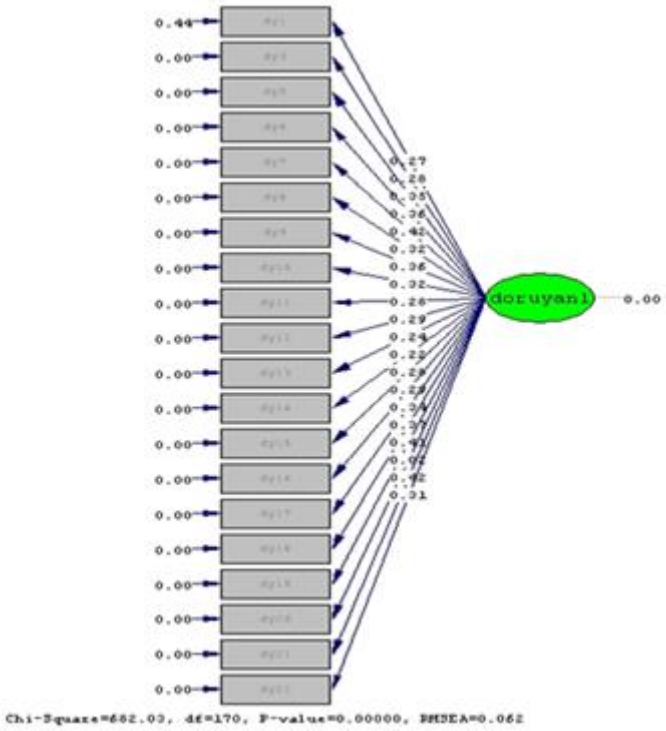
Appendix 5. Path Graph of the Confirmatory Factory Analysis of the Written Examination Subtest



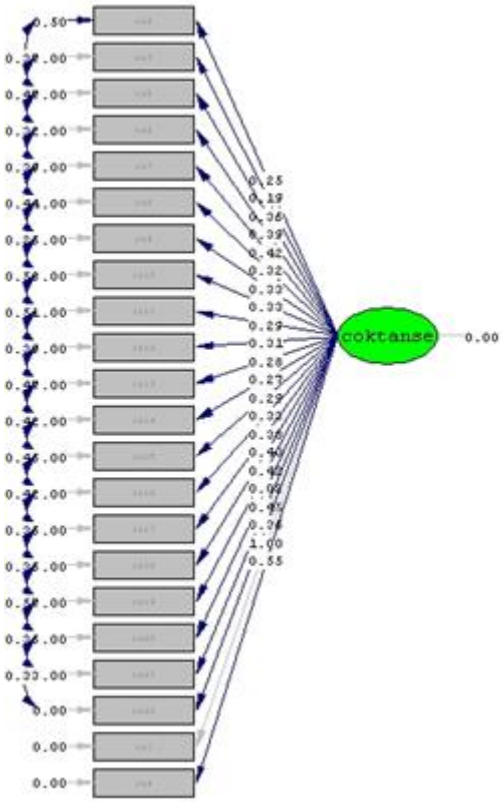
Appendix 6. Path Graph of the Confirmatory Factory Analysis of the Short-Answer Examination Subtest



Appendix 7.Path Graph of the Confirmatory Factory Analysis of the True/false Examination Subtest



Appendix 8. Path Graph of the Confirmatory Factory Analysis of the Multiple-choice Examination Subtest



Appendix 9. The Inventory of Motive of Preference for the Written, Short Answer, True/false and Multiple-choice Tests

Gender:
Grade:
Education Level of Mother:
Education Level of Father:

(1) True for me
(2) Partly true for me
(3) Totally true for me

Dear Students,
Please read the following items and mark the gap under the code with (x) indicating one of the judgments shown on the top right corner. Thank you for participating in our study.

In these examinations	**Examination Types**											
	Written Examinations			Short Answer Tests			True/false Tests			Multiple-choice Tests		
	1	2	3	1	2	3	1	2	3	1	2	3
1) I can cheat easily.												
2) I can easily let others copy from me.												
3) I have a chance in turning the wheel.												
4) I don't need to learn the subjects by heart.												
5) My preparation doesn't take time.												
6) My preparation is easy.												
7) I don't get nervous.												
8) It is easy for me.												
9) I use time efficiently.												

10) I don't have a problem focusing.												
11) I succeed.												
12) I don't feel the need to ask any questions to the teacher.												
13) I feel it is necessary to copy												
14) I don't panic if I finish early.												
15) I think the questions are difficult.												
16) I think the time to answer is not sufficient.												
17) It provides more correct results.												
18) It suits my learning style better.												
19) I easily express what I want to say.												
20) I can predict the score.												
21) I don't feel obliged to study.												
22) Reading is enough to study.												
23) Writing is enough to study.												
24) I finish answering quickly.												
25) It doesn't make me panic.												
26) I cannot be sure about my answer.												
27) I don't feel obliged to express myself/my thoughts.												
28) I get bored.												

29) I feel comfortable.													
30) I have a headache.													
31) I feel bad.													
32) I find it difficult.													
33) I trust in my response.													
34) I want to finish and get out quickly.													

Note: Bold statements are final inventory items.