Research Article

THE COMPARISON OF THE DIFFERENT STATISTICAL ANALYSIS RESULTS USED ON THE EVALUATION OF SCALES ON A REAL SAMPLING

Fatih SEZEK
Ataturk University, Kazım Karabekir Faculty of Education, Erzurum, Turkey
e-mail: fsezek@gmail.com, ORCID: 0000-0002-1841-4303

Uluhan KURT
Ministry of Education, Erzurum, Turkey
e-mail: uluhaaan@hotmail.com, ORCID: 0000-0002-0683-6875

Submission Date: 26.04.2019 Publication Date: 29.07.2019

Doi: 10.33418/ataunikkefd.558316

Abstract

The purpose of this study is to examine the effects of using different statistical analysis methods on the results of questionnaire or scales evaluation in education studies. The scale-survey method was used for data collection. In accordance with this purpose, "Students' Satisfaction of Physical Space" (SSPS) developed by researchers and measuring their school satisfaction level of students was used. The sample of the study is consisted of 720 students studying in the boarding school in the province of Erzurum. Firstly, the data obtained from the sample were scored and then, both parametric and non-parametric analyzes were performed on the total score of the test. Evaluating of tests results were statistically compared. Thus, we tried to determine which analysis method was more accurate in the evaluation of the scales. Accordingly, we were be able to achieve more reliable results. As a result, it was determined that performing statistical analyzes by scoring questionnaires and scales brought different problems. Contrary to the general belief that parametric tests are more reliable than non-parametric tests, it is also shown that the comparison with the chi-square test is one of the most suitable methods in the analysis of categorical datum obtained from the questionnaires. On the other hand, instead of collecting the scores of the different class levels students' answers to each question, it is more convenient to apply the chi-square test to the total frequencies of the answers. In addition, among students of different grades (K5-8), it was determined that there was a statistically significant difference among their satisfaction levels from the physical environment of the school.

Keywords: Parametric and non-parametric tests, Chi-square, scales, comparison of statistical tests

Öz

Bu çalışmanın amacı, eğitim çalışmalarında farklı istatistiksel analiz yöntemlerinin kullanılmasının anket sonuçları veya ölçek değerlendirme sonuçları üzerindeki etkilerini incelemektir. Veri toplamada ölçek ile tarama yöntemi kullanılmıştır. Bu amaç doğrultusunda araştırmacılar tarafından geliştirilen ve öğrencilerin okul memnuniyet düzeyini ölçen "Yatılı Bölge Ortaokullarındaki Öğrencilerin Fiziki Mekân Memnuniyetleri Ölçeği" (ÖFMMÖ) kullanılmıştır. Araştırmanın örneklemini, Erzurum ilinde YBO'larda okuyan 720 öğrenci oluşturmaktadır. İşlem sürecinde ilk olarak, örneklemden elde edilen verilere değerler verildi, sonrasında öğrencilerin testin tamamından aldıkları toplam puanlar üzerinden hem parametrik hem de parametrik olmayan analizler yapıldı. Bu sayede ölçeklerin değerlendirilmesinde hangi analiz yönteminin doğru olduğu belirlenmeye çalışılmıştır. Sonuç olarak, anket ve ölçeklerde öğrencilerin verdikleri cevaplar üzerinden puanlama yaparak istatistiksel analiz yapılması beraberinde bir takım

problemleri getirmektedir. Parametrik testlerin parametrik olmayan testlerden daha güvenilir olduğuna dair genel düşüncenin aksine ki-kare testi ile karşılaştırmanın, anketlerden elde edilen kategorik verilerin analizinde en uygun yöntemlerden biri olduğu da gösterilmiştir. Öte yandan, farklı sınıf seviyelerinde puanların toplanması yerine ki-kare testi için öğrencilerin her bir soruya verdikleri cevapların toplam sıklıklarının girilmesi daha uygundur.

Anahtar Kelimeler: Parametrik ve parametrik olmayan testler, Ki kare, ölçekler, istatistiksel testlerin karşılaştırılması

INTRODUCTION

Today, questionnares and scales are among the most popular data collection tools used for academic and social research. They have a wide range of uses in many disciplines from subjects such as interest, motivation, self-design, attitude and values to the real-world applications of academic theories. Thanks to these data collection tools, it is possible to obtain information about many subjects by asking people questions (Erkuş, 2010; Koç, 1986; Koçyiğit, 2002; Köklü, 1995; Özgüven, 2007).

Considering the widespread uses of questionnaires and scales, it is not correct to think of them as simple knowledge-collection tools. The accuracy of the data obtained depends to a large extent on the method, the structure of the questionnaire and the scale, and the correct selection of the statistical analyzes used. While determining the method of analysis of questionnaires and scales, the features to be considered are: The size of the sample, the relationship between the variables, the comparison of the groups, whether the scale or questionnaire is applied more than once in different years, how the datas are recorded (percentage, average, etc.), number of dependent and independent variables, whether the data is sufficient quality (Anderson; 1988; Beatty, 1997; Büyüköztürk, 2011; Tezbaşaran, 2004).

On the other hand, datum on psychological and social research cannot be measured as measured by physical properties such as length or volume. Because the features to be measured have no an objective and meaningful definition accepted by all people, and also there is no absolute zero point. According to Kerlinger (1973): Intelligence, talent and personality test scores are definitely the ranking scale. In this case, the scores given in the test can only be sorted. However, the actual number of units separating a score from the next score can be many, several, or approximately zero. Therefore, the analyzes that can be performed with the ranking scales are limited. For example, ‘‘very satisfied’’ and ‘’satisfied’’ can not be averaged, In addition, people with higher test scores cannot be said to be more satisfied. Because sorting scales do not give any idea about how large a value in the rankings is from another. Although the rankings in these scales are indicated by numbers or scores, these numbers are not units of measurement (Büyüköztürk, 2007).

The qualities measured in education, psychology and sociology are not constant but variable. The participants' attitudes towards a subject, event or situation, the perception of activities related to the services provided in a subject, etc. features may change over a short period of time. The features that are wanted to be measured may differ not only from person to person, but also from situation to situation, from time to time for the same person. Research shows that different answers are given to the same question even in different survey methods (mail, telephone, interview). On the other hand, the abstractness of the behaviors subject to measurement makes it difficult to determine the critical behaviors that define the feature to be measured. Moreover, it makes it difficult to use the units in the same sense or to ensure the equality of the units. Although abstract concepts were written very detailed by researcher, the phenomenon described can be

visualized in many ways in the minds of the participants (Salant ve Dillman, 1974). On the other hand, when different people observe the same phenomenon, they do not always understand the same. Different value judgments and personal expectations of individuals may cause observer prejudices to occur. In this case, people sometimes may see and hear the way they want (Baştürk, 2014).

In order to minimize the negative situations mentioned above, the researchers refer to standardization and operational definitions. Standardization is to carry out uniform and consistent operations at all stages of data collection. Thus, each participants can be asked the same questions and the answers can be scored according to predetermined rules. By sufficiently standardizing all the features of the tests, all participants can experience the same experimental conditions and the results can be printed or recorded more healthily. Thus, results can be compared with the studies of other researchers. To ensure standardization, Likert (1932) developed a new and simple measurement tool that included a list of attitudes preferred or not for use in psychology.

The situations listed in the Likert-type scales are applied to a group of participants. From the datum obtained, the substances with the highest distinguishing power are determined and brought together, and then final test was combined. The items in the scale are usually accompanied by a five-option (such as; from strongly agree to strongly disagree). Scoring is also completed by evaluating the values from 1 to 5 numerically. However, for the reasons mentioned above, the datum obtained are never equal-spaced or proportional scale (Gerrig ve Zimbardo, 2016). The question then comes to mind: While the scales in educational sciences and psychology should actually be treated according to the nominal or ordinal scale, why do the researchers process as interval scale? Probably, this situation may be due to the flexibility and diversity in statistical procedures (arithmetic mean, standard deviation, t-test, Anova, Manova, Ancova, etc.) that can be performed for interval scales.

Because descriptive tests, such as arithmetic mean, can only be used in cases where the datum are considered to be really interval or ratio scale. As mentioned before, research in the education can never be measured with intervals or ratio scale. Calculations are made assuming that only the data is such. However, a large number of different statistical analyzes (for examples median, mod, frequency tables, percentage, Spearman rank difference correlation coefficient, chi-square, graphs or tables) in nominal or ordinal scales can be made also (Büyüköztürk, 2017).

There are a few main objectives of this study: Our aim is to analyze the datum obtained from surveying instruments such as questionnaires or scales used in sociology and psychology with the most accurate methods, and is to discuss how to achieve the best results. For this purpose, the data were scored first, then data were analyzed by both parametric and non-parametric methods. In addition, the data were accepted as frequency and analyzed by Chi-Square method.

Research Problem

What is the most appropriate statistical method in determining the relationships between the physical space satisfaction levels (PSS) of the students studying at different grade levels in YBO?

Sub-Problems

When the scale items are scored;

1. Which of the parametric tests (Anova, t-test) or non-parametric tests (Kruskal Wallis, Mann Withney-U) are more accurate and appropriate in the analysis of scale items?
2. Assuming the data is normally distributed, in the application of parametric tests, which of the total score of test gives more reliable results?
3. Assuming the data is normally distributed, in the application of non-parametric tests, which of the total score of test gives more reliable results?

   Assuming datum frequency;

4. Is the application of the chi-square analysis to the total frequencies of the responses given to the scale statistically more accurate and reliable?

METHOD

Research Model and Sample

The sample of this study, in which the survey method is used, consists of 720 students studying at the secondary schools at YBO's in the districts of Erzurum in the spring term of 2017-2018 academic year. The scale was applied to a sufficient number of students in terms of the power to represent the universe (Yazıcıoğlu ve Erdoğan, 2004).

Application Process

The necessary permissions for the research were obtained from the MEB. The scales containing the information required for the study were applied to 720 volunteer students studying in 13 YBOs in the districts of Erzurum during the first period of 2017-2018 academic year. Students who were surveyed, 142 were in the 5th grade, 134 were in the 6th grade, 221 were in the 7th grade, and 221 were in the 8th grade. The same questionnaire was applied to all students. The application was made by the researchers personally.

Data Collection Tools

Demographic Information Questionnaire

The demographic information questionnaire prepared by the researchers was applied to determine the different features of the students in the boarding schools. Participants were asked questions about gender, grade level, outline grade point average, residence information and number of siblings.

Students' Satisfaction of Physical Space (SSPS)

The scale, developed by Kurt and Sezek (2018), aims to measure the level of satisfaction of the students in the regional boarding schools in terms of the opportunities offered by their schools. The scale consists of 9 items in total. It measures two sub-dimensions as attitude and competency dimension. The internal consistency coefficient (Cronbah alpha) is for the attitude size of .76, for proficiency size .71 and .82 for the whole scale. The total variance value explained by the factors in the scale was calculated as 55% and compliance indices obtained as a result of DFA were; $\chi 2/df = 1.74$, RMSEA= 0.05, RMR =0.06, GFI =0.98, AGFI = 0.96, CFI=0.97 and NFI=0.96. The scale is a Likert-type with five options. Student opinions are scored as 1- strongly disagree, 2-

disagree, 3- undecided, 4- agree and 5- strongly agree. The part correlation of each item with the whole test is at least 0.4.

FINDINGS

In this study, the Physical Space Satisfaction Scale was applied to the students of different class levels in YİBO's and their satisfaction levels were tried to be determined. In the statistical analysis, parametric and non-parametric tests were performed to determine the relationships between the variables in the scales, and to score the data given as interval scales. Then, the answers given to the test were analyzed by frequency-based chi-square analysis as in the nominal and ordernal scales. SPSS 24.00 statistical program was used to make both parametric and non-parametric analyzes of the data obtained from the students (URL 1).

Before the statistical analysis, whether the data were categorical (nominal or ordinal) or continuous (interval or ratio) were examined. Non-parametric tests were used for Categorical data, and parametric statistics were also used for continuous data. The assumptions of parametric tests can be listed as follows: First, the data should be interval or ratio. Second, the data should follow the normal distribution. Third, group variances should be equal (Daniel, 1990). According to this; We considered each item of our scale equally spaced. Students' answers were scored as follows; strongly disagree (1), disagree (2), undecided (3), agree (4), strongly agree (5).

The total test scores of the students were calculated by collecting the scores of the answers given to each question. The normal distribution of data was examined. Thus, the following calculations were obtained. According to this:

When the Data Is Scored, What Is the Normal Distribution Curve of the Test?

We have examined each different grade level in terms of the total test score, respectively. According to the total test scores, we found that the significance levels of all the analysis results were the same. Therefore, we have given them all in a table (Table 1). According to Büyüköztürk (2011), Kolmogrov-Simirnov value was examined because the sample number was 50 and above. When we examine different grade levels, data colected from other grade levels are not distributed normally, except for 6th grade ($p < 0.05$). On the other hand, even if we consider all grade levels as a single sample, we can say that the distribution is not normal (Table 1).

Table 1.
Results of the Kolmogorov-Smirnov Normality Test for the Total Scores of the Scale

Grades	N	Sig.	Skewness	Kurtosis
5th grade	142	.00	-0.54	-0.16
6th grade	134	.20	-0.05	-0.35
7th grade	221	.00	-0.30	-0.50
8th grade	221	.00	-0.35	-0.37
Total	720	.00	-0.35	-0.30

However, by looking at a single result, the normality of distribution cannot be determined. The whole of Kurtosis, skewness, Kolmogrov-Simirnov, Shapiro-Wilks, histogram, P-P, Q-Q tests should be examined. Huck (2008) states that Skewness value should be between +1 and -1 and Kurtosis should be between +2 and -1. According to the values of skewness and kurtosis, all of the research data showed normal distribution.

On the other hand, the arithmetic average of the responses of the students at all different grade levels are in 95% confidence interval for mean lower/ upper bound. As a result, according to Kolmogorov-Smirnov, data do not distribute normally, while other data seem to be dissipating normally. The statistics about the descriptive test scores of the students are presented in Table 2.

Table 2.
Descriptive Statistics for Total and Average Test Scores

Total scores of responses given to the test			
Grades	**N**	**Mean**	**Standart deviation**
5th grade	142	32.77	7.74
6th grade	134	30.60	6.18
7th grade	221	27.97	8.14
8th grade	221	28.66	7.73
Total	720	29.62	7.80

Because of the dilemma of the analysis results mentioned above, it was concluded that it is necessary to make an analysis by assuming that the data is normal and then distorted.

When We Accept that the Sample Is Distributed Normally, ANOVA Test for the Total Score of the Scale

ANOVA was applied because of 4 different levels of grade. The findings of the applied statistical procedures are presented in Table 3 and 4.

Table 3.
ANOVA Test Results of Total Test Scores

Source of Variance	**Sum of Squares**	**DF**	**Mean Square**	**F**	**p**
5-8th grade	2320	3	773.35	13.40	.00**

p<.01**

According to Table 3, there were differences in ANOVA test results. According to Levene test results, it was determined that the variances of the group distributions were not homogeneous (LF= 5.37; P =.00). Since the variances were not equal, Tamhane T2 test which is one of the paired comparison tests was used (Güriş and Astar, 2015). Thus, the following results were obtained (see Table 4).

Table 4.
Tamhane T2 Test Results for Comparisons Between Classes

Grades	**Mean Diffrence (I-J)**	**Standart Error**	**p**
5 to 6	2.13	.84	.07
5 to 7	4.75**	.85	.00
5 to 8	4.06**	.83	.00
6 to 7	2.63**	.76	.00
6 to 8	1.94	.75	.06
7 to 8	-.69	.76	.93

p<.01**

According to Tamhane T2 test results, when we compare classes at different levels as a binary, there is statistically no difference between 5th and 6th grades, 6th and 8th grades, 7th to 8th grades. Whereas, there is a significant difference between 5th and 7th grades, 5th and 8th grades, and 6th to 7th grades (Table 4).

When We Accept that the Sample Is Not Distributed Normally, Kruskall Wallis for Total Scale Score

In cases, where the assumptions of parametric hypothesis tests are not met, non-parametric tests are more appropriate. In this section, the Mann Whitney U-Test was used instead of the independent sample T-Test, and the Kruskal Wallis H-Test was used instead of one-way ANOVA (Daniel, 1990). The results of this analyses are given below (Table 5 and 6).

Table 5.
Kruskal Wallis Test Results for the Total Test Scores

Grades	N	Mean Rank	DF	Chi-Square	p
5th grade	144	443.35			
6th grade	134	379.13	3	35.48	.00**
7th grade	221	320.14			
8th grade	221	335.57			

p<.01**

According to Table 5, there are statisticaly differences among classes. Mann Whitney U test was used in order to compare between scores of differ grades. (Can, 2018).

Table 6.
Mann Whitney U Test Results for Total Test Scores

Grades	N	Mean Rank	Sum of Rank	U	p
5 to 6	144	152.75**	21996.5	7739.5	.00
	134	125.26**	16784.5		
5 to 7	144	219.69**	31636	10628	.00
	221	159.09**	35159		
5 to 8	144	215.91**	31090.5	11173.5	.00
	221	161.56**	35704.5		
6 to 7	134	196.79*	26369.5	12289.5	.01
	221	166.61*	36820.5		
6 to 8	134	192.09*	25739.5	12919.5	.04
	221	169.46*	37450.5		
7 to 8	221	216.45	47834.5	23303.5	.40
	221	226.55	50068.5		

p<.01**, p<.05*

In Table 6, When we compare classes at different levels in pairs, while there is statistically no difference between the 7th and 8th grades, there is a significant difference between the test scores of the other classes. Kruskall Wallis and Mann Whitney U analyzes were performed on the arithmetic means of our scale.

However, the significance levels were similar with the results of total test scores. Therefore, it is not mentioned here to avoid rewriting similar results.

When Data Were Considered as Frequency, in the Situation Performing Chi-Square Test

We thought that the students' answers to the test questions were on the ordinal scale. Chi-Square test was performed on the total frequency of the responses given. According to Daniel (1990); Students' answers to each question can be calculated from their frequencies without scoring. The total frequencies of the responses of the students to the test items were calculated separately for each different grade level. Thus, the following calculations were obtained (Table 7).

Table 7.
Frequency and Percentage of Responses Given by Classes

Grades	**Strongly disagree**	**Disagree**	**Undecided**	**Agree**	**Strongly agree**	**Total**
5th grade	120 %9.4	136 %10.7	209 %16.4	357 %28	453 %35.5	1275
6th grade	117 %9.9	165 %14	231 %19.5	390 %33	280 %23.6	1183
7th grade	371 %19	278 %14.3	334 %17.1	578 %29.7	388 %19.9	1949
8th grade	325 %16.8	245 %12.6	344 %17.7	633 %32.7	391 %20.2	1938
Total	933 %14.7	824 %13	1118 %17.7	1958 %30.8	1512 %23.8	6345

N = 720 students and 167 pieces of boxes are not marked.

In Table 7, Satisfaction rankings can be understood by the total percentage of those who agree (Agree + Strongly Agree) or disagree (Disagree + Strongly Disagree) answers given by students. According to this, approximately 63% of the 5th grades, approximately 57% of the 6th grades, 50% of the 7th grades and about 53% of the 8th grades said that they accepted the opinions in the survey. So they're quite happy with the the opportunities of their schools.

On the other hand, the most effective way of comparing two classes is to extract the frequencies of the given answers, respectively. For example, in table 7, Let's compare the 5th and 6th grades. The answers percentages (%) of the 5th and 6th grades were subtracted from each other in each column, respectively. In the subtraction process, the 5th classes were writen to the upper line, and the 6th class to the bottom line. If the answer percentages (%) for the 5th grades is greater than the 6th grades, then the result is positive. Otherwise, it will take a negative value. Accordingly, if we calculate the percentage differences, strongly disagree = -0.5, disagree = -3.3, undecided = -3.1, agree = -5, strongly agree = +11.7. When we examine the results of the extraction, the percentages of the 6th grades in the form of ‘‘the disagree’’ and ‘‘agree’’ are higher, while the percentages of the 5th grades in the form of ‘‘strongly agree’’ are higher. That is, it is understood from the answers given to the questionnaire that the 5th grade responds more positively. In this way, all classes can be paired respectively, and the satisfaction levels of the classes can be found. However, Chi-Square test was used to understand the difference between the classes.

In Figure 1, data were processed to SPSS for Chi-Square test. In the first column, 5th grades as 5, 6th grades as 6, 7th grades as 7, 8th grades as 8 were written. In the second column, ‘‘strongly disagree’’ as 1, ‘‘disagree’’ as 2, ‘‘undecided’’ as 3, ‘‘agree’’ as 4, ‘‘strongly agree’’ as 5 were recorded. The total frequencies of the responses of the students at each grade level were processed in the third column, respectively (Figure 1).

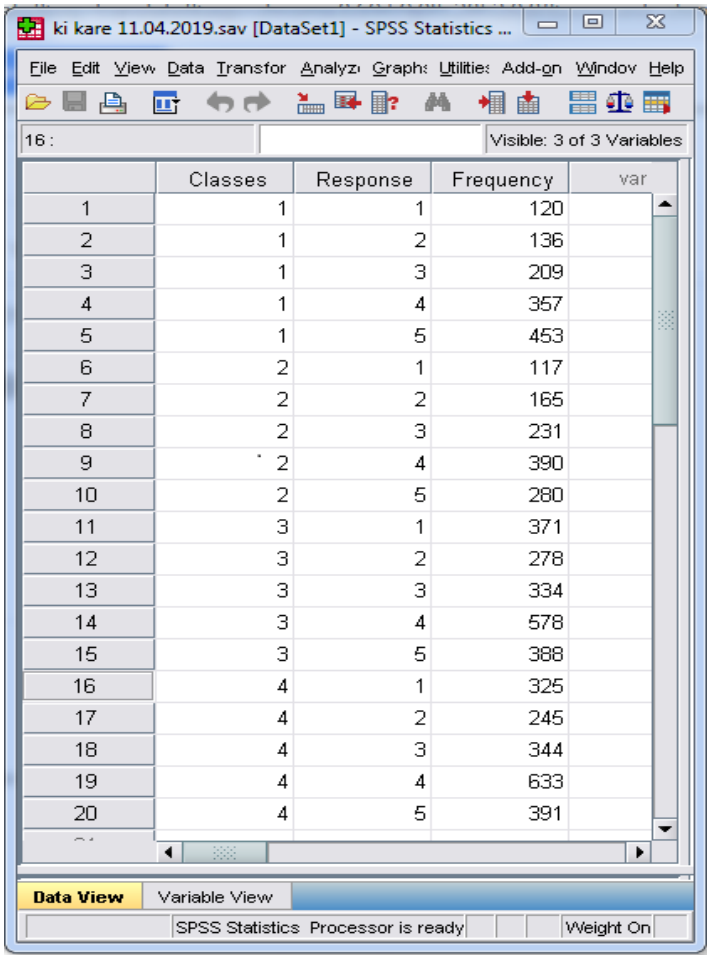

Figure 1. Screenshot of How Data is Processed in SPSS

Since there is no category with less than 5 in expected frequencies, the chi-square table value is examined (Table 8). As a result, according to different grade levels; There are a statistically significant difference between the answers (X2 = 188,126, p <0,05).

Table 8.
Chi-square Test Results of the Responses to the Scale

Groups	Value	DF	p
5^{th}-8^{th} Classes	190.71**	12	.00

The chi-square test was used to compare the different level classes in pairs (see Table 9).

Table 9.
Binary Comparison of Classes by Chi-square Test

Classes	DF	X^2	p
5 to 6		42.84**	.00
5 to 7		127.73**	.00
5 to 8	4	107.86**	.00
6 to 7		49.39**	.00
6 to 8		30.95**	.00
7 to 8		7.75*	.10

In Table 9, according to the answers to the survey, only there is no difference between the answers of 7th to 8th grades. 5th grades students have more positive opinions than the 6th grades, and 6th grades students have more positive opinions than 7th and 8th grades, statistically.

DISCUSSION

In Likert-type questionnaires and scale studies, the answers given for each question are actually considered as frequency in the ordinal scale. However, data have been converted into an interval scale before, and then they have been analyzed. In these studies, the scores given to the answers of the students (for example, "strongly disagree" as 1 or "strongly disagree" as 5) are collected and analyzed on the total test scores. Whereas, when the answers given to the questions in the questionnaire or scale are considered, Are the meaning levels of their expressions (disagree, strongly disagree, undecided, agree, and strongly agree) equal? Isn't it? It is impossible to answer this question. Because these units are arbitrary, and vary according to the respondent and the question. Due to the uncertainty of the satisfaction measurements and the dynamic nature of the desired qualities, the answers of each participant to each question are not accurate or inaccurate. We discuss a statement of the scale developed by Kurt and Sezek (2018) and used in this study. ‘‘I think the materials in the dormitory (bed, blanket, bed linen, etc.) are clean’’ First student may give a statement as agree, and second may give a statement as strongly disagree. In this case, we cannot say that the first student is correct and the second student's opinion is wrong. Because the concept of cleanliness varies from person to person. In fact, it is even wrong to say that the two participants who expressed that they were agree in this statement of the scale had the same feelings. Because each individual's personal cleaning criteria are different. While a person migth think it was clean, someone else might think the same environment was dirty. This situation is the effect of individual differences on the criteria.

Since each question in our scale is not an absolute true or false answer and no real zero point that can make a reference, scores cannot be collected. Therefore, the total score of the test in terms of measurement does not make sense. The numerical values given are not the magnitude of the individual's belief in a situation. It is not the right practice to collect the attitudes of the individual to a range determined by the researcher. In fact, the main purpose of scoring the answers of the participants and then collecting these scores by the researchers is to rank each participant in the sample in terms of the investigated feature. Another important problem is that when we score the responses given, there is 1 point difference between the answers (strongly disagree and disagree), respectively. But, the difference between the first (strongly disagree) and last stages (strongly agree) is up to 4 points. In other words, the researcher gives an arbitrary score to each answer given.

In this case, When the competence of the scores given to measure different attitudes is controversial, and then how accurate is the calculation of standard deviation and variance over the total test scores by collecting these scores? Because the measurements with a large standard deviation are actually more weight than they are given to them. This drawback can be roughly eliminated by dividing each measurement result by its own standard shift or by converting each raw score to the Z or T standard score (Tekin, 2000). But, in this situation, it brings additional processes and new problems.

One of the major problems is that the total test scores of the two participants who give different answers to the scale questions may be equal. For example, our scale consists of nine questions. Assume that the total score of two different participants is 30. There may be hundreds different responses combinations that can give this sum. The first student may respond to the first three questions as "extremel agree", and the other six questions as "undecided" or "disagree". The second student may give the answers to the first four questions as "strongly disagree", and the last five questions as "agree". In this case, both students may have the same test score. In this case, although these two participants have different opinions at the level of questions, they will be included in the same group as their total scores are equal. These differences will not be reflected in the analysis results. As this may cause significant deviations in the statistical results in small samples, it will be necessary to work with as large a sample as possible. However, if the responses are taken directly as total frequencies, since the distribution of the answers will be reflected in the total test frequency, there will be no slip situations in the test results due to the scoring of the researcher. In all analyzes conducted on scoring, the importance status in parametric and non-parametric test results varies. There is no statistically significant difference between the 5th to 6th, 6th to 8th, and 7th to 8th grades in pairwise comparisons in parametric tests. However, in the results of Kruskall Wallis and Chi Square, there is no difference between only 7th and 8th grades. In Table 6, in the Tamhane T2 test, there are no statistically significant difference between the 5th to 6th, and 6th to 8th grades, and the significance values (p) were close to each other. In Table 8, there is a statistically significant difference between the aforementioned group comparisons in the Mann Whitney U test. In addition, the difference between 5th to 6th grades are larger (p=.00), whereas the difference between 6th to 8th grades are smaller (p=0.04). In Chi-Square test, the significance levels between the above-mentioned classes in pairwise comparisons are (p=.00). One of the possible causes of different levels of significance in different statistical analysis results may be related to the issues discussed above. In order to understand the effects of the scoring on the results, we think that it should be investigated more by making comparisons over large and small sample groups.

On the other hand, the applied scale has two dimensions, high reliability and validity and all questions of the scale have positive meanings. Considering all these features, instead of collecting the scores of the answers given to each question, it is more appropriate to calculate the total frequencies of each answer option. In addition, the comparison conditions obtained from the parametric tests can be obtained by comparing between the pairs of classes at different levels, by using the chi-square test.

Another problem encountered when the answers to the scale is scored is whether the data are distributed normally. If data is normally distributed, parametric tests should be performed. Non-parametric tests should be performed if not distributed. In Table 1, it is seen that the data are normal at different grades levels in terms of kurtosis and skewness values. whereas, according to Kolmograv-Simirnov test results, other grades are not distributed normally, except for 6th grade. In this case, it is necessary to apply parametric

test to the sixth grade and non-parametric to others. To overcome this problem, all class levels were considered as a single sample and the normal distribution of the whole population was examined. As a result of this process, some analysis results show that non-parametric and parametric tests can be performed. These situations lead to confusion about the choice of analyzes to be applied to the scale. In this study, normal distribution analyze was performed for case and the result was not changed (Tablo 1-11).

Parametrical and non-parametric tests were performed due to the above-mentioned confusion. When the results of the parametric and non-parametric analysis are compared, it is seen that the importance levels of the differences between satisfaction levels between classes change. However, when working with other research data, the results of parametric and non-parametric analysis may be similar or different. In comparison of all classes, the significance levels of the Kruskall Wallis and Mann Whitney U are different from ANOVA. However, the most important problem here is whether the data are parametric or non-parametric.

Considering all the above considerations, we compared the chi-square test results by accepting students' responses to the test as a frequency. Thus, by scoring the data, we will not have to deal with many problems that we may encounter when the total test score of the test is taken. When we look at the results of Kruskall Wallis and Mann Whitney U with chi-square results, there is no difference in terms of significance levels. That question then comes to mind: If the results are the same, why are the data considered as equally spaced scale, instead of frequencies? In our opinion, this situation is based on the preference of researchers to analyze data from the past. As can be seen here, the properties investigated can be analyzed by applying chi-square method. Moreover, when this method is preferred, there are not many negative situations mentioned above.

As a result, in psychological tests and especially in Likert-type scales, it is considered that it would be more appropriate to evaluate the given answers as total frequencies. Therefore, statistical analysis of the scales; frequency distribution, frequency tables, median, mod and chi-square tests will be more appropriate. In addition to, we also think that more studies should be done on the application areas of chi-square test.

REFERENCES

Anderson, L.W. (1988). *"Attitudes and their measurement"*, New York: Educational Research, Methodology and Measurement. An International Handbook, Keeves, J.P. (Ed), 421-426.

Baştürk, S. (2014). *Eğitimde ölçme ve değerlendirme*. Ankara: Nobel Yayıncılık.

Can, A. (2018). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi*. 6. Baskı. Ankara: Pegem Akademi.

Daniel, W. (1990). *Applied nonparametric statistics* (2nd edn). Boston: PWS-Kent.

Erkuş, A. (2010). *Davranış bilimleri için bilimsel araştırma süreci*. Ankara: Seçkin Yayıncılık San. ve Tic. Ltd. A.Ş.

Beatty, J. R. (1997). *Statistical methods (Volume Two)*. Newyork: The McGraw – Hill Companies, Inc.

Büyüköztürk, Ş. (2007). *Deneysel desenler*, Ankara: Pegem A Yayıncılık

Büyüköztürk, Ş. (2011). *Sosyal bilimler için veri analizi el kitabı - istatistik, araştırma deseni, spss uygulamaları ve yorum* (15. Baskı). Ankara: Pegem Akademi

Büyüköztürk, Ş. (2017). *Sosyal bilimler için veri analizi el kitabı*. (23. Baskı). Ankara: Pegem A Yayıncılık.

Gerrig, R. J. Ve Zimbardo P. G. (2016). *Psikoloji ve yaşam.* Ankara: Nobel Yayıncılık.

Güriş, S. ve Astar, M. (2015). *Bilimsel araştırmalarda spss ile istatistik*. İstanbul: Der Yayınları.
Huck, S. W. (2008). *Reading statistics and research* (5th ed.). New York: Addison Wesley Longman.
Koç, N. (1986). Personel Seçiminde Psikolojik Testler. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi. 18*(1), 24-34.
Koçyiğit, B. K. (2002). *Likert tipi tutum ölçeklerinin geliştirilmesinde kullanılan bazı tekniklerin karşılaştırılması*. Yayınlanmamış Yüksek Lisans Tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
Köklü, N. (1995). Tutumların ölçülmesi ve likert tipi ölçeklerde kullanılan seçenekler. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi. 28* (2), 81-93.
Kerlinger, F.N. (1973). *Foundation of Behavioural Research*. New York: Holt. Rinehand and Hinston.
Kurt, U. ve Sezek, F. (2018). Yatılı bölge ortaokullarındaki öğrencilerin fiziki mekan memnuniyetlerine yönelik ölçek geliştirme çalışması. *Uşak Üniversitesi Eğitim Araştırmaları Dergisi, 4*(3), 42-57.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives in Psychology,* 140, 1-55.
Özgüven, İ. E. (2007). *Psikolojik testler*. Ankara: PDREM Yayınları.
Salant, Priscilla ve Don Dillman (1994), *How to conduct your own survey?,* New York: John Wiley & Sons, Inc.
Tekin, H. (2000). *Eğitimde ölçme ve değerlendirme* (14. Baskı). Ankara: Yargı Yayınevi.
Tezbaşaran, A. (2004). Likert tipi ölçeklere madde seçmede geleneksel madde analizi tekniklerinin karşılaştırılması. *Türk Psikoloji Dergisi, 19*(54), 77-87.
URL-1 <http://www.atauni.edu.tr/#sayfa=ibm-spss-statistics-20> Erişim tarihi: 01.02.2019
Yazıcıoğlu, Y. ve Erdoğan, S. (2004). *SPSS uygulamalı bilimsel araştırma yöntemleri*, Ankara: Detay Yayıncılık.

Ölçeklerin Değerlendirilmesinde Kullanılan Farklı İstatistiksel Analiz Sonuçlarının Gerçek Bir Örneklem Üzerinde Karşılaştırılması

Genişletilmiş Özet

Bu çalışmanın amacı, eğitim çalışmalarında farklı istatistiksel analiz yöntemlerinin kullanılmasının anket sonuçları veya ölçek değerlendirme sonuçları üzerindeki etkilerini incelemektir. Veri toplamada ölçek ile tarama yöntemi kullanılmıştır. Bu amaç doğrultusunda araştırmacılar tarafından geliştirilen ve öğrencilerin okul memnuniyet düzeyini ölçen "Yatılı Bölge Ortaokullarındaki Öğrencilerin Fiziki Mekân Memnuniyetleri Ölçeği" (ÖFMMÖ) kullanılmıştır. Araştırmanın örneklemini, Erzurum ilinde YBO'larda okuyan 720 öğrenciden oluşmaktadır.

Bu araştırmada, ölçeklerin test maddelerinin puanlanması veya verilen cevapların toplam frekanslarının mı daha doğru sonuç vereceği ölçme bilimi açısından tartışılmıştır. Bunun için, ilk olarak örneklemden elde edilen veriler puanlanmış ve daha sonra testin toplam puanlarında hem parametrik hem de non-parametrik analizler yapılmıştır. Anket ve ölçek puanlamalarıyla istatistiksel analiz yapılmasının pek çok tartışmalı meseleyi

beraberinde getirdiği tespit edilmiştir. Bunların en önemlilerinden birisi duygu ve düşüncelerin puanlanabilirliği meselesidir. Çünkü eğitimde, psikolojide ve sosyolojide ölçülen nitelikler değişkendir. Deneklerin bir konuya, olaya veya duruma karşı tutumları, bir konuda verilen hizmetlerle ilgili algıları ve ölçülmek istenen özellikler sadece kişiden kişiye değil, aynı kişi için farklı durum ve zamanlarda bile farklılıklar gösterebilirler. Araştırmalar; aynı soruya farklı anket yöntemlerinde (posta, telefon, mülakat) bile farklı cevaplar verildiğini göstermektedir (Dillman, 1978). Diğer yandan ölçmeye konu olan davranışların soyut olması ölçülecek niteliği tanımlayan kritik davranışların belirlenmesini, birimlerin herkesçe aynı anlamda kullanılmasını veya anlaşılmasını ve eşitliğinin sağlanmasını zorlaştırmaktadır. Araştırmacı soyut kavramları çok ayrıntılı yazsada, tanımlanan olgu katılımcıların zihninde çok farklı şekillerde canlandırılabilmektedir (Salant ve Dillman, 1974). Diğer yandan farklı insanlar aynı olayı gözlemlediklerinde, sahip oldukları farklı değer yargıları ve kişisel beklentileri farklı anlamalara sebep olabilir. İnsanlar bazen de kendi istedikleri şekilde görür ve duyarlar (Baştürk, 2014). Duygu ve düşünceler sınıflama ve sıralama ölçeğinde olduklarından mutlak veya tanımlanmış sıfır noktaları bulunmaz. Bu nedenle bu verileri puanlamak ve daha sonra bunları toplamak ne kadar doğrudur.

Diğer bir mesele de veriler puanlandığında farklı düşüncelere sahip iki kişinin aynı puana sahip olduklarında aynı kategoride değerlendirilmeleridir. Ayrıca parametrik testlerin uygulanabilmesi içinde verilerin normal dağılım göstermesi gerekmektedir. Bunun için normallik testleri yapılmıştır. Verilerimizin farklı sınıf seviyelerinde basıklık ve çarpıklık değerlerine göre normal, Kolmograv-Simirnov test sonucuna göre ise altıncı sınıflar hariç diğerlerinin normal dağılmadığı görülmektedir. Normal dağılım gösteren altıncı sınıflara parametrik, diğerlerine non-parametrik test uygulanması gerekmektedir. Bu şekilde iki farklı test uygulamak sonuçları karşılaştırmamızı zorlaştıracaktır. Sorunu aşmak için popülasyonun tamamı tek bir örneklem kabul edip normal dağılımına bakılmıştır. Bu işlem sonucunda da bazı veriler non-parametrik, bazı veriler ise parametrik testlerin yapılabileceğini göstermektedir. Bu durumda uygulanacak test analizlerine karar vermeyi son derece güçleştirmektedir. Yukarıda bahsettiğimiz bu karmaşadan dolayı, verilere önce parametrik sonra non-parametrik analizler yaptık. Daha sonra, test sonuçları istatistiksel olarak karşılaştırıldı. Parametrik testlerden ANOVA sonuçlarında testin toplam puanları arasında önem seviyeleri açısından herhangi bir fark bulunamamıştır. Ancak aynı verilere non-parametrik testlerden Kruskal Wallis ve Mann Withney U testi uygulandığında önem seviyeleri açısından fark olduğu görülmüştür. Thamne T2 sonuçlarına göre; 5. ile 6., 6. ile 8. ve 7. ile 8. Sınıflar arasında fark yokken, Mann Withney U sonuçlarına göre yalnızca 7. ile 8. sınıflar arasında fark yoktur. Yani parametrik ve non-parametrik testlerin önem seviyeleri arasında fark bulunmuştur.

Diğer bir analiz şekli ise farklı sınıf seviyelerinde puanların toplanması yerine, öğrencilerin her soruya verdikleri cevapların toplam frekanslarını bularak ki-kare testi uygulanabilir (Tablo 9). Çünkü uygulanan ölçek iki boyutlu, yüksek bir güvenirlik ve geçerliliğe sahip ve bütün soruların olumlu anlam içermesi gibi özelliklerini göz önüne aldığımızda, her bir soruya verilen cevapların puanlarını toplamak yerine her bir şıkkın toplam frekansları üzerinden hesap yapılması yöntem olarak daha uygun gözükmektedir. Ayrıca, öğrencilerin verdikleri cevaplardan katılanların (Katılıyorum + Tamamen Katılıyorum) veya katılmayanların (Katılmıyorum + Hiç Katılmıyorum) toplam % oranları üzerinden memnuniyet sıralamaları tespit edilmiştir. Diğer yandan, sınıfları ikili karşılaştırmanın en etkili yolu cevap yüzdeleri sırasıyla birbirinden çıkartılmıştır. Tablo 9'a göre, çıkartma işleminde üstte 5. sınıflar, alta 6. sınıflar bulunmaktadır. Eğer 5.

sınıfların cevap %'si büyükse pozitif, 6. Sınıfların cevap %'si daha büyükse negatif bir değer çıkacaktır. Buna göre % farklarını hesaplarsak Hiç katılmıyorum= -0,5, Katılmıyorum= -3,3, Karasızım= -3,1, Katılıyorum= -5, Tamamen Katılıyorum= +11,7 sonucu bulunur. Burada 6. sınıfların katılmadıkları ve katıldıkları yönde cevap %'leri fazla iken, 5. sınıfların ise tamamen katıldıkları yönünde cevap % daha fazladır. Yani ankete verilen cevaplardan 5. sınıfların daha olumlu yönde cevap verdikleri anlaşılmaktadır. Bu şekilde bütün sınıflar sırasıyla ikili karşılaştırılarak sınıfların memnuniyet sıralamaları yüzde farklardan bulunmuştur. Ancak hangi sınıflar arasında istatistiksel fark olduğunu anlamak için Ki-Kare testi yapılmıştır. Şekil 1'e baktığımızda, veriler Ki-Kare testi yapmak için SPSS'e işlenirken, **birinci sütuna;** 5. Sınıflar 5, 6. Sınıflar 6, 7. Sınıflar 7 ve 8. Sınıflar 8, **ikinci sütuna;** Hiç Katılmıyorum (1), Katılmıyorum (2), Karasızım (3), Katılıyorum (4), Tamamen Katılıyorum (5) şeklinde kodlanmıştır. Her bir sınıf düzeyindeki öğrencilerin verdikleri cevapların toplam frekansları sırasıyla **üçüncü sütuna** işlenmiştir. Ki-Kare testi sonuçlarına göre yalnızca 7 ile 8. sSınıflar arasında fark yoktur.

Sonuç olarak yukarıda bahsedilen bütün hususlar göz önüne alındığında, öğrencilerin teste verdikleri cevapları frekans olarak kabul edip Ki-Kare testi sonuçlarını karşılaştırdığımızda, verileri puanlayarak toplam test puanı alındığında karşılaşabileceğimiz pek çok problemle uğraşmak zorunda kalmayacağız. Diğer yandan, farklı sınıflarda memnuniyet durumları arasındaki farkların önem seviyeleri Kruskall Wallis ve Mann Whitney U testiyle aynı sonuçları vermektedir.

Parametrik testlerin parametrik olmayan testlerden daha güvenilir olduğuna dair genel düşüncenin aksine ki-kare testi ile karşılaştırmanın, anketlerden elde edilen kategorik verilerin analizinde en uygun yöntemlerden biri olduğu tespit edilmiştir. Ayrıca, farklı sınıflardaki öğrenciler arasında (K5-8), okulların fiziksel ortamlarından memnuniyet düzeyleri arasında istatistiksel olarak anlamlı bir fark olduğu tespit edilmiştir.