Comparing Performance of Different Equating Methods in Presence and Absence of DIF Items in Anchor Test*

Neşe Gübeş [i]
Mehmet Akif Ersoy University

Şeyma Uyar [ii]
Mehmet Akif Ersoy University

Abstract

This study aims to compare the performance of different small sample equating methods in the presence and absence of differential item functioning (DIF) in common items. In this research, Tucker linear equating, Levine linear equating, unsmoothed and presmoothed (C=4) chained equipercentile equating, and simplified circle arc equating methods were considered. The data used in this study is 8th-grade mathematics test item responses which obtained from Trends in International Mathematics and Science Study (TIMSS) 2015 Turkey sample. Item responses from Booklet-1 (N=199) and Booklet-14 (N=224) are chosen for this study. Data analyses were completed in four steps. In the first step, assumptions for DIF detection and test equating methods were checked. In the second step, DIF analyses were conducted with Mantel Haenszel and logistic regression methods. In the third step, Booklet 1 was chosen as base form and Booklet 14 chosen as a new form, then test equating was conducted under common item nonequivalent groups design. Test equating was done in two phases: the presence and absence of DIF items in the common items. Equating results were evaluated based on standard error of equating (se), bias and RMSE indexes. DIF analyses showed that there were two sizeable DIF items in anchor test. Equating results showed that performances of equating methods are similar in presence and absence of DIF items from anchor test and there is no notable change in se, bias and RMSE values. While the circle arc equating method outperformed other equating methods based on se, 4-moment presmoothed chained equipercentile equating method outperformed other methods based on bias and RMSE evaluation criteria.

Keywords: Test Equating, Small Samples, Differential Item Functioning

DOI: 10.29329/ijpe.2020.248.8

---

*This study was presented as an oral presentation at the 6th International Congress on Measurement and Evaluation in Education and Psychology

[i] **Neşe Gübeş,** Assist. Prof., Education Faculty, Department of Educational Sciences, Mehmet Akif Ersoy University, ORCID: 0000-0003-0179-1986

Correspondence: nozturk@mehmetakif.edu.tr

[ii] **Şeyma Uyar,** Assist. Prof. Dr., Faculty of Education, Educational Sciences Department, Mehmet Akif Ersoy University

INTRODUCTION

In national and international testing programs, multiple forms of a single test are used to provide test security or to allow sampling a large of items without having each student answer all of the items. Alternative test forms which developed considering the same construct and blueprint almost differ somewhat in their difficulty. If one form is more difficult than other form, examinees would be expected to get higher scores from the easier form and get lower scores from the more difficult form. Test equating is required to remove effects on scores of these undesirable differences in test form difficulty (Dorans, Moses, & Eignor, 2010). As Kolen and Brennan (1995, p. 2) defined "equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably."

In testing programs like Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA) common-item nonequivalent groups (CINEG) design is used to equate test scores. In CINEG design, common items from different test forms are used to equate test forms. Like other statistical analysis methods, common item test equating is exposing to sampling error. One way to reduce sampling error is to conduct equating with large samples of examinees (Kurtz & Dwyer, 2013). Random equating error is directly related to sample size. The sample sizes required to conduct test equating accurately vary based on equating designs and equating methods. For example, "a random-groups design requires a much larger sample than a common-item design, which requires a larger sample than a single-group design" (Kim & Livingston, 2010, p. 286). Kolen and Brennan (2004) suggest that the minimum sample size for linear equating should be 400 and the minimum sample size for equipercentile equating should be 1500. However, large samples always may not be accessible in real test situations. Hence, a variety of methods has been recommended to cope with equating problem in small samples. These methods can be listed as identity, linear, chained equipercentile equating with log-linear presmoothing, circle arc, and synthetic equating (Babcock, Albano & Raymond, 2012). In this study, chained equipercentile equating, linear and circle arc methods are considered so information about these methods is given below.

Chained Equipercentile Equating

Chained equipercentile equating method is an alternative equipercentile equating method. Firstly, this method was described by Angoff (1971) and then Dorans (1990) named this method as chained equipercentile equating. In this method, Form X scores are equated to common items scores. Then scores of common items are equated to the Form Y scores. Assume that Form A is an anchor test for Form X and Form Y. Population P takes the Form X and Population Q takes the Form Y. The scores of Form X are equated to scores of anchor test A using examinees from Population P. Then anchor test A scores are equated to Form Y scores using Population Q. Because of including a chain of two equipercentile equating, it is called as chained equipercentile equating (Kolen & Brennan, 2004).

As Livingston (1993) reported smoothing in equipercentile equating is decreasing sample size requirements by about one half. Presmoothing and postsmoothing are two types of smoothing method which used in equipercentile equating. While the score distributions are smoothed in presmoothing, the equipercentile equivalents are smoothed in postsmoothing. Presmoothing can be done with a polynomial log-linear model or a strong true score model. In this study, we consider presmoothing with the polynomial log-linear model. For the polynomial log-linear presmoothing method, choosing the degree of the polynomial (C) is important because it limits how much smoothing is done. The C parameter is generally chosen from numbers from 1 to 10. After presmoothing, the fitted distribution has the moment preservation property. This means that first C moments of the fitted distribution are the same to sample distribution's first C moments. For instance, if C=2, the mean and standard deviation of the fitted distribution are the same to the mean and standard deviation of the observed distribution. Likelihood ratio chi-square goodness of fit statistic can be used for choosing the C parameter. For instance, the difference between chi-square statistics for C=3 and C=4 can be examined with one degree of freedom. A significant difference between chi-square values means that the more

complex model (C=4) fits the sample data more than the more simple model (C=3). If the two models fit the data adequately the simplest model should be chosen (Kolen & Brennan, 2004).

Linear Equating Methods

Linear equating assumes "apart from differences in means and standard deviations, the distributions of the scores on Form X and Form Y are the same" (Crocker & Algina, 1986, p.458). Tucker and Levine are most prevalent (Kolen & Brennan, 2004) linear equating method in CINEG design and their use in small samples are supported by prior researches for small sample sizes (Parshall, Du Bose Houghtan, & Kromrey, 1995; Skaggs, 2005). In this study, we consider Tucker and Levine linear equating methods. Tucker equating was described by (Gulliksen, 1950) and he attributed it to Ledyard Tucker (as cited in Kolen & Brennan, 2004). Tucker equating method makes two important assumptions: regression slopes of the total test scores on the common item score for both populations are equal and variance on the common item score between both examinee populations are equal (Kurtz & Dwyer, 2013). Levine observed score equating is another equating method which used with CINEG design. There are three assumptions in Levine observed score equating which related to the observed scores in classical test theory: (1) there is a perfect correlation between the true scores of total test and true scores of anchor test in the old and new form populations, (2) the total test true scores' regression on to the anchor test true scores are assumed to be the same linear function for the old form and the new form populations, (3) the measurement error variance for X is the same for Populations 1 and 2 (Kolen & Brennan, 2004).

Circle-Arc Equating

Livingston and Kim (2008, 2009) suggested the circle arc method for small-sample data. This method has a curvilinear equating function. There are two kinds of circle arc equating method: the symmetric circle arc and simplified circle arc. Circle arc equating gets its equating function due to arc connecting three points in a Cartesian coordinate system (Babcock et al., 2012). "The upper end of the curve is determined by the maximum possible score on each form. The lower endpoint of the curve is determined by the lowest meaningful score on each form. The middle point on the curve is determined from the data, by equating at one point in the middle of the score distribution. If those three points happen to lie on a straight line, that line is the estimated equating curve. If three points do not lie on a straight line, they determine an arc of a circle." (Livingston & Kim, 2009, p.332). Livingston and Kim (2008) reported that the circle arc method typically yielded more precise and less biased results than other methods (mean, linear and smoothed equipercentile equating) in small samples.

Differential Item Functioning

The other issue that should be considered in national and international testing programs differential item functioning (DIF). The purpose of DIF analysis is to determine items that function differently for examinees which have the same underlying ability from different subgroups. DIF studies are usually carried out regarding to reference and focal groups that are established by considering manifest (observed) group characteristics such as gender and ethnicity. It is supposed that the observed groups are homogeneous subgroups. In line with this assumption, an item containing DIF is considered advantageous or disadvantageous for all individuals in any manifest groups. Therefore, with these studies, once a DIF item has been determined, there is little knowledge about the examinees for which the item functions differentially (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002). However, there is a low relationship between the manifest characteristic associated with DIF and the actual advantaged or disadvantaged groups. Therefore, comparisons of item responses for manifest groups may lack sensitivity to determine the true source(s) of DIF (Cohen and Bolt, 2005; De Ayala et al., 2002; Oliveri, Ercikan, & Zumbo, 2013; Samuelsen, 2005).

Latent group means that to set the group membership to an unknown homogeneous subgroup which can be determined by mixture modeling (McLachlan & Peel, 2000). In mixture modeling, while

the item functions the same in a latent group, it functions differently among latent groups (Fieuws, Spiessens, & Draney, 2004). So the use of mixture IRT models can overcome this problem that rises with the use of manifest groups. They can make it possible to detect latent groups for which the items function differently (Cohen & Bolt, 2005). In this study, DIF analyses were conducted based on latent classes.

Latent DIF

The mixture model is defined by Rost (1990) as a "Mixture Rasch" model. It is a combination of latent class and the Rasch model. In this model, it is assumed that a population of examinees can be grouped into several discrete latent classes based on examinees response patterns. With this model, item parameters can be simultaneously estimated with individual's ability and the class he/she belongs (Alexeev, Templin, & Cohen, 2011; Cohen and Bolt, 2005; Mislevy and Verhelst, 1990; Rost, 1990). In these models, each latent class fits Rasch model but classes have different item difficulty parameters. Therefore, MixIRT models can simultenously determine subpopulations that display qualitative differences and quantify the differences in the ability within the groups (Mislevy & Verhelst, 1990; Rost, 1990). According to the model, the possibility of giving a correct answer is as follows.

$$P\left(y_{ijg}=1 \mid g,\theta_{jg}\right)=\frac{1}{1+\exp[-(\theta_{jg}-\beta_{jg})]} \qquad (1)$$

In equation 1, g=1,…,G refers to index with specified latent class; j=1,…,J refers to index with specified responders; θ_{jg}: j. refers to examinees latent ability in g latent class; β refers to difficulty parameter of i. item in class g.

If the DIF detection is carried out during the test construction process, the test developers usually delete flagged items from the test. However, in many situations DIF can be detected after data have been collected. In these situations, deleting DIF detected items may not be a good idea because item deletion can affect test reliability and validity negatively (Elosua & Hambleton, 2018). Hu and Dorans (1989) reported that deleting both minimal and sizable DIF items resulted in different scaled scores after IRT true score re-equating and Tucker re-equating. They also noted that the deleting item itself had a noticeable effect on scale score and the effect size of the DIF item had a less prominent effect on the scale scores (cited in Kolen & Brennan, 2005). Therefore, it is important to determine DIF items during test equating process and apply methods which reduce the effect of these items on test equating constants (Hidalgo-Montesinos & Lopez-Pina, 2002).

In literature, there are some researches (Atalay-Kabasakal & Kelecioğlu, 2015; Chu, 2002; Demirus & Gelbal, 2016; Turhan, 2006; Yurtçu & Güzeller, 2018) which have been compared equating methods in presence of DIF items. In these studies, IRT equating methods were compared in the presence or absence of DIF items. There is not any study which compares small samples equating methods in the presence and absence of DIF items in tests. Therefore this study aims to compare the performance of different small sample equating methods in the presence and absence of DIF in common items.

METHOD

In this study, performance of existing small equating methods was compared in the presence and absence of DIF in common items with using real data. Therefore, this study was designed as descriptive survey. In descriptive survey design, there is not any attempt to change or influence the study situation, existing situation is described (Karasar, 2009).

Data

The data used in this study is 8th-grade mathematics test item responses which obtained from Trends in International Mathematics and Science Study (TIMSS) 2015 Turkey (N=6079) sample. Item responses from Booklet-1 and Booklet-14 are chosen for this study. There are 14 dichotomous scored items common for both booklets. Booklet 1 consists of totally 32 dichotomously scored items and the maximum score which can be obtained are 32. Booklet 14 consists of totally 26 dichotomous scored and 1 polytomous scored (0-1-2) items and the maximum score which can be taken from Booklet 14 is 28. There are 199 students who took Booklet 1 and 224 students who took Booklet 14.

Table 1. Summary of Data

	Common Items	Total of Items	Maximum Score
Booklet-1 (N=199)	14 MC	32 MC	32
Booklet-14 (N=224)	14 MC	26 MC + 1 PS	28

Note. MC: Multiple Choice Item; PS: Polytomous Scored Items

Data Analyses

In this study, data analyses were completed in four steps. In the first step, the confirmatory factor analyses are carried out for Booklet 1 and Booklet 14 to assure the unidimensionality requirement for DIF detection and test equating methods. In the second step, DIF analyses are conducted with Mantel Haenszel (MH) and logistic regression (LR) methods based on latent class. In the third step, Booklet 1 is chosen as a base form and Booklet 14 chosen as a new form, then test equating is conducted under common item nonequivalent groups design. Test equating is done in two phases: the presence of DIF items in anchor test and removing sizable (C level) DIF items from anchor test. In this study, Tucker linear equating, Levine observed score equating, unsmoothed chained equipercentile equating, chained equipercentile equating with presmoothing (C=4), and simplified circle arc equating methods are considered. These equating methods are chosen because the researches showed that they gave accurate equating results in small samples (Babcock et. al, 2011, Kim & Livingston, 2010). Equating results are evaluated based on the standard error of equating, bias and root mean squared error (RMSE) index which provided from 1000 bootstrapped samples.

The Mplus (Muthen & Muthen, 1998-2012) computer program is to assess unidimensionality assumption; the WINMIRA (reference) computer program is used to find how many latent groups exist in the data and to estimate item parameters MixRasch analysis; "difR" R package (Magis, Beland and Raiche, 2015) is used to find DIF items across latent class; "equate" R package (Albano, 2017) is used for test equating and calculating bootstrapped standard error, bias, and RMSE indexes. In "equate" package of R, as reported Albano (2017) "standard errors are calculated as standard deviations over replications for each score point; bias is the mean equated score over replications, minus the criterion; RMSE is the square root of the squared standard error and squared bias combined." (p. 5).

RESULTS

Descriptive Statistics and Testing Assumptions

The descriptive statistics for Booklet 1 and Booklet 14 are reported in Table 2. As seen in Table 2, Booklet 1 has 32 multiple choices (MC) items and Booklet 14 has 26 MC and 1 polytomous scored (0-1-2) items. In both forms, the 14 MC items are common. The mean for Booklet 1 is 14.51 and for Booklet 14 is 12.61 and mean test difficulty is equal for both forms ($z=0.00$, $p>0.05$). Cronbach alpha reliability is .93 for Booklet 1 and .91 for Booklet 14 and there is no statistical difference between forms' reliability level ($z=.75$, $p>.05$). Item discrimination of items in each form was calculated by using point-biserial correlation coefficient. The mean of point-biserial correlations was the same and .50 for Booklet 1 and Booklet 14 (see Table 2).

Table 2. Descriptive Statistics for Booklet 1 and Booklet 14

	Booklet 1	Booklet 14
N	199	214
Number of items	32 MC	26 MC + 1 PC
Common items	14 MC	14 MC
Minimum score	3	1
Maximum score	32	27
Mean	14.51	12.61
Mode	7	7
Median	11.00	11.50
SD	8.24	7.14
Mean difficulty	.45	.45
Mean r_{pb}	.50	.50
Cronbach's Alpha	.93	.91

Note. N: Total number of students; SD: Standard deviation; rpb: Point biserial correlation

Prior to DIF analyses and equating test forms, confirmatory factor analyses (CFA) is carried out for Booklet 1 and Booklet 14 by Mplus (Muthén & Muthén, 1998-2012). Comparative fit index (CFI), Tucker Lewis index (TLI) and Root mean square error of approximation (RMSEA) indexes for Booklet 1 (CFI= .99, TLI= .99, RMSEA= .027) and Booklet 14 (CFI= .99, TLI= .99, RMSEA= .019) support that each form measures a unidimensional trait (Byrne, 2010; Hu & Bentler, 1999; Kline, 2005).

Estimation of Model Parameters

To determine fitted latent classes to the model, results of model comparison criteria for mixture Rasch solutions given in Table 3 is examined.

Table 3. Results of Model Comparison Information Criteria for Mixture Rasch Solutions

Number Of Classes	BIC
1	6505.23
2	6452.29
3	6512.80

In the MixIRT model applications, information criteria Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been widely used to select the model. Li, Cohen, Kim, and Cho (2009) suggested that the smallest BIC result should be used to determine the number of classes. Based on BIC values in Table 3, we can say that a model with two latent classes' best fit the data.

Results of DIF Analyses

DIF analyses were conducted by using Mantel Haenszel (MH) and logistic regression (LR) methods based on two latent classes. The DIF results are reported in Table 4.

Table 4. The Results of DIF Analyses

	MH			**Logistic Regression**		
	ΔMH	p	DIF Level	R^2	p	DIF Level
Item 7	-3.63	.02	C	.070	.00	C
Item 13	12.68	.00	C	.047	.00	B

As seen in Table 4, the DIF analyses results showed that two items (7 and 13) sizable (C level) DIF based on Mantel Haenszel and logistic regression methods. In Mantel Haenszel method, Dorans and Holland's (1993) effect size and in Logistic regression method Gierl, Khaliq and Boughton's (1990) DIF cut points are used.

Results of Equating

In this study, five equating methods are considered: Tucker linear equating, Levine observed score equating, chained equipercentile equating no smooth, chained equipercentile equating with presmoothing (C=4), and simplified circle arc equating methods with nominal weights. As mentioned before for the polynomial log-linear presmoothing method, choosing the degree of the polynomial (C) is important. For this study, we compare chi-square values under different smoothing parameter. The moments and fit statistics for presmoothing are given in Table 5.

Table 5. The Moments and Fit Statistics

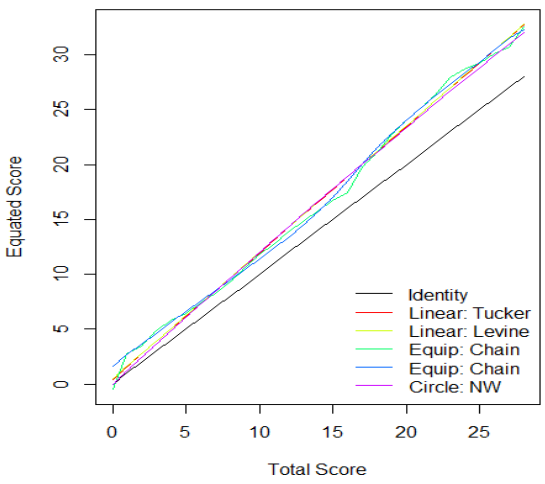
Form	Smoothing Parameter	$\bar{\mu}$	$\bar{\sigma}$	$\overline{sk}$	$\overline{ku}$	$X^2(df)$	$X_C^2 - X_{C+1}^2$
Booklet 1	C=5	14.51	8.22	.57	1.91	16.12 (27)	2.38
	C=4	14.51	8.22	.57	1.91	16.93 (28)	0.81
	C=3	14.51	8.22	.57	2.42	76.15 (29)	59.22
	C=2	14.51	8.22	.16	2.13	105 (30)	28.97
	C=1	14.51	9.45	.19	1.85	120 (31)	15.63
Booklet 14	C=5	12.60	7.13	.28	1.74	15.43 (27)	0.68
	C=4	12.60	7.13	.28	1.74	15.73 (28)	0.29
	C=3	12.60	7.13	.28	2.43	75.63 (29)	59.90
	C=2	12.60	7.13	.31	2.47	75.78 (30)	0.15
	C=1	12.60	9.15	.44	2.06	121.88 (31)	46.10

Note. The chosen C parameter for presmoothing is shown in boldface.

As seen in Table 5, for Booklet 1, C=4 the overall X^2 statistic is not significant ($X^2_{(28)}$ =16.93, p>.05) and the difference statistics for chi-square $X^2_{C=}4$-$X^2_{C=5}$ equals .81 and it is not significant at .05 level for one degree of freedom (X^2<3.84). Based on results, for C≥4 model fit the data and C=5 do not improve the fit of data. For Booklet 14, as seen in Table 5, C=4 the overall X^2 statistic is not significant ($X^2_{(28)}$ =15.73, p>.05) and the difference statistic for chi-square $X^2_{C=4}$-$X^2_{C=5}$ equals .29 and it is not significant at .05 level for one degree of freedom (X^2<3.84). Based on results, again for C≥4 model fit the data and C=5 do not improve the data fit. For Booklet 1 and Booklet 14, C=4 is chosen for presmoothing.

In this study, Booklet 1 is a base form and Booklet 14 is a new form and test equating is conducted under CINEG design. Test equating is done in two phases: the presence of DIF items in the anchor test and removing sizeable DIF items from anchor test. The Figure 1 shows that equated scores versus total scores in the presence and absence of sizeable DIF items in the anchor test.

a) DIF items in anchor test



b) After removed DIF items from anchor test

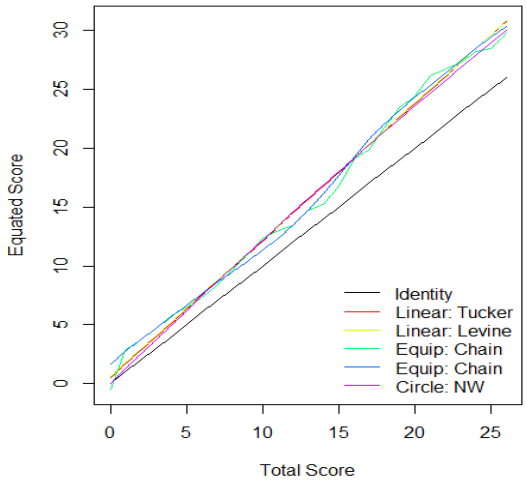


Figure 1. Equated Scores versus Total Scores

In Figure 1, the red line represents Tucker linear equating, the yellow line represents Levine observed score equating, the green line represents chained unsmoothed equipercentile equating, the blue line represents chained presmoothed (C=4) equipercentile equating and the purple line represents simplified circle arc equating method. Also in Figure 1, the black line belongs to identity equating and it means that there is no equating between old form and new form scores. As seen for both conditions Tucker, Levine and circle arc methods yield similar equated scores; their lines in the graphics are almost top of each other. In unsmoothed chained equipercentile equating method, there are some irregularities between equated scores and total scores (see the green line in Figure 1). The random error in estimating equivalent scores causes these irregularities. As seen in Figure 1, these irregularities got lost in presmoothed equipercentile equating (see the blue line).

The performance of different equating methods in presence and absence of DIF items in anchor test was evaluated based on standard errors of equating, bias and RMSE values which provided from 1000 bootstrapped samples and reported in Table 6.

Table 6. Equating Results

Equating Methods	Presence of DIF Items in Anchor Test			Absence of DIF Items in Anchor Test		
	se	Bias	RMSE	se	Bias	RMSE
Tucker Linear Equating	0.36	0.52	0.63	0.35	0.52	0.63
Levine Observed Score Equating	0.37	0.52	0.63	0.39	0.51	0.64
Unsmoothed Chained Equipercentile Equating	0.77	0.54	0.94	0.72	0.61	0.94
Presmoothed (C=4) Chained Equipercentile Equating	0.49	0.19	0.52	0.51	0.17	0.53
Simplified Circle - Arc Equating	0.19	0.69	0.72	0.19	0.72	0.74

As seen in Table 6, when the common items include DIF items, simplified circle arc equating method has the least (.19) standard error of equating (se) and unsmoothed chained equipercentile equating method has the largest (.77) standard error of equating. On the other hand, when we consider bias as a criterion simplified circ-arc method has the largest (.69) amount of bias, 4-moments presmoothing chained equipercentile equating has the smallest amount of bias value. Levine and Tucker equating methods have the same (.52) and smaller bias values than unsmoothed equipercentile equating method. We can say that Tucker linear equating and Levine observed score equating methods show similar and better performance than the unsmoothed chained equipercentile equating method. According to last criteria of RMSE values, again smoothed chained equipercentile equating method has the smallest (.52) RMSE value and the unsmoothed equipercentile equating method has the largest (.94) RMSE value. Tucker linear equating and Levine observed score linear equating methods had the same (.63) and smaller RMSE values than simplified circle-arc equating method (.72). Again we can say that Tucker linear equating and Levine observed score linear equating methods show similar and better performance than the simplified circle-arc equating method.

After removing two sizeable DIF items from anchor test, the similar results have been found (See Table 6). Again based on se criteria, the simplified circle arc method was the best and the unsmoothed chain equipercentile equating method was the worst. On the other hand, based on bias criteria the best equating method is presmoothed equipercentile equating method and the worst one is simplified circle arc equating method. Concerning RMSE values, the best one is again presmoothed chained equipercentile equating method and the worst one is unsmoothed chained equipercentile equating method. Also, according to results, we can say that performances of equating methods are similar with the presence and not presence of DIF items in anchor test and we can say that there is no notable change in se, bias and RMSE values.

Another result of this study is that whether or not common items include DIF items, unsmoothed chained equipercentile equating method has larger se, bias and RMSE values than presmoothed (C=4) chained equipercentile equating method.

CONCLUSION AND DISCUSSION

In this study, five equating methods were considered: Tucker linear equating, Levine observed score linear equating, unsmoothed chained equipercentile equating, chained equipercentile equating with presmoothing (C=4), and simplified circle arc equating methods with nominal weights. Equating methods were compared in two phases: the presence of DIF items in anchor test and removing sizeable DIF items from anchor test. The results show that performances of equating methods are similar to presence and absence of DIF items in anchor test and there is no notable change in se, bias and RMSE values. Also, results show that according to the standard error of equating criteria, the circle arc equating method outperformed other equating methods but based on bias evaluation criteria its performance was the worst one in both situations.

As Kolen and Brennan (2004) reported standard error of equating is the standard deviation of equivalent scores over replications of the equating process and random error indexed by the standard error of equating. Standard error equating is closely related with sample size and as the sample size becomes larger it becomes smaller. The result of this study showed that the circle arc method has the minimum se among other equating methods. The circle-arc method especially suggested for small samples (Livingston, & Kim, 2009) and in their study, Kim and Livingston (2010) showed that in small samples the circle arc method clearly outperformed other equating methods (chained equipercentile, Levine, chained linear, Tucker) based on bias, RMSD and, se evaluation indexes. The results of our study supported Kim and Livingston (2010) only in terms of random equating error. Also, among other equating methods the unsmoothed chained equipercentile equating has the largest se value and we can say that based on random equating error its performance was the worst. As seen in Figure 1 there are some irregularities between equated scores and total scores and Kolen and Brennan (2004) noted that the reason for these irregularities is random equating error. Also in our study, the results of unsmoothed chained equipercentile equating method based on se support this view.

Based on bias and RMSE evaluation criteria, smoothed chained equipercentile equating method is the best equating method and unsmoothed chained equipercentile equating method is the worst method. In one of a simulation study, Aşiret and Sünbül (2016) shows that for sample size 200 presmoothed equipercentile equating method produced more accurate results than other methods (linear, circle-arc, mean). Our study results supports this finding with real data which has sample size roughly 200. Tucker linear equating and Levine observed score equating methods show similar and better performance than the unsmoothed chained equipercentile equating method. To all evaluation indexes, smoothed chained equipercentile equating has lower values than the unsmoothed equipercentile equating method. We can say that presmoothing tended to decrease random and systematic equating error as in shown other studies (Aşiret & Sünbül, 2016; Özdemir, 2017; Kelecioğlu & Öztürk Gübeş, 2013; Livingston, 1993; Skaggs, 2005). As a result, we can also say that presmoothed chained equipercentile equating yields more precise and accurate equating results than unsmoothed chained equipercentile equating as is assumed (Kolen and Brennan, 2004).

RECOMMENDATIONS

This study is limited with 8th-grade mathematics data from Booklet 1 and Booklet 14 in TIMSS 2015 Turkey sample. The results showed that performances of equating methods are similar to the presence and not the presence of DIF items in anchor test and there is no notable change in the error of equating. This result should be interpreted carefully and in further researches effects of DIF on small sample equating methods should be examined with real and simulated data sets more detailed.

Acknowledgements or Notes

This study was presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology, KOSOVA.

REFERENCES

Albano, A. (2017). *equate: Observed –score linking and equating.* [Computer software].

Alexeev, N., Templin, J., & Cohen, A. (2011). Spurious latent classes in mixture rasch model. *Journal of Educational Measurement*, *48*(3), 313-332.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Kuram ve Uygulamada Eğitim Bilimleri*, 16(2), 647-668.

Atalay-Kabasakal, K. & Kelecioğlu, H. (2015). Effect of differential item functioning on testequating.*Educational Sciences: Theory & Practice, 15*(5), 2015, 1229-1246.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement, 72*(4), 608-628.

Byrne, B. M. (2010). *Structural equation modeling with AMOS*, (2nd ed.). New York: Routledge.

Chu, K. L. (2002). *Equivalent group test equating with the presence of differential item functioning* (Doctoral dissertation). Available from ProOuest Dissertations and Theses database.

Cohen, A.S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning *Journal of Educational Measurement, 42*(2), 133-148.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

De Ayala, R.J., Kim, S.H., Stapleton, L.M., & Dayton, C.M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing,3*(4), 243-276.

Demirus, K. B., & Gelbal, S. (2016). The study of the effect of anchor items showing or not showing differential item functioning to test equating using various methods. *Journal of Measurement and Evaluation in Education and Psychology 7*(1), 182-201.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*(1), 3-17.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NH: Erlbaum.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS Research Report No. RR-10-29). Princeton, NJ: ETS.

Elosua, P., & Hambleton, R. K. (2018). Psychological and educational test scorecomparability across groups in the presence of item bias. *Journal of Psychology and Education, 13*(1), 23-32.

Fieuws, S., Spiessens, B., & Draney, K. (2004). Mixture models. In P. de Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. (pp.317-340). New York: Springer.

Gierl, M., Khaliq, S. N., & Bougthon, K. (1999). Gender differential item functioning in mathematics and science: Prevalence and policy implications. Paper presented at the *Improving large-scale assessment in education*. Symposium conducted at the Annual Meeting of Canadian Society for the Study of Education, Canada.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the lord statistic. *Educational and Psychological Measurement, 62*(1), 32–44.

Hu, L. & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Karasar, N. (2009). *Bilimsel araştırma yöntemi: Kavramlar, ilkeler, teknikler.* Ankara: Nobel Yayınları.

Kelecioğlu, H., & Öztürk Gübeş, N. (2013). Comparing linear equating and equipercentile equating methods using random groups design. *International Online Journal of Educational Sciences*, *5*(1), 227-241.

Kim, S. & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common item design. *Journal of Educational Measurement, 47*(3), 286-298.

Kline, R. (2005). *Principles and practices of structural equation modeling* (2n ed.). New York: Guilford Press.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and Practices*. New York: Springer Verlag.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Kurtz, A. M., & Dwyer, A. C. (2013). *Small sample equating: Best practices using a SAS Macro*. Retrieved from http://analytics.ncsu.edu/sesug/2013/BtB-11.pdf

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*(5), 353-373. doi: 10.1177/0146621608326422

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

Livingston, S. A., & Kim, S. (2008). *Small sample equating by the circle-arc method* (ETS Research Report No. RR-08-39). Princeton, NJ: ETS

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*(3), 330–343.

Magis, D., Beland, S., & Raiche, G. (2015). *difR: Collection of methods to detect dichotomous differential item functioning (DIF)*. [Computer software].

Mclachlan, G. & Peel, D., (2000). *Finite Mixture Models.* John Wiley & Sons, Inc. New York.

Mislevy, R. J. & Norman, V. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195-215.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Oliveri, M. E., Ercikan, K. Zumbo, B. (2013). Analysis of Sources of Latent Class Differential Item Functioning in International Assessments. *International Journal of Testing*, *13*(3), 272–293. doi: 10.1080/15305058.2012.738266

Özdemir, B. (2017). Equating TIMSS mathematics subtests with nonlinear equating methods using NEAT design: circle-arc equating approaches. *International Journal of Progressive Education, 13*(2), 116-132.

Parshall, C. G., Du Bose, P., Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*(1), 37–54.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271-282. doi: 10.1177/014662169001400305

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. (Doctoral dissertation, Faculty of Graduate School of the University of Maryland, College Park). Retrieved from https://drum.lib.umd.edu/bitstream/handle/1903/2682/umi-umd-2604.pdf?sequence=1&isAllowed=y

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309–330.

Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning* (Doctoral Dissertation). Available from ProOuest Dissertations and These database.

Von Davier, M. (2001). *WINMIRA* [Computer Software]. Groningen, the Netherlands: ASCAssessment Systems Corporation. USA and Science Plus Group.

Yurtçu, M. & Güzeller, C.O. (2018). Investigation of Equating Error in Tests with Differential Item Functioning. *International Journal of Assessment Tools in Education*, *5*(1), 50-57.

Career Planning Scale of Students Studied in Sports Sciences (CPS): Validity and Reliability Study

Sultan Yavuz Eroglu [i]
Siirt University

Erdem Eroğlu [ii]
Siirt University

Abstract

The purpose of this study was to develop a scale towards career planning of sports sciences students. Study group consisted of 543 students who were attending in physical education and sports teaching, sports management, and coaching departments in Siirt University. Construct validity of scale was tested through factor analysis and confirmatory factor analysis. Reliability of scale was measured through Cronbach Alpha and test-retest test. Discrimination of scale was tested between down %27 and up %27. Correlation analysis was made between scale factors. In order to calculate the reliability of 23 items in Career Planning Scale, Cronbach Alpha which is a inner consistency coefficient was calculated. General reliability of scale were found to be high as alpha=0.885. Analysis results have demonstrated that adjustment statistics calculated through confirmatory factor analysis showed a significant adjustment and positive correlations were determined between scale sub-dimensions and general scores as a result of correlation analysis ($p<0,05$).

Keywords: Career, Career Planning, Sports, Physical Education and Sport

DOI: 10.29329/ijpe.2020.248.9

---

[i] **Sultan Yavuz Eroglu,** Assoc. Prof. Dr., Physical Education and Sports, Siirt University, ORCID: 0000-0001-5875-2836

[ii] **Erdem Eroğlu,** Assist. Prof. Dr., Physical Education and Sports, Siirt Univerity, ORCID: 0000-0002-6301-9257

Correspondence: erdemeroglu@siirt.edu.tr

INTRODUCTION

The term ‘’career’’ is used in the meaning of acquiring skills, continuous and step by step progression in any field of interest of an individual (Tortop, 1994: 92). Career is a series of works which have continuity for a lifetime equipped with behaviour motives of an individual and is to gain prestige and power, having a better status, and earning money as a result of breaking through in a selected work area (Bayram, 2008:19).

In the stage of beginning of a career (explore), individuals tend to have knowledge of about the careers and professions they are interested in by trying to define what kind of skills they had. In accordance with these, they follow the education process related to professions they are interested in. Therefore, this stage is a process that will go on after starting to work. After starting to work, individuals who are aware of their responsibilities, appear to be important with regard both to make a progress about their careers and to reach to company targets easily.

Individuals can set goals for themselves and exert efforts to reach these goals during university education. In accordance with that if a student can get to know himself-herself better, he/she will enable himself/herself to find a job that is satisfactory.

Sports phenomenon is a process which is a continuing and open for improvement. Thus, while career goals are being defined, it is thought that sports departments are preferred due to both being interested in sports and having a sportsmenship identity in the past.

Sports education/training, is an education process including hypothetic and applied education. During this education process, while student is given theoretical knowledge for his/her profession, they are tried to turn these knowledges into skills in their fields of profession. One of the crucial and initial conditions in developing quality in sports education is the quality of education. Reaching the targets in sports education is depended on using international rules and methods in it.

When considered as a clear approach, an organization process must be established to support and theoretical and applied knowledge with experiences and following innovations all the time. To tell sports education apart from general education, will make targets to reach impossible. Essentially the purpose of education, is to raise qualified human power through education. A qualified individual is well-developed both physically and mentally and who have a proficiency in getting in touch socially within society he/she lived. In this sense, physical education and sports is highly effective in raising qualified individuals (Kızılet, 2018).

As a result of education given in physical education and sports faculties, students can become sports specialist, physical education teacher, coach and also can found an enterprise in sports business. They are expected to have their career plans made in order to be beneficial to society and make difference in these areas during preparation to physical education schools or in the beginning of their education.

For example, as different comparing to other branch teachers; the responsibilities of physical education teachers are not confined to weekly course hours. Physical education teachers also busy themselves with relations with people and other teachers and administrators in addition to extra-curricular activities, need more time for these responsibilities (Altuntaş, 2016).

Sports phenomenon is a social structure where competition environment exists. It has also a significant place in educating and training mind and body. Therefore raising physically and mentally healthy individuals, and raising elite athletes with regard to performance; will contribute positively to people and country.

For this reason, for physical education teachers, knowing purposes well and evaluate them will make physical education teachers contribute to country as sports specialist, coach, educator who

are able to evaluate the performance, whose communication skills are perfect and who understand athlete's psychology after their graduation from school. But if physical education department students would not define their purposes in advance i.e. if they would not target being a good coach, educator and sports specialist and intend only to graduate from their schools, this will cause them not to fullfil their professions better. So this decision must have been given in early years of university education.

Therefore, beginning of a career (explore) consitutes our studies' general framework. While exploring process lasts in mid-twenties, their school life ends and working life starts. An individual mostly in a struggle to understand himself/herself; after that, an individual consider his/her own conditions, and make research related to what kind of job they will be successful.

They determine their weakness and strengths. One of the most important expectations of individuals in the stage of exploring while entering into working life is a long term and productive career. Exploring stage is a very important stage for both individuals and organizations. Because, meeting the expectations of organizations and needs of individuals occur in this stage. In this study which we purposed to measure how they describe the adequacy of the education they took, profession, and expectation they had, it is important for students to select the right professions. As a result of literature search, some studies were reviewed related to career goals and planning. But when considered in terms of sports, there were no any scales found out of athletes career scale. Therefore, to develop a scale towards career planning of sports sciences students and contribute to this field in accordance with this, underlies our research.

METHOD

Research Model

This research was designed in descriptive survey model. As developing a scale was first intended, determining the properties to be tested, writing the items for scale, taking expert opinion and rearranging items, and making validity and reliability stages were followed in research (Cronbach, 1984; Altun ve Büyüköztürk, 2011).

Study Group

Study group consisted of 543 students who were attending in physical education and sports teaching, sports management, and coaching departments in Siirt University. Desciptive statistics related to study group were given in Table 1 below.

Table 1. Desciptive statistics

Groups	Frequency(n)	Per cent (%)
Gender		
Female	184	33,9
Male	359	66,1
Marital Status		
Married	21	3,9
Single	522	96,1
Department		
Physical Education Teaching	159	29,3
Sports Management	311	57,3
Coaching	73	13,4
Class		
1	166	30,6
2	179	33,0
3	87	16,0
4	111	20,4

Education Level of Mother		
Illiterate	241	44,4
Primary School	151	27,8
Secondary School	70	12,9
High School	33	6,1
Associate Degree	6	1,1
Bachelor's Degree/Undergraduate	40	7,4
Graduate	2	0,4
Education Level of Father		
Illiterate	82	15,1
Primary School	166	30,6
Secondary School	108	19,9
High School	125	23,0
Associate Degree	11	2,0
Bachelor's Degree/Undergraduate	42	7,7
Graduate	9	1,7

Students were distributed as 184 (33,9) female, 359 (66,1) male according to gender and Marital status; 21 (3,9%) married, 522 (96,1%) single. Departments; 159 (29,3%) physical education and sports teaching, 311 (57,3%) sports management, 73 (13,4%) coaching. Classes; 166 (30,6%) 1st class, 179 (33,0%) 2nd class, 87 (16,0%) 3rd class, 111 (20,4) 4th class. Education level of mother; 241 (44,4%) illiterate, 151 (27,8%) primary school, 70 (12,9%) secondary school, 33 (6,1%) high school, 6 (1,1%) associate degree, 40 (7,4%) bachelor's degree/undergraduate, 2 (0,4%) graduate. Education level of father; 82 (15,1) illiterate, 166 (30,6%) primary school, 108 (19,9%) secondary school, 125 (23,0%) high school, 11 (2,0%) associate degree, 42 (7,7%) bachelor's/undergraduate, 9 (1,7%) graduate.

Data Collection Tools

In the first stage of developing the scale, literature associated with career and career planning were reviewed and the importance of career planning were tried to determine. Based on related literature search, a 35 item, item repository was established. Trial form of scale established was broached to 5 experts from Physical Education and Sport Department and 3 from Faculty of Economics and Administrative Sciences.

Experts have evaluated the suitability of career planning for properties of physical education and sports students and clarity of items. In accordance with expert opinions, 5 items were omitted from scale and some were tried to be corrected. Following this correction, a 30 item trial form was composed. Participants were asked express their opinions as ''Strongly agree'', ''agree'', ''undecided'', ''disagree'', ''strongly disagree'' in a 5 likert type form. Before scale was applied to students, explanations were made to participants related to career planning and the purpose of research was explained. Application of scale lasted in 10 minutes.

Statistical Analysis

Construct validity of scale was tested through exploratory factor analysis and confirmatory factor analysis. Reliability of scale was measured through Cronbach Alpha and test-retest test. Discrimination of scale was tested between down %27 and up %27. Correlation analysis was made between scale factors.

FINDINGS

In order to reveal construct validity of scale, exploratory factor analysis method was applied. As a result of Barlett test, ($p=0.000<0.05$) a relationship was found between variables included in factor analysis. As a result of Kaiser-Mayer-Olkin test, ($KMO=0.904>0,60$) it was determined that sample size was adequate for factor analysis to apply. In applying factor analysis varimax method was selected and construct of relationship between factors were provided to stay same. As a result of factor analysis, variables were collected under 5 factors which had 55.496% variance.

Items 6, 8, 9, 19, 20, 22, 28, were omitted as factor loading was under 0,4. In order to calculate reliability of 23 items in career planning scale; internal coefficient''Cronbach Alpha'' was calculated. General reliability of scale was found very high with alpha=0.885. According to explained variance ratio and alpha coefficient it can be said that Career Planning Scale is reliable tool. Factor construct was shown below.

Table 2. Construct of Career Planning Scale

Dimension	Factor Load
Career Awareness ((Eigen value=7,243; Explained variance=14,973; Alpha=0,833)	
4- I want to create differences for the company i work for and be dynamic.	0,653
10- I determine my career goals according to my interests and skills	0,651
11- I think that i focus on my goals according to my career plans	0,647
5- I am aware the way i will follow in order to reach my career goal	0,644
13- While planning my career, I know that not only i will get promotion hierarchically but also i will improve my skills.	0,604
14- My sportive skills are pathfinders for making a career planning for me.	0,490
12- I know positive and negative sides of place I work, i searched it.	0,486
18- I think that i am able to overcome the obstacles that will come in my career way	0,455
1-I am aware of my weak and strong sides and skills	0,433
Professional Awareness (Value=1,925; Variance determined=12,661; Alpha=0,780)	
16- I am aware of my profession's progress facilities	0,757
15- I have knowledge of the future of my profession	0,697
17- I am aware of knowledge and skills asked in my profession	0,692
23- I follow the events related to my profession	0,475
Faith Towards Career (Value=1,341; Variance determined= 11,688; Alpha=0,784)	
30- Thinking about my career inspires me	0,727
29- I believe that i will overcome any obstacles will come in my way in reaching to my career	0,698
27- I know career planning is important for being successful in my profession	0,633
26- I think that i am compatible with the profession that i chose	0,617
Accuracy of Selection (Value=1,211; Variance Determined= 8,143; Alpha=0,649	
2- the department i study enables to plan my career and improve it and make me reach my goal	0,817
3- Position i selected, made me to build my career plan	0,803
7- I think that making a career planning, specifies the choices and resolves uncertainties	0,431
Education Proficiency (Value=1,044; Variance Determined= 8,031; Alpha=0,641	
25- Facilities in school, are adequate to actualize my career planning	0,802
24- I participate in seminars, courses and symposiums for my career	0,738
21- I think that the education i received is adequate to reach my career goals	0,678
Total Variance=55.496%; General Reliability (Alpha)=0.885	

Factor construct obtained in exploratory factor analysis of scale was tested through confirmatory factor analysis. Diagram is given related to confirmatory factor analysis below

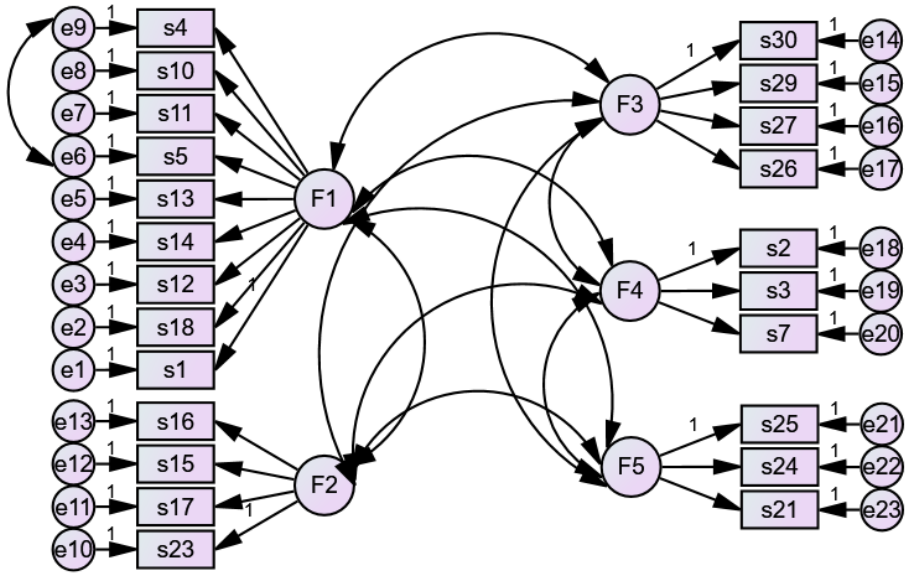


Figure 1. Diagram related to Confirmatory Factor Analysis

Criterion related to Confirmatory Factor Analysis given below

Table 3. Confirmatory Factor Analysis Model Fit Values for CPS

Index	Normal Value*	Acceptable Value**	CPS
$\chi 2$/sd	<2	<5	2.42
GFI	>0.95	>0.90	0.92
AGFI	>0.95	>0.90	0.90
CFI	>0.95	>0.90	0.92
RMSEA	<0.05	<0.08	0.05
RMR	<0.05	<0.08	0.05

*, ** References: (Şimşek, 2007; Hooper and Mullen 2008; Schumacker and Lomax, 2010; Waltz, Strcikland and Lenz 2010; Wang and Wang, 2012; Sümer, 2000; Tabachnick ve Fidel, 2007).

As a result of analysis; it was determined that model fit statistics calculated with factor analysis were seen to be fit. Standardised factor loadings, t values and explained variance ratios ($\boldsymbol{R^2}$) were given below in Table 4.

Table 4. Confirmatory Factor Analysis Factor Loadings and Regression Coefficients related to items

Items		Factors	β	Std. β	S.Error	t	p	R^2
s1	<---	F1	1,000	0,511				0,487
s18	<---	F1	1,453	0,628	0,141	10,276	p<0,001	0,483
s12	<---	F1	1,347	0,517	0,148	9,124	p<0,001	0,569
s14	<---	F1	1,377	0,609	0,136	10,103	p<0,001	0,566
s13	<---	F1	1,309	0,529	0,141	9,258	p<0,001	0,610
s5	<---	F1	1,526	0,619	0,150	10,173	p<0,001	0,473
s11	<---	F1	1,564	0,667	0,147	10,620	p<0,001	0,599
s10	<---	F1	1,623	0,679	0,151	10,723	p<0,001	0,562

s4	<---	F1	1,387	0,607	0,138	10,052	p<0,001	0,568
s23	<---	F2	1,000	0,549				0,504
s17	<---	F2	1,185	0,744	0,100	11,883	p<0,001	0,574
s15	<---	F2	1,228	0,737	0,104	11,819	p<0,001	0,543
s16	<---	F2	1,262	0,757	0,105	11,987	p<0,001	0,554
s30	<---	F3	1,000	0,710				0,522
s29	<---	F3	1,035	0,753	0,068	15,320	p<0,001	0,513
s27	<---	F3	1,006	0,750	0,066	15,257	p<0,001	0,461
s26	<---	F3	0,850	0,568	0,072	11,877	p<0,001	0,445
s2	<---	F4	1,000	0,688				0,538
s3	<---	F4	1,052	0,781	0,094	11,219	p<0,001	0,533
s7	<---	F4	0,583	0,447	0,068	8,523	p<0,001	0,577
s25	<---	F5	1,000	0,754				0,603
s24	<---	F5	0,734	0,553	0,089	8,288	p<0,001	0,544
s21	<---	F5	0,695	0,553	0,084	8,289	p<0,001	0,597

When standardised coefficients were examined, it was determined that factor loadings were high, standart error values were low, t values were significant (p<0,001), R^2 values were high. These results confirm construct validity related to factor construct determined previously.

Discrimination of scale were analysed with the difference between 27% lower and 27% upper groups (Table 5). According to analysis findings, it was determined that scale was able to distinguish upper and lower group with sub-dimensions (p<0,05).

Reliability results of scale were given in Table 5. Reliability of scale depended on time were provided (p>0,05).

Table 5. Discrimination of Scale and Test- Retest Findings

	Discrimination				**Test-Retest**	
	Lower 27% Avg±Sd)	**Upper 27% (Avg±Sd)**	**t**	**p**	**t**	**p**
Career Awareness	3,481±0,639	4,697±0,234	-21,674	**0,000**	0,357	0,722
Professional Awareness	3,357±0,732	4,738±0,316	-21,016	**0,000**	-1,054	0,295
Faith in Career	3,456±0,816	4,799±0,271	-18,947	**0,000**	1,482	0,142
Accuracy of Choice	3,243±0,751	4,456±0,521	-16,096	**0,000**	-0,820	0,414
Education Proficiency	2,696±0,810	3,862±0,906	-11,625	**0,000**	-0,292	0,771
General Career Planning	3,322±0,437	4,582±0,153	-32,998	**0,000**	-0,413	0,681

Table 6. Average Scores and Correlation Matrix

	Average	Standart Deviance	Career Awareness	Career Awareness	Career Awareness	Career Awareness	Career Awareness	Career Awareness
Career Awareness	4,154	0,615	1,000					
Professional Awareness	4,129	0,744	0,649**	1,000				
Faith in Career	4,222	0,750	0,610**	0,615**	1,000			
Accuracy of Choice	3,822	0,791	0,436**	0,360**	0,354**	1,000		
Education Proficiency	3,228	0,988	0,166**	0,186**	0,171**	0,299**	1,000	
General Career Planning	3,997	0,538	0,875**	0,794**	0,772**	0,631**	0,457**	1,000

*<0,05; **<0,01

"Career awareness" of students was average 4,154±0,615 (Min=1.67; Max=5), "professional awareness" was average 4,129±0,744 (Min=1; Max=5), "faith towards career" was average 4,222±0,750 (Min=1; Max=5), "accuracy of choice" was average 3,822±0,791 (Min=1; Max=5), "education proficiency" was average 3,228±0,988 (Min=1; Max=5), "career planning general" was average 3,997±0,538 (Min=1.7; Maks=5).

As a result of correlation matrix; positive correlations were determined between scale sub-factors and general scores ($p<0,05$).

RESULTS AND FUTURE RECOMMENDATIONS

As a result of education given in sports science faculties, students are able to establish a company as entrepreneur, coach, physical education teacher or sports specialist in sports business. In order to be the best, create difference and being beneficial to society, they are expected to be made their career planning when they were attended in a physical education department and started to receive education.

However, as students attended in physical education and sports departments work in other business sectors, made us think that there might be issues related to career planning. In addition, as there was no any valid and reliable measurement tool towards career planning, demonstrated that there was a need for such a tool. In accordance with this, in this study we have carried out, it was aimed to develop a scale to determine career plannings of physical education and sports students and to test validity and reliability of it.

In this study, 23 items and 5 sub-factor scale were developed consisting of Career Awareness, Professional Awareness, Faith towards Career, Accuracy of Choice, Education Proficiency in order to determine career planning features of sports science students. Carreer awareness dimension consisted of 9 item, Professional awareness dimension; 4 item, faith towards career dimension; 4 item, accuracy of choice dimension; 3 item, and education proficiency dimension consisted of 3 item. Analysis results toward construct validity of scale revealed that scale items had acceptable level of factor loading and scale was in a five factor form.

It was also revealed that internal consistency was generally on acceptablelevel. There was a significant relationship between the dimensions of scale. These results demonstrated that **CPS**'s validity and reliability was on a sufficient level. It was thought that with this shape of the scale can be used to determine career planning features of sports sciences students, and also thought to contribute to literature as a valid and reliable measurement tool. It can be recommended that this study can be carried out with larger groups as it has limited groups in current study. The scale was also recommended to be used for sports high school students and graduate students as well as using it for undergraduate students by repeating their validity and reliability studies.

REFERENCES

Altuntaş EA. (2016). Beden eğitimi öğretmeni adaylarının öğretmenlik mesleğine ilişkin tutumları ile öz yeterlikleri arasındaki ilişki. Bartın Üniversitesi, Eğitim Bilimleri Enstitüsü,

BAYRAM, C.(2008). Kariyer Planlama ve Yönetimi, Kum Saati Yayın Dağıtım LTD. ŞTİ., İstanbul.

Brown, T.A. (2006). Confirmatory Factor Analysis for Applied Research. The Guilford Press, New York, USA.

Büyüköztürk, Ş. (2007). *Sosyal Bilimler için Veri Analizi El Kitabı*, Ankara: Pegem A Yayıncılık.

Cronbach, L. J. (1984). Essentials of psychological testing (4th ed), New York: Harper Row. *Journal of Educational Measurement*, 23 (2). 175-183.

Hooper D, Coughlan J, Mullen MR. Structural Equation Modelling: Guidelines for Determining Model Fit. Electronic Journal of Business Research Methods 2008; 6(1): 53-60.

Hox, J. J., & Bechger, T. M. (1998). An Introduction to Structural Equation Modeling. Family Science Review, 11, 354–373

Kızılet, T. (2018). Drama Ve Diksiyon Öğretiminin Spor Eğitimcisi Adayı Öğrencilerin Mesleki Yeterliliklerine Etkisi. Yüksek Lisans Tezi. İstanbul.

Mels G. (2006), "LISREL for Windows: Getting Started Guide", http://www.ssicentral.com/lisrel/techdocs/GSWLISREL.pdf

Schumacker RE, Lomax RG. A Beginner's Guide to Structural Equation Modeling. New Jersey: Taylor & Francis; 2004. p.1-8.

Sümer, N. (2000). Yapısal Eşitlik Modelleri. Türk Psikoloji Yazıları. No.3, S.6, 49-74.

Şimşek ÖF. Yapısal Eşitlik Modellemesine Giriş, Temel İlkeler ve LİSREL Uygulamaları. Ankara: Ekinoks; 2007. p.4-22.

Tabachnick, B. G. and Fidel, L. S. (2007). Using Multivariate Statistics. Pearson Education Inc. Boston.

Tortop N. (1994). Personel Yönetimi, 5. Basım, Ankara: Yargı Yayınları.

Waltz CF, Strcikland OL, Lenz ER. Measurement in Nursing and Health Research. New York: Springer Publishing Company; 2010. p.176-8.

Wan, T. T. (2002). Evidence-based health care management: Multivariate modeling approaches. Springer: Netherlands.

Wang J, Wang X. Structural Equation Modeling: Applications Using Mplus: methods and applications. West Susex: John Wiley & Sons; 2012. p.5-9.

International Journal of Progressive Education, Volume 16 Number 3, 2020

Why Do Students Prefer Different Question Types?

Nihan Sölpük Turhan [i]
Fatih Sultan Mehmet Vakif University

Abstract

Measurement tools that are used in education are important factors that affect course success and motivation of students. This study aims to determine the opinions of high school students on different question types. As the subgoals of the research, the study aims to determine the reasons for multiple choice test preference and its effect on learning motivation level according to the grade. Study group consists of 355 students who are 10^{th}, 11^{th} and 12^{th} graders in state schools in Istanbul province center in spring term of 2018-2019 school year. Mixed method and convergent parallel design were utilized for the study. "Academic Motivation Scale (AMS)" that was developed by Bozanoğlu (2004) and "Inventory of Motives to Prefer Written, Short-Answer, True-False and Multiple-Choice Questions (IMP)" that was developed by Eser (2011) were used for data collection in the study. Interview method was utilized to determine the opinions of teachers on test types. Therefore, semi-structured interview form was prepared as a data collection tool. Data analysis was made by using Multivariate Analysis of Variance (MANOVA). As a result, the study found that the motives to prefer multiple-choice questions and averages of learning motivation vary significantly in favor of 10^{th} grade students and final year students in high school. The study revealed that student performance varies by question type. The study also found that multiple-choice questions can be considered as a motivation factor for high school students and a good way of testing the goals and achievements.

Keywords: High School Students, Learning Motivation, Mixed Method, Question Types

DOI: 10.29329/ijpe.2020.248.10

---

[i] **Nihan Sölpük Turhan,** Assoc. Prof. Dr., Educational Science, Fatih Sultan Mehmet Vakif University, ORCID: 0000-0002-9279-7699

Correspondence: nsolpuk@fsm.edu.tr

INTRODUCTION

Exam is a method that is applied to measure the accumulation of knowledge and development of students in the education process. Nowadays, it has become an obligation in some way and also one of the efficient methods particularly in receiving feedbacks for the results of the education system. Students encounter exams at all stages of their education life beginning from primary school. In addition to this, exams continue for the whole life after university education is complete. For this reason, exams have an important place in people's life. Considering such an important factor in detail, we encounter different types of exams. In a study on students, student motivation is split into four types: external, internal, social and achievement motivation (Jenkins,2001).

Common question types encountered in the education system are classified as written, multiple-choice, open-ended, gap-filling, matching, classification, true-false and short-answer questions. Students may have different opinions on these exam types. According to Kılıç and Çetin (2018), one of the reasons for the students to have different opinions on the exam types is that exam type preferences of students may be affected by the difference in their strengths and weaknesses. Besides, it is emphasized that exam type preferences of students may vary by their understanding of learning or the level of exam anxiety. In addition, success of students is largely attributed to their awareness of their own learning style and being able to direct their learning. While some students think that they express themselves better in open-ended questions, others think that they find the definitive answer in multiple-choice questions. Among such exam types, multiple-choice and open-ended questions come to the forefront.

While students choose the answer among the predetermined options in multiple-choice exam questions, there is no such option for the answer in open-ended questions. The answer totally depends on the knowledge and skills of students. Therefore, differences of opinion particularly on these two types have arisen among the students; while some students turn towards multiple-choice exam type, others consider open-ended questions suitable (Gronlund, 1998). Clarke et al. (2005) began studying the research question "Is there any difference between student preferences and evaluation types?" and verified the fact that there is a significant difference in student preferences and MCQ test is the most preferred option on average.

Academic motivation is an effective factor that determines the determination and energy of the individual, guides their behavior and ensures their continuation (Dunn & Stephens, 1972). They experience difficulties as the negative attitudes and beliefs about learning change in later years. Therefore, the change of motivation for learning at an early age plays an important role (Patrick et al., 2008).

In literature, studies related to exam type preferences (Kılıç, 2016; İlhan Beyaztaş & Senemoğlu, 2015; Bal, 2013; Demir, 2012; Eser, 2011; Büyüköztürk & Gülbahar, 2010; Bayrak, 2007) were found. In the research conducted by Grandt (1987), the gender effect on success in the multiple choice exam type was analyzed. Accordingly, it has been revealed that male students perform better in the multiple choice exam type than female students. Zeinder (1987), in his study on high school students, stated that the students found multiple choice questions more simple, understandable, interesting and fair. In addition, they found that they preferred the multiple choice question type to open-ended questions.

Purpose and Importance of the Study

The purpose of this study is to find out the opinions of high school students on different exam types and reveal the motives of their opinions. In the meantime, the study aimed to determine the motives of high school students to prefer multiple-choice questions depending on their grades and their effect on the level of learning motivation. In order to hold the optimal exam for students, it is important to take their opinions on different exam types and reveal the motives for them. In addition, it

is also important to analyse the motives of students to prefer multiple-choice questions depending on their grades and their effect on the level of learning motivation.

METHOD

Research Model

In this study, the mixed method in which qualitative and quantitative research methods were employed together was used and the study was developed by using convergent parallel design. According to this design, qualitative and quantitative findings were obtained simultaneously in the study (Creswell, 2012). Phenomenological design which is one of the qualitative research methods was used. A study with phenomenological design focuses on the common meaning of the experiences of several people with respect to a phenomenon or concept (Creswell, 2012). Scanning design is used as quantitative research method. According to this design, it is aimed to reveal a situation as is in a study (Karasar, 2000).

Study Group

Study group of the qualitative part of the study consists of 12 high school students in total in the state schools in Istanbul province, Uskudar district; of the students, four are 10^{th} graders, four are 11^{th} graders and four are 12^{th} graders. Of them, five are male and seven are female.

Study group of the quantitative part of the study consists of 355 students in total who are 10^{th}, 11^{th} and 12^{th} graders in the state schools in Istanbul province, Uskudar district. While 80.8% of the students in the study group (287 students) are female, 19.2% of them (68 students) are male. Besides, 23.1% of the students (82 students) are 10^{th} graders, 26.5% (94 students) are 11^{th} graders and 50.4% (179 students) are 12^{th} graders.

Data Collection Tools

In the part where qualitative method is used, semi-structured interview technique was used as a data collection tool. A pilot study was made beforehand and then final interview questions were obtained. Final semi-structured interview questions consist of five open-ended questions, and flexibility is ensured in these questions according to understanding levels of the students. Face-to-face interviews were made with the students, and notes were taken in the interviews with each student. The interviews lasted for about three hours in total.

Semi-structured interview questions are as follows;

1. Do you think that different exam types have any impact on your success? What are the reasons for thinking or not thinking this way?

2. Do different exam types increase your anxiety or excitement? Could you please explain this along with the reasons?

3. Among multiple-choice, gap-filling, true-false, matching, classification, short-answer questions, open-ended question types, which type would you prefer and what is the reason for this preference?

4. Do you think that multiple-choice questions reflect your actual success? Could you please explain the reasons for your positive or negative thought?

5. If you become a teacher in the future, which exam type would you use in the questions you prepare for your students? Why?

Two scales were used in quantitative method part of the study. One of the scales is "Academic Motivation Scale" that was developed by Bozanoğlu (2004). This scale consists of 20 items and 1 reverse item. 5-point likert scale is used for rating, and the rates are as follows: 1= Strongly not applicable, 2= Not applicable, 3=Neutral, 4=Applicable, 5=Strongly applicable. In the studies made on scale reliability, test-retest method in which 101 high school students participate was applied, and the correlation between the two application was found .87 (Bozanoğlu, 2004).

Another scale that was used in this study is the mesurement tool "Inventory of Motives to Prefer Written, Short-Answer, True-False and Multiple-Choice Questions (IMP)" that was developed by Eser (2011) . This tool was used to measure exam type preference levels of high school students and determine the motives. 3-point Likert scale was used for IMP separately for each exam type, and the rates are as follows: (1) Not true for me, (2) Partly true for me, (3) Entirely true for me (Eser, 2011).

Data Analysis

In the qualitative method part of the study, the data that were obtained from the semi-structured interview results were used in data analysis. Descriptive analysis was used for data analysis. In this analysis method, a descriptive analysis was made based on the words and the language used in qualitative analysis (Kümbetoğlu, 2005). The participants in the study group, for instance, were coded as follows: 1st female student was coded as (F, student 1), 2nd male student was coded as (M, student 2). The data obtained from all of the participants were coded in this way and presented in the findings section.

In the quantitative method part, the data were analysed by using SPSS program. Multivariate Analysis of Variance (MANOVA) was used to test if the dependent variables of the motives to prefer multiple-choice questions and the learning motivation vary by the independent variables of grades.

FINDINGS

In this study, the data were obtained and the findings were analysed by using qualitative and quantitative method simulatenously. In this section, the findings that were obtained by using qualitative method were presented first.

In order to obtain findings through semi-structured interview questions; the students were asked "Among multiple-choice, gap-filling, true-false, matching, classification, short-answer questions, open-ended question types, which type would you prefer and what is the reason for this preference?". The findings of this question are as follows:

> *"I prefer multiple-choice questions. In such exams, the questions are asked from where I have studied. Besies, the chance to find the correct answer is higher and cheating is easier."* (F, Student 3)

> *"I prefer multiple-choice exam type most. Because the answer to the questions is written in one of the options and one can easily understand the question. I arrive at answer by ruling out options."* (F, Student 2)

> *"My preference is multiple-choice type. I can find the correct answer even if I don't know anything about the question. I turn the wheel and find the correct answer." (F, Student 5)*

> *"I prefer open-ended type. The reason is that one can get full score by writing what the teacher covered in the class. But we directly lose the full score if we give wrong answer in multiple-choice questions."* (F, Student 8)

"I mostly do well in open-ended exams. I provide all I know for the question, and options don't misguide me." (F, Student 11)

Upon viewing the answers of the students, it was found that the majority of the students prefer multiple-choice exam type. The reason for this is generally attributed to the chance of finding the correct answer even if they do not know anything about the question. Other students prefer open-ended exam type. These students stated that they directly lose the full score when they fail to give the correct answer in multiple-choice questions, but they can get a certain score to the extent of their knowledge in open-ended questions. Some students stated that open-ended exams are easier and they are not torn between two options as in the case of mutiple-selection exams.

Another question that was asked to the students is; "Do you think that different exam types have any impact on your success? What are the reasons for thinking or not thinking this way?". The findings of this question are as follows:

"I think question type is highly effective in testing achievement, I cannot fully express my thoughts and knowledge in open-ended questions. But I often do well in multiple-choice exams because of being more comfortable." (M, Student 6)

"I don't think so, people may make mistakes, and may not exactly reflect their knowledge during exam." (F, Student 2)

"I get stressed a lot in multiple-choice exams and all other types of exams. So, it doesn't have any impact." (M, Student 4)

"There is always a chance to find the correct answer in multiple-choice questions. One cannot give any answer to open-ended questions if nothing comes to mind, so it is more stressful." (M, Student 9)

"Yes, my success varies by the type of exam. I get better scores especially in open-ended exams." (F, Student 7)

When the data above were analysed, students were found to state that they are mostly more successful in multiple-choice exams. Some students stated that success doesn't vary by exam types. The students emphasized that all exam types have the same content and cause stress equally.

Another question that was asked to the students is: "Do different exam types increase your anxiety or excitement? Could you please explain this together with the reasons?" The findings of this question are as follows:

"I am mostly more comfortable in multiple-choice exams." (F, Student 5)

"I get panicked in exams regardless of the question type." (F, Student 2)

"I'm afraid of open-ended exam questions. Sometimes, a question looks complicated." (F, Student 8)

"If I know anything about the subject, I can comfortably answer any types of questions." (M, Student 1)

"I get stressed in all exam types, I generally experience exam stress." (M, Student 9)

When the data above were analysed, majority of the students were found to express that different exam types increase their exam anxiety and excitement. Some students stated that their anxiety and stress don't vary by exam types. Other students emphasized the importance of knowledge and stated that there is no need to get stressed as long as the question is known. In addition, the

students expressed that their anxiety and excitement are not related to question types, but "the exam", and they are excited in all exam types.

Another question that was asked to the students is; "Do you think that multiple-choice questions reflect your actual success? Could you please explain the reasons for your positive or negative thought?" The findings of this question are as follows:

> *"I think open-ended questions can reveal the actual success. Because, I can express my knowledge and thoughts clearly in this question type."* (F, Student 12)
>
> *"I think the speed of reading, comprehension and solving increase as I solve multiple-choice questions. I think it affects success positively."* (F, Student 10)
>
> *"Neither actual success nor inadequacies can be revealed only through multiple-choice question type. There must be different questin types in exams."* (M, Student 10)
>
> *"Open-ended questions may bring actual success. I think there isn't such opportunity in multiple-choice type."* (F, Student 7)
>
> *"Absolutely no. I don't believe single exam type can reveal actual success."* (F, Student 2)

When the data above were analysed, it was found that while majority of the students stated that multiple-choice questions reflect actual success, others stated that multiple-choice questions don't reflect actual success.

A different question that was asked to the students is as follows: "If you become a teacher in the future, which exam type would you use in the questions you prepare for your students? Why?" Below are the findings of this question:

> *"I would use different question types. I would ensure my students to get used to all question types. Besides, preparing different question types would be efficient in better evaluating the knowledge of my students."* (F, Student 5)
>
> *"Sometimes, I would mostly prepare open-ended exams in case there might be some students that give wrong answer to multiple-choice questions although they know the subject."* (M, Student 9)
>
> *"I would prefer open-ended questions to evaluate the extent of knowledge of students."* (M, Student 4)
>
> *"If I were a teacher, I wouldn't hold an exam by using a single question type. I would use multiple-choice questions for easy subjects and open-ended questions for the hard ones."* (F, Student 3)
>
> *"I am mostly comfortable in giving answers to open-ended questions. So I would prepare open-ended questions if I were a teacher."* (M, Student 4)

When the data above were analysed, majority of the students were found to state that they would prepare questions in open-ended type for students if they become teachers in the future. While some students stated that they would hold a type of exam for each grade depending on their level, others stated that they can prepare exams in both multiple-choice and open-ended question types.

The findings that were obtained by using quantitative method in this study are presented below.

In this section, we tested if the motives of the students to prefer multiple-choice exams and their learning motivation vary by grades, and presented the analysis results.

The study hypothesis is "Do problem-solving skills and learning motivation of students regarding multiple-choice exams vary by grades?". As seen in Table 1, 10th graders have the highest level of problem-solving skills regarding multiple-choice exams (X=3.60) in the evaluation that was made by grades. On the other hand, 11th graders were found to have the highest level of learning motivation ($\overline{X}$=2.24).

Table 1 Descriptive Statistics of Scores of the Students by Grades

Dependent Variable	Grade	N	$\bar{x}$	SD
The motives to prefer multiple-choice exams	10	82	3.60	.434
	11	94	3.56	.409
	12	179	3.47	.418
Total		355	3.52	.422
Learning motivation	10	82	2.31	.253
	11	94	2.24	.292
	12	179	2.23	.288
Total		355	2.25	.283

N: Number of individuals $\overline{X}$: Average SS: Standard deviation

In table 2, multivariate analysis of variance was used to test if the differences between the averages as a result of the evaluation that was made by grades are statistically significant, and it was found that averages are significantly different from each other (Wilks Lambda (L)=.971, F=.2.559; p>.05).

Table 2 Multivariate Analysis of Variance (MANOVA) Results of the Scale Scores

	Value	F	Hypothesis sd	Error sd	p
Wilks' Lambda	.971	2.559b	4.000	702.000	.038

It was found as a result of the two-way trail analysis in Table 3 that averages of learning motivation vary significantly by grades (F=3.045 p<.001), and there isn't any significant difference among the averages of the motives to prefer multiple-choice exams by grades (F=2.462; p>.05).

Table 3 Two-Way Trail Test Results of the Scale Scores

Dependent Variable	Sum of Squares	df	Mean Square	F	p
The motives to prefer multiple-choice exams	1.073	2	.537	3.045	.049
Learning motivation	.390	2	.195	2.462	.087

When the average differences of the motives to prefer multiple-choice exams and learning motivations were analysed by grades according to the multiple comparison test for finding the source of difference among the averages in Table 4, significant differences were found among 10th and 12th graders.

Table 4 LSD Paired Comparison Results Regarding Scale Scores

Dependent Variable	Grade (I)	Grade (J)	Difference of Average (I-J)	SE	p
The motives to prefer multiple-choice exams	10	12	.1266*	.05597	.024
Learning motivation	10	12	.0823*	.03752	.029

* p<.001 SE: Standard error

Findings on the correlations among the motives to prefer multiple-choice exams and learning motivations

Results are presented fr the Pearson Correlation analysis that was made in order to find if there is a significant correlation among the scores of the motives of the students to prefer multiple-choice exams and their learning motivation.

Positive significant correlations were found among the motives of the students to prefer multiple-choice exams and their learning motivation [r=.95].

CONCLUSION AND DISCUSSION

Results of the findings that were obtained from the study fit for purpose. Opinions of the students on exam types were obtained and the motives for these opinions were revealed. Considering in general terms, the students put emphasis on multiple-choice and open-ended exam types. Of these two exam types, they preferred multiple-choice exam type most. However, there isn't any significant difference between the number of students that prefer multiple-choice and open-ended exam types. In addition, high correlation was found between the motives to prefer multiple-choice exams and the learning motivation as a result of the study. The motives to prefer multiple-choice exams and the learning motivation differ among the final year students and 10^{th} graders. The results of this study have parallels with those of the studies in this field. According to Kılıç and Çetin (2010), multiple-choice exam type is used in many important exams in our country. Therefore, multiple-choice is one of the exam types which the students that prepare for such exams are mostly familiar with. This case is considered as one of the motives of students to prefer multiple-choice exams more. Learning approaches can be defined as certain strategies that students adopt for studying and their different learning characteristics. It was found that various tools were developed in order to measure them (Cassidy, 2004). Coffield et al. (2004), emphasized various academic foundations of the study on learning styles. These foundations are grouped in three approaches; theoretical, pedagogical and commercial.

Majority of the students admit the hypothesis that different exam types have impact on their success. Also, most of them stated that their exam anxiety and excitement don't vary by exam types. These students indicated that their anxiety is not related to the exam type, but "the concept of exam" itself. Students feel a certain level of anxiety regardless of exam type, because they feel that they are evaluated in exams. Most of the students stated that multiple-choice exams are not sufficient alone in evaluating their knowledge and skills. But still, they prefer multiple-choice exams for various reasons.

However, there are contradictions between the answers they give as a student and the answers they give when they consider themselves as teachers. Most of the students that prefer multiple-choice exam type stated that they would hold open-ended exams for their students when they become a teacher. Besides, there are also students who think that "they would apply both exam types and hold a different exam type for each grade". When we take into account all of these, the students know which exam types benefit them, but they refrain from such exams. Teachers may explain the qualitative and quantitative characteristics of different exam types to students in order to prevent such negative attitudes towards some exam types. In open-ended questions, teachers may avoid using the question types that may unsettle students, and make evaluations by taking account of their interpretation skills besides measuring their accumulation of knowledge. They can make inferences upon exam evaluations and apply the activities and excersises that are suitable for these inferences in the class. In this way, their problems related to different exam types may decrease by means of improving the skills of students as well as making up the deficiencies in their accumulation of knowledge. In addition, the benefit of multiple-choice question type can be emphasized for the students that refrain from such questions. According to Traub (1990), students take a more positive attitude towards multiple-choice exams compared to free response tests. Because, they think that preparation for these tests is easier, they are easier to solve, decrease stress and anxiety and thus bring them relatively higher scores.

Furnham, Batey and Martin (2011) carried out studies in which multiple-choice exams and continuous assessment methods are preferred by stuents. They found that these methods promote participation and increase the motivation and learning of students. In addition, Bandarage et al. (2009) stated that multiple-choice exams are the sources of motivation for students in terms of continuous assessment.

While multiple-choice and open-ended exams can be applied, different exam types can also be applied. Every exam type has a purpose of application. Hence, teachers should not limit themselves to a single exam type for their students, they should also apply different types. Thus, target achievements can be measured through different question types.

REFERENCES

Bal, P.A. (2013). Lisans öğrencilerinin matematik dersine ilişkin değerlendirme tercihleri ile öğrenme stratejileri arasındaki ilişkinin incelenmesi. *International Online Journal of Educational Sciences, 5*(1), 242-257.

Bandarage, G., Zoysa, M.K., & Wijesinghe, L.P.(2009). Effect of continuous assessment tests with multiple choice questions on motivation for meaningful learning. *In Proceedings of the Annual Academic Sessions of the Open University of Sri Lanka*. 8–10.

Bayrak, R. (2007). *Ölçme ve değerlendirmenin öğrenme üzerindeki etkisi*. Yüksek lisans tezi, Karadeniz Teknik Üniversitesi, Trabzon.

Bozanoğlu, İ. (2004). Academic motivation scale: Development, validity, reliability. *Ankara University Faculty of Educational Sciences Journal, 37*(2), 83-98.

Büyüköztürk, Ş. ve Gülbahar, Y. (2010). Yükseköğretim öğrencilerinin değerlendirme tercihleri, *Eurasian Journal of Educational Research, 41*, 55-72.

Cassidy, S. (2004). Learning styles: an overview of theories, models and measures. *Educ Psych 24,* 419–444.

Clarke, D.P., Heaney, J. G., & Gatfield. T.J.(2005). Multiple choice testing: A preferred assessment procedure that is fair to all our business students? *In ANZMAC 2005 Conference: Marketing Education*. 51–57.

Coffield, F., Moseley, D., & Hall, E., et al. (2004). Learning Styles and Pedagogy in Post-16 Learning: A Systematic and Critical Review. *London: Learning and Skills Research Centre*

Creswell, J. W. (2012). *Educational Research: Planning Conducting and Evaluating Quantative and Qualitative Research.* Boston: Pearson Publicatio

Demir, M. K. (2012). Sınıf öğretmeni adaylarının arasınav ve dönem sonu sınavları hakkındaki görüşleri. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 193-214

Dunn, D.J., & Stephens, C.E. (1972). *Management of Personel Manpower- Management and Organizational Behaviour.* New York: McGraw Hill Book Co.

Eser, M. T. (2011). *Examination of causes of some of the factors which affect students' exam type preference.* (Unpublished masters dissertation). Hacettepe University, Ankara.

Furnham, A., Batey, M. & Martin, N. (2011). How would you like to be evaluated? The correlates of students' preferences for assessment methods. *Personality and Individual Differences 50*, 2 (2011), 259–263.

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 Teesting Service.

Gronlund, N.E. (1998). *Assessment of student achievement*. Boston: Allyn and Bacon.

İlhan Beyaztaş, D & Senemoğlu, N. (2015). Başarılı öğrencilerin öğrenme yaklaşımları ve öğrenme yaklaşımlarını etkileyen faktörler. *Eğitim ve Bilim, 40*(179), 193-216

Jenkins, T. (2001). *The Motivation of Students of Programming.* Master's thesis. University of Kent at Canterbury.

Karasar, N. (2000). *Scientific research method (12th edition*). Ankara: Nobel Publication Distribution

Kılıç, Z. (2016). *Öğrencilerin sınav türü tercihlerinin çeşitli değişkenlerle ilişkisi.* Yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Kılıç, Z. & Çetin, S. (2018). Examination of Students' Choice of Exam Type in Terms of Various Variables. *İlköğretim Online Dergisi*, 17(2), 1051-1065.

Kümbetoğlu, B. (2005). *Qualitative methods and research in sociology and anthropology.* İstanbul: Bağlam Press.

Patrick, H., Mantzicopoulos, P., Samarapungavan, A., & French, B.F. (2008). Patterns of Young Children's Motivation for Science and Teacher–Child Relationships. *The Journal of Experimental Education*, 76(2), 121–144.

Traub, E.R. (1990). Multiple-Choice vs. free-response in the testing of scholastic achievement. *Ontario Institute for Studies in Education*

Zeinder, M. (1987). Essay Versus Multiple-Choice Type Classroom Exams: The student's Perpective. *Journal of Education Research. 80*, 352- 358.