Çukurova Üniversitesi Eğitim Fakültesi Dergisi
Vol: 47 No: 1 Sayfa: 260-284
www.cufej.com

Comparison of Classification Judgments Scaling Methods*

Serap BÜYÜKKIDIK[a]
Gazi Üniversitesi, Eğitim Fakültesi, Ankara/Türkiye

Article Info

DOI: 10.14812/cuefd.288600

Article history:
Received 14.02.2017
Revised 16.08.2017
Accepted 18.08.2017

Keywords:
Rubric,
Scaling,
Judgmental decisions scaling,
Classification judgments

Abstract

In this study, 264 performance tasks 6, 7 and 8th grade secondary school students answered was examined by the scaling method with classification criteria to determine whether classification judgements differ or not obtained from the scores of 0-15 by four raters with analytical and holistic rubrics ("0-3" beginning level, "4-6" developable, "7-9" apprentices, "10-12" headworker, "13-15" master).The research is based on descriptive research with the dimension of not aiming generalization and revealing the present situation. In the study, data were collected with two performance tasks, including evidence of validity and reliability and rubrics related to these tasks. As a result of the research, it is seen that scale values and orders of classification judgements of four raters for 264 performances in terms of problem solving ability can be partially changed according to rubric type and the method used. While the scaling method does not cause any difference in the order of scale values in the case of using holistic rubric; it is seen that the rank order of the scale obtained from the B full data matrix and the scale sequences obtained from the B rank numerical solution and D state full matrices are different in the case of using analytical rubric.

Sınıflama Yargılarıyla Ölçekleme Yöntemlerinin Karşılaştırılması

Makale Bilgisi

DOI: 10.14812/cuefd.288600

Makale Geçmişi:
Geliş 14.02.2017
Düzeltme 16.08.2017
Kabul 18.08.2017

Anahtar Kelimeler:
Dereceli puanlama anahtarı,
Ölçekleme,
Yargıcı kararlarına dayalı ölçekleme,
Sınıflama yargıları

Öz

Bu araştırmada ortaokul 6-7-8. sınıf öğrencilerinin yanıtladığı 264 performans görevinin analitik ve bütünsel dereceli puanlama anahtarıyla dört hakem tarafından 0-15 arasında puanlanması sonucunda elde edilen sınıflama yargılarının ("0-3" başlangıç düzeyinde, "4-6" geliştirilebilir, "7-9" çırak, "10-12" kalfa, "13-15" usta) farklılaşıp farklılaşmadığı sınıflama yargılarıyla ölçekleme tekniğiyle incelenmiştir. Araştırma var olan durumu ortaya çıkarması ve genelleme amacı gütmemesi boyutuyla betimsel bir araştırma niteğindedir. Çalışmada geçerliği ve güvenirliğine ilişkin kanıtların toplandığı iki adet performans görevi ve bu görevlere ilişkin dereceli puanlama anahtarları ile veriler toplanmıştır. Araştırma sonucunda dört hakemin 264 performansı problem çözme becerisi bakımından sınıflama yargılarının ölçek değerlerinin ve sırasının kullanılan anahtar türü ve yönteme göre kısmen değişiklik gösterebileceği görülmüştür. Bütünsel dereceli puanlama anahtarı kullanılması durumunda kullanılan ölçekleme yöntemi ölçek değerlerinin sırasında farklılaşmaya neden olmazken, analitik dereceli puanlama anahtarı kullanılması durumunda B Hali tam veri matrisiyle elde edilen ölçek sıraları ile B Hali sayısal çözümü ve D Hali tam verili matristen elde edilen ölçek sıralarının farklılık gösterdiği görülmüştür.

*A part of this study is presented as a verbal statement at 4th International Conference On New Trends in Education and Their Implications

[a] **Author:** sbuyukkidik@gmail.com

Introduction

National Council of Teachers of Mathematics (2000) emphasizes that the evaluation of students should change according to widely accepted learning theories in recent years. The evaluation principle suggests the use of a variety of assessment techniques and tools that assess what learners can do, evaluate what they know, support their mathematics learning and enable learners perform their performance by writing, verbally and actively and methods that allow them to present information in different, unique ways. Performance-based assessment method allows children to understand how they think and put their knowledge into practice. Teachers can blend teaching processes with performance-based assessment to provide additional learning experiences (Brualdi, 1998).

Shepard (2000) states that, evaluation format and its content should be changed to represent each field of problem-solving skills and thinking styles better so that classroom evaluations can be consistent with the constructivist approach. According to Wiggins (1989: p. 41), educational reform achieves the best success with the changes in evaluation system, because the assessment systems "decide what teachers actually teach and what students really know". It has been the sharpest reform to organize open-ended events in which students can solve problems, practice reasoning, practice knowledge in everyday life and therefore determine the performance with the activities implemented (Shepard, 2000).

In the constructivist approach, it is important for students to gain different perspectives, and students are encouraged to find different ways of solving problems with no single correct answer (Vrasidas, 2000). According to Resnick (1987); traditional assessment methods used by teachers examine the cognitive abilities of students in a narrow field and are not related to how they use their knowledge in other disciplines and in their daily lives. As Zollman and Jones (1994) reported; the traditional measurement methods are not devoid of true learning but rather are based on competition, focus on exam skills rather than measuring the accumulation of students, resulting in devaluation in education because students are named as losers compared to others, unreserved, test anxious, have negative attitudes towards schools and teachers, and because of the lack of preliminary knowledge and thinking skills, different orientations such as performance-based assesment have emerged in measuring the characteristics of the learners. As for educational and educational programs reform created by those who make educational policies; It should be taken into account that the performance-based assesment is a valuable tool (Linn, 1993).

Since the rater judgements also come into play in determining performance based situation, the preferred scoring method is important. According to Mertler (2001), the most preferred scoring methods are checklist and scoring scales. In the case of scoring scales, the most commonly used is rubric in performance-based asssessment. A rubric is a one or two-page document, usually listing the criteria for determining a status and grade of quality from the competent to the weak (Andrade, 2001). Deciding whether a rubric is the appropriate method in a performance-based assessment practice is related to the purpose of the measure rather than to which course or class level it is applied (Moskal, 2000). Rubric is divided into two: an analytical rubric that evaluates performance in pieces and a holistic rubric that focuses on performance as a whole (Mertler, 2001). Scoring guidelines are usually formed in two types: result focused (holistic) and process (analytical) focused. It is possible to say that neither the analytical rubric is better than the holistic rubric, nor holistic rubric is better than the analytical rubric. Both are included in the performance evaluation, and it is important to note basic situation and states such as the evaluation objective (process or product), the measured quality (whether it is separated or not), learners raters etc. should be taken into consideration to decide on the type of rubric used (Atılgan, Doğan ve Kan, 2009). Problems have several components (recognition of the problem, definition of the problem, analysis, suggested answers, experience, result of the problem). Each component can be measured with a specific item or it can holistically demonstrate the purpose (solution) of problem solving with an item. This depends on what we are interested in, whether we want to determine our student ability through the steps in the process, or whether we want to follow student ability to make sure it has reached the right answer or not (Haladyna, 1997).

When the studies in literature comparing the analytical and holistic rubrics are analyzed, it is seen that more reliability comparisons were made (Follman & Anderson, 1967; Bauer, 1981; Klein et al., 1998; Boring, 2002; Alharby, 2006; Jonsson & Svingby, 2007). No studies were conducted in literature on the scaling rubrics and comparison of scale values used in performance-based assessment.

The approaches used in scaling can be grouped into two groups, one based on subject reactions and the other based on judge decisions (Turgut and Baykul, 1992). In scaling methods based on judge decisions, observers objectively determine the relative status of each stimulus relative to other stimuli, whereas in approaches based on subject responses, the scaling of responses is aimed, not the substance or stimulus (Torgerson, 1958). Scaling approach with classification judgements used in this study is the scaling approach based on judge decisions.

Classification Judgments Method

Classification judgments method/law is a statistical model that determines the relationship between range limits and the scale values of the stimuli when the stimuli are classified at consecutive intervals (Turgut and Baykul, 1992).

In this study, the sequential interval method, which is one of the applications of the classification judgements law, is used. Sapphire (1937), one of the first to use the sequential interval method, notes that Thurstone refers to this method for the first time. Guilford's absolute scaling method (1938) and Attneave's dual grades method (1949) are in fact considered as the first applications of the sequential interval method. It appears that Thurstone's comparative judicial law was the major influence in the development of the sequential interval method, and Edwards and Thurstone (1952) contributed this method (Turgut and Baykul, 1992).

When the scaling studies in the literature are examined, it is seen that scaling studies are mostly done with pair-wise comparison and rank-order judgment methods (Anıl and Güler, 2006; Nartgün, 2006; Kan, 2008; Öğretmen, 2008; Güler and Anıl, 2009; Bal, 2011; Özer and Acar, 2011; Ekinci, Bindak, ve Yıldırım, 2012; Öztürk, Özdemir and Gelbal, 2012). There is no study on literature on the comparison of the findings obtained numerical solution and full data matrix B case solution and D full data matrix solution of classification judgements scaling studies. There is also no study on the classification of the scores obtained from rubric. In this respect, it is thought that this study will contribute to the literature.

Aim of Research

In this study, it is aimed to scale ratings of classification judgements of student performances which are scored with analytical and holistic rubric with B and D cases and to compare the scale values obtained. With this aim, the following questions were asked:

1. Are the scale values differentiated as a result of classification judgements scaling of the scores obtained from the analytical and holistic rubrics with B case solution?

2. Are the scale values differentiated as a result of classification judgements scaling of the scores obtained from the analytical and holistic rubrics with D case solution?

3. How is the comparison of scale values obtained with B and D cases?

Method

Design

The research is based on descriptive research with the dimension of not aiming generalization and revealing the present situation.

Participants

The study group is comprised of 132 6,7 and 8th grade students in a secondary school in Kütahya in 2011-2012 education year.

Table 1.

Student Demographic Information Frequency and Percentages

Class Levels	Girl		Boy		Total	
	f	%	f	%	f	%
6th Class	28	46,6	32	53,4	60	45,45
7th Class	17	40,4	25	59,5	42	31,81
8th Class	13	43,3	17	56,6	30	22,72
Total					**132**	**100**
Total Performance					**264**	

The students identified in the study were rated by four raters on the basis of 264 performance outcomes in which they demonstrated problem-solving skills. The raters are volunteer mathematics teachers who work in different regions and whose experiences vary from 0 to 5 years.

Data Collection Instrument

First, data were collected with the performance task aimed at transferring students' mathematical problem solving skills to everyday life and later with analytical and holistic rubrics were used to provide scoring. Measuring instruments were developed and improved for validity by taking expert opinion from four measurement and evaluation specialists, a linguist, a mathematics education specialist and nine elementary school mathematics teachers for performance tasks and rubrics. The cronbach alpha reliability coefficients obtained by scoring two performance tasks that measure one-dimensional structure ranged from 0,839 to 0,873 for the analytical rubric, and from 0,834 to 0,863 for the holistic rubric. When we investigated interrater realibility coefficient with intraclass correlation coefficient for analytical rubric it was 0,930 and for holistic rubric it was 0,874. When all these coefficients are examined, it is seen that the measurements are reliable.

Data Analysis

Microsoft Office Excel 2010 program was used for analysis of the data. The following steps were followed in the analysis of the data.

Common Steps in Scaling with B and D State

Stage 1: The scores of the analytical and holistic rubrics yield a frequency matrix of 264 performances obtained by four referees/raters according to five ratings. Table 2 and Table 3 show the frequency matrix.

Table 2.

Frequency Matrix for the Classification of Performances by Using Holistic Rubric

	CLASSES				
Uj	1	2	3	4	5
1	48	47	62	57	50
2	54	51	66	54	39
3	52	48	76	49	39
4	84	55	43	38	44

Table 3.

Frequency Matrix for the Classification of Performances by Using Analytical Rubric

	CLASSES				
Uj	1	2	3	4	5
1	44	35	51	50	84
2	38	41	58	55	72
3	43	33	63	60	65
4	50	42	64	47	61

Stage 2: Cumulative frequency matrix is created. Table 4 and Table 5 show the matrix of cumulative frequencies

Table 4.

Cumulative Frequencies for the Classification of Performances by Using a Holistic Rubric

	CLASSES				
Uj	1	2	3	4	5
1	48	95	157	214	264
2	54	105	171	225	264
3	52	100	176	225	264
4	84	139	182	220	264

Table 5.

Cumulative Frequencies for the Classification of Performances by Using a Analytical Rubric

	CLASSES				
Uj	1	2	3	4	5
1	44	79	130	180	264
2	38	79	137	192	264
3	43	76	139	199	264
4	50	92	156	203	264

Stage 3: Construct a matrix of cumulative ratios. Table 6 and Table 7 show the matrix of the cumulative ratios.

Table 6.

Cumulative Ratios Matrix For The Classification Of Performances by Using Holistic Rubric

	CLASSES			
Uj	1	2	3	4
1	0,182	0,360	0,595	0,811
2	0,205	0,398	0,648	0,852
3	0,197	0,379	0,667	0,852
4	0,318	0,527	0,689	0,833

Table 7.

Cumulative Ratios Matrix For The Classification Of Performances by Using Analytical Rubric

	CLASSES			
Uj	1	2	3	4
1	0,167	0,299	0,492	0,682
2	0,144	0,299	0,519	0,727
3	0,163	0,288	0,526	0,754
4	0,189	0,348	0,591	0,769

Stage 4: The unit normal deviations matrix is formed. Table 8 and Table 9 show the matrix of unit normal deviations.

Table 8.

Unit Normal Deviations Matrix For The Classification Of Performances By Using Holistic Rubric (Z)

	CLASSES				
Uj	1	2	3	4	Zjg
1	-0,908	-0,359	0,240	0,880	-0,037
2	-0,825	-0,259	0,379	1,046	0,085
3	-0,852	-0,309	0,431	1,046	0,079
4	-0,473	0,067	0,494	0,967	0,264

Table 9.

Unit Normal Deviations Matrix For The Classification Of Performances By Using Analytical Rubric (Z)

	CLASSES				
Uj	1	2	3	4	Zjg
1	-0,967	-0,527	-0,019	0,473	-0,260
2	-1,063	-0,527	0,047	0,605	-0,234
3	-0,983	-0,560	0,066	0,686	-0,197
4	-0,88	-0,389	0,230	0,735	-0,076

Scaling from Full Data Matrix by B State

Stage 5: Provide the correct graphic for the two adjacent rows and provide the location where the ordinate axis stops. From the correct graphical equation (y = ax + b), values Sj and t are calculated by using a and b values.

Stage 6: Scale values are calculated using the formula Sj = tg-aj.Zj obtained from $t_g - s_j = Z_{jg}.\sqrt{\sigma_j^2 + B}$ Calculation of scale values for both rubrics is given in Table 10 and Table 11.

Table 10.

Calculation of Scale Values Using Holistic Rubric

Uj	aJ	Zj	aj*Zj	Sj	Sc
1	1,000	-0,037	-0,037	0,256	0,246
2	1,049	0,085	0,089	0,130	0,120
3	1,079	0,079	0,085	0,134	0,124
4	0,792	0,264	0,209	0,010	0

Table 11.

Calculation of Scale Values Using Analytical Rubric

Uj	aJ	Zj	aj*Zj	Sj	Sc
1	1,000	-0,260	-0,260	0,154	0,175
2	1,154	-0,234	-0,270	0,165	0,186
3	1,167	-0,197	-0,230	0,125	0,145
4	1,116	-0,076	-0,085	-0,021	0

Scaling with Numerical Solution B State

Stage 7: In the unit normal deviation matrix, row and column sums are found and averaged. The column averages are the upper bounds of the class.

Stage 8: Standard deviations of the class boundaries (σt) are calculated by taking the deviations of the column averages from the general averages.

Stage 9: Calculate the standard shifts (σzj) of the line elements by taking deviations from the line averages of the line elements or by any other suitable method. Calculation of row-column mean and standard shifts with numerical solution B are given in Table 12 and Table 13.

Table 12.

Calculation Of Row-Column Average And Standard Deviations with B State Numerical Solution Using Holistic Rubric

	SINIFLAR (g)							
Uj	1	2	3	4	topZj	Zj	ss Zj	
1	-0,908	-0,359	0,240	0,880	-0,148	-0,037	0,770	
2	-0,825	-0,259	0,379	1,046	0,341	0,085	0,808	
3	-0,852	-0,309	0,431	1,046	0,316	0,079	0,832	
4	-0,473	0,067	0,494	0,967	1,055	0,264	0,614	
topZj	-3,059	-0,860	1,544	3,940	1,564	Ss t	t ort	r
ORT Zj	-0,765	-0,215	0,386	0,985	0,391	0,755	0,098	0,7555

Table 13.

Calculation Of Row-Column Average And Standard Deviations with B State Numerical Solution Using Analytical Rubric

	SINIFLAR (g)							
Uj	1	2	3	4	topZj	Zj	ss Zj	
1	-0,967	-0,527	-0,019	0,473	-1,040	-0,260	0,624	
2	-1,063	-0,527	0,047	0,605	-0,937	-0,234	0,720	
3	-0,983	-0,560	0,067	0,686	-0,789	-0,197	0,730	
4	-0,880	-0,389	0,230	0,735	-0,304	-0,076	0,706	
topZj	-3,893	-2,002	0,325	2,499	-3,071	Ss t	t ort	r
ORT Zj	-0,973	-0,501	0,081	0,625	-0,768	0,695	-0,192	0,694592

Stage 10: a values and then aj*Zj values are found with $a_j = \frac{\sigma_t}{\sigma_{zj}}$ formula.

Stage 11: The scale values are calculated with $S_j = \bar{t} - a_j.\bar{Z}_j$ formula , the necessary transitions are made so that the starting point is zero. Calculation of the aj and scale values for numerical solution B for both rubrics is given in Table 14 and Table 1.

Table 14.

Calculation of aj and Scale Values with B State Numerical Solution Using Holistic Rubric

aj	aj*Zj	Sj	Sc
0,981	-0,036	0,134	0,361
0,935	0,080	0,018	0,245
0,908	0,072	0,026	0,253
1,231	0,325	-0,227	0,000

Table 15.

Calculation of aj and Scale Values with B State Numerical Solution Using Analytical Rubric

aj	aj*Zj	Sj	Sc
1,114	-0,290	0,098	0,215
0,965	-0,226	0,034	0,151
0,951	-0,188	-0,004	0,113
0,983	-0,075	-0,117	0,000

Scaling from Full Data Matrix with D State

Stage 12: Estimating the class boundary values by taking the column averages of the Z matrix.

Stage 13: The general average of the matrix is then calculated with $S_j' = \bar{z}.. - z.;_j$ formula , and the scale values of the stimuli are estimated by subtracting the line averages from this average. Calculation of the scale values with D for both rubrics is given in Table 16 and Table 17.

Table 16.

Calculation Of Scale Values with D State by Using Holistic Rubric

	SINIFLAR (g)							
Uj	1	2	3	4	totalZj	Zj	Sj	Sc
1	-0,908	-0,359	0,240	0,880	-0,148	-0,037	0,135	0,301
2	-0,825	-0,259	0,379	1,046	0,341	0,085	0,013	0,179
3	-0,852	-0,309	0,431	1,046	0,316	0,079	0,019	0,185
4	-0,473	0,067	0,494	0,967	1,055	0,264	-0,166	0,000
totalZj	-3,059	-0,860	1,544	3,940		1,564		General avarage
Av. Zj	-0,765	-0,215	0,386	0,985		0,391		0,09776

Table 17.

Calculation Of Scale Values with D State by Using Analytical Rubric

	SINIFLAR (g)							
Uj	1	2	3	4	totalZj	Zj	Sj	Sc
1	-0,967	-0,527	-0,019	0,473	-1,040	-0,260	0,068	0,184
2	-1,063	-0,527	0,047	0,605	-0,937	-0,234	0,042	0,158
3	-0,983	-0,560	0,067	0,686	-0,789	-0,197	0,005	0,121
4	-0,880	-0,389	0,230	0,735	-0,304	-0,076	-0,116	0,000
totalZj	-3,893	-2,002	0,325	2,499		-0,768		General avarage
Av. Zj	-0,973	-0,501	0,081	0,625		-3,071		-0,192

After all these steps, the scale values of the four raters that score performance tasks using holistic and analytical rubrics were obtained.

Findings

In a performance based assessment practice, four maths teacher raters are asked for quantization of 264 performances at a "0-3" beginning level, "4-6" developable, "7-9" apprentice, "10-12" headworker, "13-15" master by using analytical and holistic rubrics from 0 to 15 in terms of problem solving skills. Scaling of the obtained classifications by classification criteria in terms of liking/rating from the point of view of the problem solving ability of the four raters was made by means of B State, B State Numerical Solution and D State, and it was tried to investigate whether the findings obtained differ or not.

Findings Related to Scaling with B State Full Data Matrix from Classification Judgements

In Figure 2 and Figure 3 below, the scale values obtained with the B state full data matrix are shown on the number line.

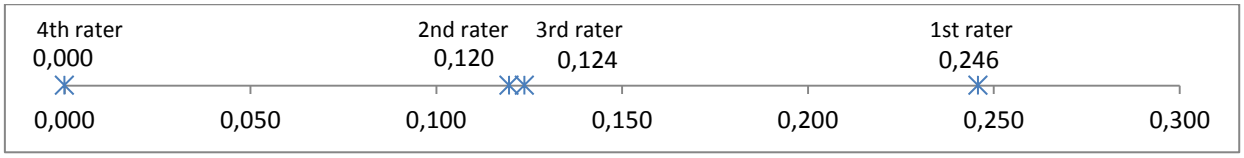


Figure 2. *Displaying The Scale Values Related to Classifications of Each Rater By Using Holistic Rubric On The Number Line (B State Full Data Matrix)*

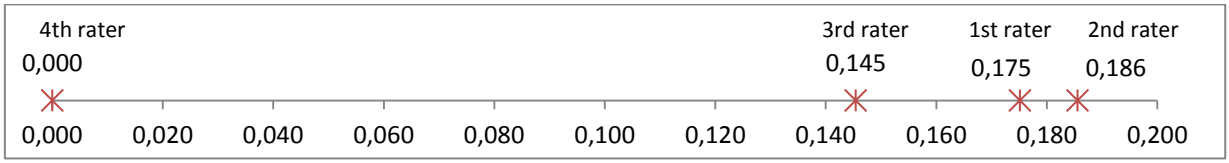


Figure 3. *Displaying The Scale Values Related to Classifications of Each Rater By Using Analytical Rubric On The Number Line (B State Full Data Matrix)*

When Figure 2 and Figure 3 are examined, it is seen that the scale value orders and ranges of the classifications made using holistic and analytical rubrics differ from each other. In the classification made by using the holistic rubric, the scale value of the 4th rater is the lowest, followed by the scale values of the 2nd rater, 3rd rater and 1st rater in terms of classification judgements. In the classification made by using the analytical rubric, the scale value of the 4th rater is the lowest, followed by the scale values of the 3rd rater, 1st rater and 2nd rater in terms of classification judgements. This is also seen when the matrix of frequencies is analyzed in the case of using both two rubrics, with the lowest scaling value 4th rater ranked 264 performances in the lowest classification category. It is seen that rubric used causes a change in the classification status of raters except the fourth rater.

Findings Related to Scaling with Classification Judgements B State Numerical Solution

In the figures below (Figure 4 and Figure 5), the scale values obtained by numerical solution B are shown on the number line.

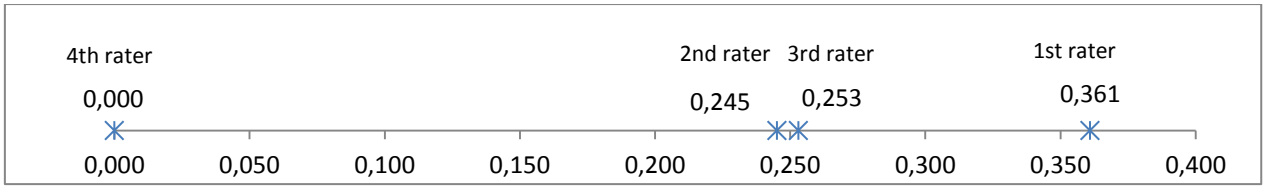


Figure 4. *Displaying The Scale Values Related to Classifications Of Each Rater By Using Holistic Rubric On The Number Line (B State Numerical Solution)*

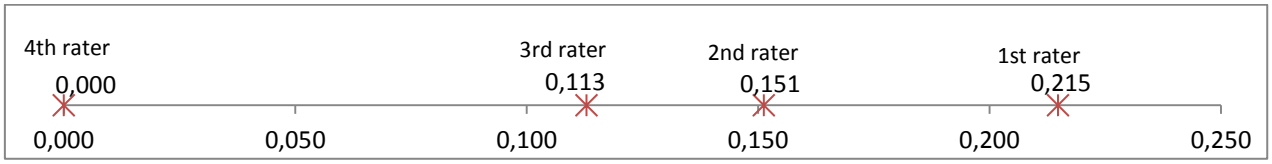


Figure 5. *Displaying The Scale Values Related to Classifications Of Each Rater By Using Analytical Rubric On The Number Line (B State Numerical Solution)*

When Figure 4 and Figure 5 are examined, it is seen that the scale value orders and ranges of the classifications made using holistic and analytical rubrics differ from each other. In the classification made by using the holistic rubric, the scale value of the 4th rater is the lowest, followed by the scale values of the 2nd rater, 3rd rater and 1st rater in terms of classification judgements. In the classification made by using the analytical rubric, the scale value of the 4th rater is the lowest, followed by the scale values of the 3rd rater, 2nd rater and 1st rater in terms of classification judgements. This is also seen when the matrix of frequencies is analyzed in the case of using both two rubrics, with the lowest scaling value 4th rater ranked 264 performances in the lowest classification category. It is seen that rubric used causes a change in the classification status of raters except the fourth and first raters.

Findings Related to Scaling with Classification Judgements D State

In the following figures (Figure 6 and Figure 7), the scale values obtained by D State Full data matrix solution are shown on the number line.

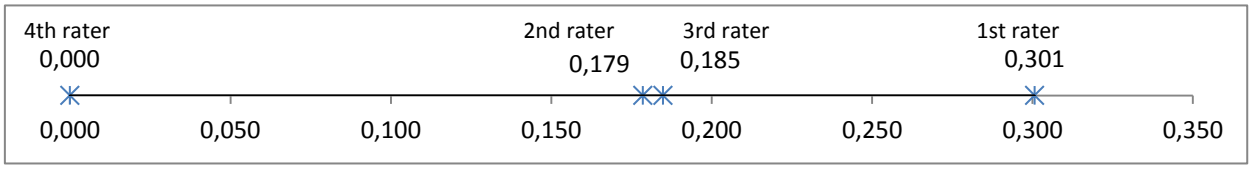


Figure 6. *Displaying The Scale Values Related to Classifications of Each Rater By Using Holistic Rubric on The Number Line (D State Full Data)*

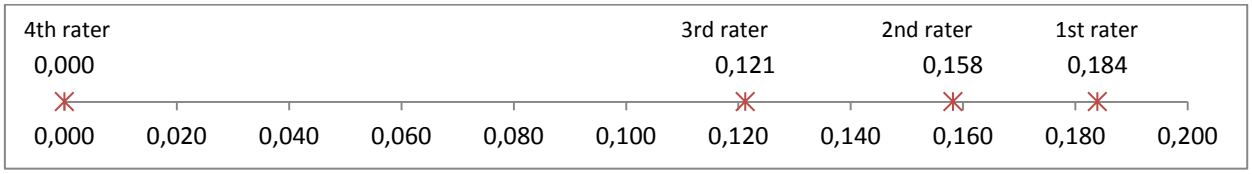


Figure 7. *Displaying The Scale Values Related to Classifications of Each Rater By Using Analytical Rubric on The Number Line (D State Full Data)*

When Figure 6 and Figure 7 are examined, it is seen that the scale value orders and ranges of the classifications made using holistic and analytical rubrics differ from each other. In the classification made by using the holistic rubric, the scale value of the 4th rater is the lowest, followed by the scale values of the 2nd rater, 3rd rater and 1st rater in terms of classification judgements. In the classification made by using the analytical rubric, the scale value of the 4th rater is the lowest, followed by the scale values of the 3rd rater, 2nd rater and 1st rater in terms of classification judgements. This is also seen when the matrix of frequencies is analyzed in the case of using both two rubrics, with the lowest scaling value 4th rater ranked 264 performances in the lowest classification category. It is seen that rubric used causes a change in the classification status of referees/raters except the fourth and first referees/raters.

Comparison Of Scaling Findings with Classification Judgements B And D State

Table 18 compares the scale values obtained by different methods with the holistic and analytical rubrics.

Table 18.

Comparison of Scale Values Obtained By Different Methods with The Holistic and Analytical Rubrics

HOLISTIC RUBRIC				ANALYTICAL RUBRIC			
Referee/ Rater Number	B State Full Data Matrix	B State Numerical Solution	D State	Referee/ Rater Number	B State Full Data Matrix	B State Numerical Solution	D State
Rater1	0,246	0,361	0,301	Rater1	0,175	0,215	0,184
Rater2	0,120	0,245	0,179	Rater2	0,186	0,151	0,158
Rater3	0,124	0,253	0,185	Rater3	0,145	0,113	0,121
Rater4	0,000	0,000	0,000	Rater4	0,000	0,000	0,000

Examining Table 18, it appears that the classification of 264 performances in terms of problem-solving skills using a holistic rubric by four raters produces only a difference in the scale interval and does not cause a change in the order of scale values according to the preferred classification judgements scaling method. It is seen that the classification of 264 performances in terms of problem-solving skills using a analytical rubric by four raters produces a difference in the order of scale values between solution from B full data matrix and other methods according to the preferred classification judgements scaling method and changes intervals partially. The Spearman rho (rs) correlation coefficient was calculated to examine the consistency between the scale values obtained as a result of the scaling operations performed on the B state full data matrix, B state numerical solution, and D state full data matrix. This finding is statistically significant/meaningful ($p < 0,01$), while the findings of the study show that there is a perfect correlation ($rs = 1,00$) between the scaling operations performed on the points obtained with the holistic rubric. There is a partially high relationship between the values obtained from the B full data matrix and other scaling methods ($rs = 0,80$) while the significance level is at 0.05 and 0,01 level non- significant relationship in the points obtained with the analytical rubric. This relationship is statistically significant/meaningful ($p < 0,01$), while there is a perfect correlation ($rs = 1,00$) between B state numerical solution and the scaling operations with D state performed on the points obtained with the analytical rubric.

Discussion, Conclusion, and Suggestions

In this study, it is tried to examine whether rubric type and scaling method with classification judgements are effective on the differentiation of scale values. In the result of the study, it is seen that scale values and orders differ by using analytical and holistic rubric. When classification judgements B state full data scaling are applied on the data obtained from analytical and holistic rubrics, it is seen that the type of rubric used only does not cause a change in the classification of the fourth referee/rater. There is a change in scale orders of the referees except the fourth and the first ones in result of scaling data obtained from analytical and holistic rubrics with B state numerical solution and D state classification judgements. While the scaling method used in the case of using the holistic rubric does not cause any difference in the scale values; it is seen that the rank order of the scale obtained from the B full data matrix and the scale sequences obtained from the B rank numerical solution and D state full matrices are different in the case of using the analytical rubric. Another important result of the study is that when the analytical rubric is used, the scale orders obtained from the B state numerical solution and D state full data matrix are not differentiated, whereas scale orders obtained from the B state full data matrix are different, that is, B state is giving in itself inconsistent scale orders. It is thought that the absence of assumptions in this differentiation may be effective.

When the literature is examined it is seen that there is a limited number of scaling studies (Anıl and Güler, 2006; Nartgün, 2006; Kan, 2008; Güler and Anıl, 2009; Bal, 2011; Özer and Acar, 2011; Ekinci et al., 2012; Öztürk et al., 2012). When these studies are examined, there are few studies to compare the two scaling methods or approaches (Kan, 2008; Öztürk et al., 2011). In addition, no scaling study with classification judgements was found.

In the future studies in scaling with classification judgements of standard setting methods, whether scale values differ or not and the reasons if there is any difference can be examined. Studies can be made to compare different scaling methods. In addition, researchers should consider the assumptions of methods when conducting scaling studies.

Türkçe Sürümü

Giriş

Matematik Ögretmenleri Ulusal Konseyi ([NCTM], 2000) yayımlamış olduğu ilkelerde öğrencinin değerlendirilmesinin son yıllarda yaygın bir şekilde kabul edilen öğrenme teorilerine bağlı olarak değişmesi gerektiğini vurgulamaktadır. Değerlendirme ilkesi, öğrencilerin neyi yapıp yapamadığının yanında, neyi bildiğini değerlendiren, öğrencinin matematik öğrenmelerini destekleyen, öğrencilerin farklı, özgün yollarıyla gösterdikleri bilgilere imkân tanıyan yöntemlerle, öğrencinin yazılı, sözlü ve eylemsel olarak performansını açığa çıkaran çeşitli değerlendirme teknik ve araçların kullanılmasını önermektedir. Performansa dayalı durum belirleme yöntemi çocukların nasıl düşündüğünü ve bilgilerini uygulamaya nasıl koyduğunu anlamayı sağlar. Öğretmenler ilave öğrenme deneyimleri sunmak için öğretim süreçlerini performansa dayalı durum belirleme ile harmanlayabilir (Brualdi, 1998).

Shepard (2000) sınıf içerisinde yapılan değerlendirmelerin yapılandırmacı yaklaşımla uyumlu olması için, ilk olarak; değerlendirme biçiminin ve içeriğinin, her alanda problem çözme becerilerini ve düşünme biçimlerini daha iyi temsil edecek şekilde değiştirilmesi gerektiğini belirtmektedir. Wiggins'e göre (1989: s. 41); "eğitim reformu değerlendirme sistemindeki değişikliklerle en iyi başarıya ulaşır", çünkü değerlendirme sistemleri "öğretmenlerin gerçekten ne öğrettiğine ve öğrencilerin gerçekten neler öğrendiğine karar verir.". Öğrencilerin problem çözebilecekleri, akıl yürütebilecekleri, bilgilerini günlük hayatta uygulayabilecekleri açık uçlu etkinliklerin düzenlenmesi ve dolayısıyla, uygulanacak etkinliklerle performansın belirlenmesi en keskin reform olmuştur (Shepard, 2000).

Yapılandırmacı yaklaşımda, öğrencilerin farklı bakış açıları kazanmaları önemlidir ve öğrenciler tek doğru yanıtı olmayan problemlerle değişik çözüm yolları bulmaya teşvik edilirler (Vrasidas, 2000). Resnick'e göre (1987) ise; öğretmenlerin kullandıkları geleneksel değerlendirme yöntemleri öğrencilerin bilişsel yeteneklerini dar bir alanda incelemekte ve onların edindikleri bilgileri diğer disiplinlerle ve günlük hayatlarında nasıl kullandıkları ile ilişkili olmamaktadır. Zollman ve Jones (1994)'un aktardığı gibi; geleneksel ölçme yöntemlerinin hakiki (gerçek) öğrenme yerine, rekabete dayalı başarıyı sunması, öğrencilerin birikimlerini ölçmek yerine, sınav becerisine odaklanması, bunun neticesinde öğrencilerden bazılarının diğerlerine oranla kaybeden olarak adlandırıldığı, özgüvensiz, sınav kaygısı olan, okul ve öğretmenlere karşı olumsuz tutuma sahip bireyler eğitimde değeri düşürmesi ve ön bilgiyi ve düşünme becerilerini göz ardı etmesi nedeniyle öğrencinin özelliklerinin ölçülmesinde performansa dayalı durum belirleme gibi farklı yönelimler ortaya çıkmıştır. Eğitim politikalarını yapanlar tarafından oluşturulan eğitim ve eğitim programları reformu için; performansa dayalı durum belirlemenin, değerli araç olduğu göz önüne alınılmalıdır (Linn, 1993).

Performansa dayalı durum belirlemede puanlayıcı/hakem kanıları da devreye girdiğinden tercih edilen puanlama yöntemi önemlidir. Mertler'e (2001) göre; puanlama yöntemlerinin en çok tercih edilenleri kontrol listesi ve puanlama ölçekleridir. Puanlama ölçeklerinden ise performansa dayalı durum belirlemede en çok kullanılan dereceli puanlama anahtarı(rubric)dır. Dereceli puanlama anahtarı, genelde bir durum ve nitelik seviyesi belirleme için ölçütleri yetkinden zayıfa listeleyen bir iki sayfalık dökümandır (Andrade, 2001). Bir performansa dayalı durum belirleme uygulamasında dereceli puanlama anahtarının uygun yöntem olup olmadığına karar vermek, hangi ders ya da sınıf düzeyine uygulandığından çok, ölçmenin amacı ile ilgilidir (Moskal, 2000). Puanlama yönergeleri genellikle sonuç (holistik/bütünsel) ve süreç (analitik) odaklı olmak üzere iki türde oluşturulmaktadır. Ne analitik dereceli puanlama anahtarının bütünsel dereceli puanlama anahtarından daha iyi olduğu, ne de bütünsel dereceli puanlama anahtarının analitik dereceli puanlama anahtarından daha iyi olduğunu söylemek mümkündür. Her ikisi de performans değerlendirmenin içinde yer alır ve hangi tip dereceli puanlama anahtarının kullanılacağına karar verirken, değerlendirmenin amacı (Süreç mi? Ürün mü?), ölçülen nitelik (Ögelere ayrışıp, ayrışmadığı), öğrenenler, hakemler vb. gibi değerlendirmeye temel teşkil eden durum

ve koşullar göz önüne alınmalıdır (Atılgan, Doğan ve Kan, 2009). Problemler birkaç bileşene (problemi tanıma, problemi tanımlama, analiz, önerilen yanıtlar, deneyimleme, problemin sonucu) sahiptir. Her bir bileşen özel maddeyle ölçülebilir ya da bir maddeyle problem çözmenin amacını (çözüm) bütünüyle (holistically) gösterebilir. Bu bizim ne ile ilgilendiğimize: süreçteki adımlarla öğrenci yeteneğini belirlemeyi isteyip, istemediğimize ya da öğrenci yeteneğinin doğru yanıta ulaşıp, ulaşmadığı takip etmek istememize, bağlıdır (Haladyna, 1997).

Analitik ve bütünsel dereceli puanlama anahtarlarının karşılaştırılmasına yönelik alanyazındaki çalışmalar incelendiğinde daha çok güvenirlik karşılaştırmaları yapılmıştır (Follman & Anderson, 1967; Bauer, 1981; Klein ve diğerleri, 1998; Boring, 2002; Alharby, 2006; Jonsson & Svingby, 2007). Performansa dayalı durum belirlemede kullanılan dereceli puanlama anahtarlarının ölçeklenmesi ve ölçek değerlerinin karşılaştırılmasına yönelik alanyazında herhangi bir çalışmaya rastlanmamıştır.

Ölçeklemede kullanılan yaklaşımlar denek tepkilerine dayalı ve yargıcı kararlarına dayalı yaklaşımlar olmak üzere iki grupta toplanabilir (Turgut ve Baykul, 1992). Yargıcı kararlarına dayalı ölçekleme yöntemlerinde gözlemciler tarafsız olarak her bir uyarıcının diğer uyarıcılara göre göreceli durumunu belirlerken, denek tepkilerine dayalı yaklaşımlarda ise madde ya da uyarıcının değil cevapların ölçeklenmesi amacını güdülür (Torgerson, 1958). Bu araştırmada kullanılan sınıflama yargılarıyla ölçekleme yaklaşımı yargıcı kararlarına dayalı ölçekleme yöntemlerindendir.

Sınıflama Yargılarıyla Ölçekleme

Sınıflama yargıları kanunu, uyarıcıların ardışık aralıklarla sınıflandığı durumlarda, aralık sınırlarıyla uyarıcıların ölçek değerleri arasındaki ilişkileri belirleyen bir istatistiksel modeldir (Turgut ve Baykul, 1992).

Bu çalışmada sınıflama yargıları kanunun uygulamalarından biri olan ardışık aralıklar yöntemi kullanılmıştır. Ardışık aralıklar yöntemini ilk kullananlardan biri olan Saffir (1937), bu yöntemden ilk defa Thurstone'un söz ettiğini belirtmektedir. Guilford'un mutlak ölçekleme metodu (1938) ve Attneave'in ikili derecelemeler metodu (1949), aslında ardışık aralıklar yönteminin ilk uygulamaları olarak nitelendirilmektedir. Ardışık aralıklar yönteminin geliştirilmesinde Thurstone'nun karşılaştırmalı yargılar kanunun büyük etkisinin olduğu ve ayrıca, Edwards ve Thurstone'un (1952) bu yönteme katkısı olduğu görülmektedir (Turgut ve Baykul, 1992).

Alanyazındaki ölçekleme çalışmaları incelendiğinde daha çok ikili karşılaştırmalar ve sıralama yargılarıyla ölçekleme çalışmalarının yapıldığı görülmektedir (Anıl ve Güler, 2006; Nartgün, 2006; Kan, 2008; Öğretmen, 2008; Güler ve Anıl, 2009; Bal, 2011; Özer ve Acar, 2011; Ekinci, Bindak ve Yıldırım, 2012; Öztürk, Özdemir ve Gelbal, 2012). Sınıflama yargılarıyla ölçekleme çalışmalarının B Hali tam veri matrisi ve sayısal çözüm ve D Hali tam verili matrisle elde edilen bulgularının karşılaştırılmasına ilişkin alanyazında bir çalışma bulunmamaktadır. Ayrıca dereceli puanlama anahtarından elde edilen puanların sınıflama yargılarıyla ölçeklenmesine ilişkin de herhangi bir çalışma yoktur. Bu bakımdan, bu çalışmanın alanyazına katkı sağlayacağı düşünülmektedir.

Araştırmanın Amacı

Bu araştırmada analitik ve bütünsel dereceli puanlama anahtarıyla puanlanılan öğrenci performanslarının beğenilme durumlarının sınıflama yargılarının B ve D Halleriyle ölçeklenmesi ve elde edilen ölçek değerlerinin karşılaştırılması amaçlanmıştır. Bu amaçla aşağıdaki sorulara yanıt aranmıştır:

1. Analitik ve bütünsel dereceli puanlama anahtarından elde edilen puanların sınıflama yargıları B Hali çözümüyle ölçeklenmesi sonucunda ölçek değerleri farklılaşmakta mıdır?

2. Analitik ve bütünsel dereceli puanlama anahtarından elde edilen puanların sınıflama yargıları D Hali çözümüyle ölçeklenmesi sonucunda ölçek değerleri farklılaşmakta mıdır?

3. B ve D Halleriyle elde edilen ölçek değerlerinin karşılaştırması nasıldır?

Yöntem

Araştırma Türü

Araştırma var olan durumu ortaya çıkarıp, genelleme amacı gütmemesi boyutuyla betimsel bir araştırma niteğindedir.

Çalışma Grubu

Araştırmanın çalışma grubunu, 2011-2012 eğitim-öğretim yılında Kütahya İline bağlı bir ortaokulda öğrenim gören 132 ilköğretim ikinci kademe 6.,7.,8. sınıf öğrencisi oluşturmuştur.

Tablo 1.

Öğrenci Demografik Bilgiler Frekans ve Yüzdeleri

Sınıf	**Kız**		**Erkek**		**Toplam**	
Düzeyleri	**f**	**%**	**f**	**%**	**f**	**%**
6. Sınıf	28	46,6	32	53,4	60	45,45
7. Sınıf	17	40,4	25	59,5	42	31,81
8. Sınıf	13	43,3	17	56,6	30	22,72
Toplam					**132**	**100**
Toplam Performans					**264**	

Çalışmada belirlenen öğrencilerin problem çözme becerilerini gösterdikleri 264 performans doğrultusunda dört hakem/puanlayıcı tarafından puanlanmıştır. Hakemleri, farklı bölgelerde görev yapan, 0-5 yıl arasında deneyimleri değişen, gönüllü matematik öğretmenleri oluşturmuştur.

Veri Toplama Aracı

Önce öğrencilerin günlük yaşama matematiksel problem çözme becerisini aktarmayı hedefleyen iki adet performans göreviyle uygulama yapılmıştır. Ardından puanlamayı yapmak için analitik ve bütünsel dereceli puanlama anahtarı kullanılarak veriler toplanmıştır. Performans görevleri ve dereceli puanlama anahtarları için dört ölçme ve değerlendirme uzmanından, bir dilbilimciden, bir matematik eğitimi uzmanından ve dokuz ilköğretim matematik öğretmeninden uzman görüşü alınarak, ölçme araçları geçerlik yönünden gözden geçirilip, geliştirilmiştir. Dört puanlayıcının tek boyutlu yapıyı ölçen iki performans görevini puanlamasından elde edilen ölçümlerde Cronbach alpha güvenirlik katsayıları analitik dereceli puanlama anahtarı için 0,839-0,873 arasında değişirken, bütünsel dereceli puanlama anahtarı için 0,834-0,863 arasında değişmektedir. Yine puanlayıcılar arası tutarlılıklar sınıf içi ilişki katsayısı ile incelendiğinde analitik dereceli puanlama anahtarı için 0,930, bütünsel dereceli puanlama anahtarı için 0,874 olduğu görülmüştür. Tüm bu katsayılar incelendiğinde ölçümlerin güvenilir olduğu görülmektedir.

Verilerin Analizi

Verilerin analizinde Microsoft Office Excel 2010 programından yararlanılmıştır. Verilerin analizinde aşağıdaki basamaklar izlenmiştir.

B ve D Haliyle Ölçeklemede Ortak Aşamalar

1. Aşama: Analitik ve bütünsel dereceli puanlama anahtarından elde edilen puanlara 264 performansın dört hakem tarafından beş sınıflı beğenilme durumlarına göre elde edilen frekanslar matrisinin elde edilir. Tablo 2 ve Tablo 3'te frekanslar matrisine yer verilmiştir.

Tablo 2.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Frekanslar Matrisi

	SINIFLAR				
Uj	1	2	3	4	5
1	48	47	62	57	50
2	54	51	66	54	39
3	52	48	76	49	39
4	84	55	43	38	44

Tablo 3.

Analitik Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Frekanslar Matrisi

	SINIFLAR				
Uj	1	2	3	4	5
1	44	35	51	50	84
2	38	41	58	55	72
3	43	33	63	60	65
4	50	42	64	47	61

2. Aşama: Yığmalı frekanslar matrisinin oluşturulur. Tablo 4 ve Tablo 5'te yığmalı frekanslar matrisine yer verilmiştir.

Tablo 4.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Yığmalı Frekanslar Matrisi

	SINIFLAR				
Uj	1	2	3	4	5
1	48	95	157	214	264
2	54	105	171	225	264
3	52	100	176	225	264
4	84	139	182	220	264

Tablo 5.

Analitik Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Yığmalı Frekanslar Matrisi

	SINIFLAR				
Uj	1	2	3	4	5
1	44	79	130	180	264
2	38	79	137	192	264
3	43	76	139	199	264
4	50	92	156	203	264

3. Aşama: Yığmalı oranlar matrisinin oluşturulur. Tablo 6 ve Tablo 7'de yığmalı oranlar matrisine yer verilmiştir.

Tablo 6.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Yığmalı Oranlar Matrisi

	SINIFLAR			
Uj	1	2	3	4
1	0,182	0,360	0,595	0,811
2	0,205	0,398	0,648	0,852
3	0,197	0,379	0,667	0,852
4	0,318	0,527	0,689	0,833

Tablo 7.

Analitik Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Yığmalı Oranlar Matrisi

	SINIFLAR			
Uj	1	2	3	4
1	0,167	0,299	0,492	0,682
2	0,144	0,299	0,519	0,727
3	0,163	0,288	0,526	0,754
4	0,189	0,348	0,591	0,769

4. Aşama: Birim normal sapmalar matrisinin oluşturulur. Tablo 8 ve Tablo 9'da birim normal sapmalar matrisine yer verilmiştir.

Tablo 8.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Birim Normal Sapmalar Matrisi (Z)

	SINIFLAR				
Uj	1	2	3	4	Zjg
1	-0,908	-0,359	0,240	0,880	-0,037
2	-0,825	-0,259	0,379	1,046	0,085
3	-0,852	-0,309	0,431	1,046	0,079
4	-0,473	0,067	0,494	0,967	0,264

Tablo 9.

Analitik Dereceli Puanlama Anahtarı Kullanılarak Performansların Sınıflanmasına Ait Birim Normal Sapmalar Matrisi (Z)

	SINIFLAR				
Uj	1	2	3	4	Zjg
1	-0,967	-0,527	-0,019	0,473	-0,260
2	-1,063	-0,527	0,047	0,605	-0,234
3	-0,983	-0,560	0,066	0,686	-0,197
4	-0,88	-0,389	0,230	0,735	-0,076

B Haliyle Tam Veri Matrisinden Ölçekleme

B Hali tam veri matrisiyle ölçeklemede ek olarak 5. ve 6. aşamalar izlenir.

5. Aşama: Birbirine komşu iki satır için doğru grafiği oluşturulup, ordinatlar eksenini kestiği yerin temin edilmesi. Doğru grafiği denkleminden (y=ax+b) a ve b değerleri kullanılarak Sj ve t değerleri hesaplanır.

6. Aşama: $t_g - s_j = Z_{jg} . \sqrt{\sigma_j^2 + B}$ formülünden elde edilen Sj=tg-aj.Zj formülü kullanılarak ölçek değerleri hesaplanır. Ölçek değerlerinin her iki dereceli puanlama anahtarı için hesaplanmasına Tablo 10 ve Tablo 11'de yer verilmiştir.

Tablo 10.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Ölçek Değerlerinin Hesaplanması İşlemleri

Uj	aJ	Zj	aj*Zj	Sj	Sc
1	1,000	-0,037	-0,037	0,256	0,246
2	1,049	0,085	0,089	0,130	0,120
3	1,079	0,079	0,085	0,134	0,124
4	0,792	0,264	0,209	0,010	0

Tablo 11.

Analitik Dereceli Puanlama Anahtarı Kullanılarak Ölçek Değerlerinin Hesaplanması İşlemleri

Uj	aJ	Zj	aj*Zj	Sj	Sc
1	1,000	-0,260	-0,260	0,154	0,175
2	1,154	-0,234	-0,270	0,165	0,186
3	1,167	-0,197	-0,230	0,125	0,145
4	1,116	-0,076	-0,085	-0,021	0

B Haliyle Sayısal Çözümle Ölçekleme

7. Aşama: Birim normal sapmalar matrisinde satır ve sütun toplamları bulunup, ortalaması alınır. Sütun ortalamaları sınıf üst sınırlarıdır.

8. Aşama: Sütun ortalamalarının genel ortalamadan sapmaları alınarak sınıf üst sınırlarının standart kayması (σt) bulunur.

9. Aşama: Satır elemanlarının satır ortalamasından sapmaları alınarak ya da uygun diğer bir yöntemle, satır elemanlarının standart kaymaları (σzj) hesaplanır. B Hali sayısal çözümüyle satır-sütun ortalama ve standart kaymaların hesaplanması işlemlerine Tablo 12 ve Tablo 13'te yer verilmiştir.

Tablo 12.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak B Hali Sayısal Çözümüyle Satır-Sütun Ortalama ve Standart Kaymaların Hesaplanması İşlemleri

	SINIFLAR (g)							
Uj	1	2	3	4	topZj	Zj	ss Zj	
1	-0,908	-0,359	0,240	0,880	-0,148	-0,037	0,770	
2	-0,825	-0,259	0,379	1,046	0,341	0,085	0,808	
3	-0,852	-0,309	0,431	1,046	0,316	0,079	0,832	
4	-0,473	0,067	0,494	0,967	1,055	0,264	0,614	
topZj	-3,059	-0,860	1,544	3,940	1,564	Ss t	t ort	r
ORT Zj	-0,765	-0,215	0,386	0,985	0,391	0,755	0,098	0,7555

Tablo 13.

Analitik Dereceli Puanlama Anahtarı Kullanılarak B Hali Sayısal Çözümüyle Satır-Sütun Ortalama ve Standart Kaymaların Hesaplanması İşlemleri

	SINIFLAR (g)							
Uj	1	2	3	4	topZj	Zj	ss Zj	
1	-0,967	-0,527	-0,019	0,473	-1,040	-0,260	0,624	
2	-1,063	-0,527	0,047	0,605	-0,937	-0,234	0,720	
3	-0,983	-0,560	0,067	0,686	-0,789	-0,197	0,730	
4	-0,880	-0,389	0,230	0,735	-0,304	-0,076	0,706	
topZj	-3,893	-2,002	0,325	2,499	-3,071	Ss t	t ort	r
ORT Zj	-0,973	-0,501	0,081	0,625	-0,768	0,695	-0,192	0,694592

10. Aşama: $a_j = \frac{\sigma_t}{\sigma_{zj}}$,formülüyle a değerleri bulunur, ardından aj*Zj değerleri bulunur.

11. Aşama: $S_j = \bar{t} - a_j . \bar{Z}_j$ formülüyle ölçek değerleri hesaplanır, başlangıç noktası sıfır olacak şekilde gerekli ötelemeler yapılır. Her iki dereceli puanlama anahtarı için B Hali sayısal çözümüyle aj ve ölçek değerlerinin hesaplanması işlemlerine Tablo 14 ve Tablo 15'te yer verilmiştir.

Tablo 14.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak B Hali Sayısal Çözümüyle aj ve Ölçek Değerlerinin Hesaplanması İşlemleri

aj	aj*Zj	Sj	Sc
0,981	-0,036	0,134	0,361
0,935	0,080	0,018	0,245
0,908	0,072	0,026	0,253
1,231	0,325	-0,227	0,000

Tablo 15.

Analitik Dereceli Puanlama Anahtarı Kullanılarak B Hali Sayısal Çözümüyle aj ve Ölçek Değerlerinin Hesaplanması İşlemleri

aj	aj*Zj	Sj	Sc
1,114	-0,290	0,098	0,215
0,965	-0,226	0,034	0,151
0,951	-0,188	-0,004	0,113
0,983	-0,075	-0,117	0,000

D Haliyle Tam Veri Matrisinden Ölçekleme

12. Aşama: Z matrisinin sütun ortalamaları alınarak sınıf sınır değerleri kestirilir.

13. Aşama: Daha sonra $S_j' = \bar{z}.. - z.;_j$ formülüyle matrisin genel ortalaması hesaplanmış ve bu ortalamadan satır ortalamaları çıkarılarak uyarıcıların ölçek değerleri kestirilir. Her iki dereceli puanlama anahtarı için D Haliyle ölçek değerlerinin hesaplanması işlemlerine Tablo 16 ve Tablo 17'de yer verilmiştir.

Tablo 16.

Bütünsel Dereceli Puanlama Anahtarı Kullanılarak D Haliyle Ölçek Değerlerinin Hesaplanması İşlemleri

	SINIFLAR (g)							
Uj	1	2	3	4	topZj	Zj	Sj	Sc
1	-0,908	-0,359	0,240	0,880	-0,148	-0,037	0,135	0,301
2	-0,825	-0,259	0,379	1,046	0,341	0,085	0,013	0,179
3	-0,852	-0,309	0,431	1,046	0,316	0,079	0,019	0,185
4	-0,473	0,067	0,494	0,967	1,055	0,264	-0,166	0,000
topZj	-3,059	-0,860	1,544	3,940		1,564		Genel ortalama
ORT Zj	-0,765	-0,215	0,386	0,985		0,391		0,09776

Tablo 17.

Analitik Dereceli Puanlama Anahtarı Kullanılarak D Haliyle Ölçek Değerlerinin Hesaplanması İşlemleri

	SINIFLAR (g)							
Uj	1	2	3	4	topZj	Zj	Sj	Sc
1	-0,967	-0,527	-0,019	0,473	-1,040	-0,260	0,068	0,184
2	-1,063	-0,527	0,047	0,605	-0,937	-0,234	0,042	0,158
3	-0,983	-0,560	0,067	0,686	-0,789	-0,197	0,005	0,121
4	-0,880	-0,389	0,230	0,735	-0,304	-0,076	-0,116	0,000
topZj	-3,893	-2,002	0,325	2,499		-0,768		Genel ortalama
ORT Zj	-0,973	-0,501	0,081	0,625		-3,071		-0,192

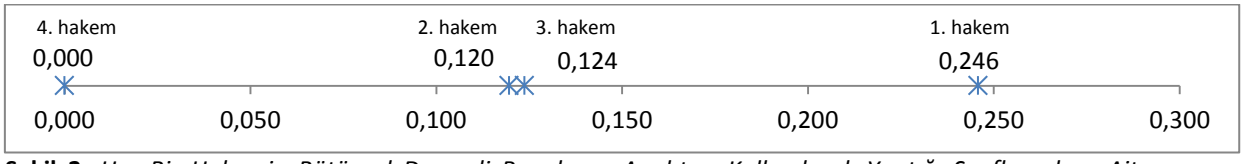
Tüm bu aşamaların ardından bütünsel ve analitik dereceli puanlama anahtarlarını kullanarak performans görevlerini puanlayan dört hakemin kararlarının ölçek değerleri elde edilmiştir.

Bulgular

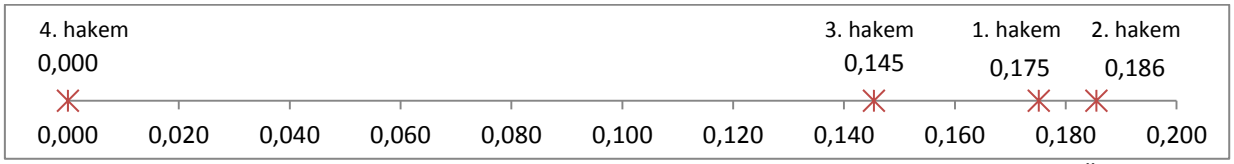
Bir performansa dayalı durum belirleme uygulamasında matematik öğretmeni dört hakemin 264 performansı "0-3" başlangıç düzeyinde, "4-6" geliştirilebilir, "7-9" çırak, "10-12" kalfa, "13-15" usta sınıflandırma düzeylerini göstermek üzere; problem çözme becerileri bakımından 0-15 arasında analitik ve bütünsel dereceli puanlama anahtarı kullanarak puanlamaları istenmiştir. Elde edilen sınıflandırmaların dört hakemin problem çözme becerisi bakımından beğenme bakımından sınıflama yargılarıyla ölçeklenmesi B Hali, B Hali sayısal çözümü ve D Haliyle yapılmış ve elde edilen bulguların farklılaşıp farklılaşmadığı incelenmeye çalışılmıştır.

Sınıflama Yargıları Tam Verili Matristen B Haliyle Ölçekleme İle İlgili Bulgular

Aşağıdaki Şekil 2 ve Şekil 3'te B Hali tam veri matrisiyle elde edilen ölçek değerlerinin sayı doğrusunda gösterimlerine yer verilmiştir.



Şekil 2. *Her Bir Hakemin Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Yaptığı Sınıflamalara Ait Ölçek Değerlerinin Sayı Doğrusunda Gösterilmesi (B Hali Tam Veri Matrisi)*

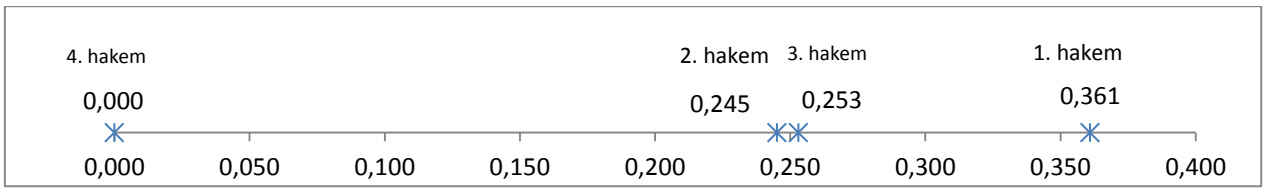


Şekil 3. *Her Bir Hakemin Analitik Dereceli Puanlama Anahtarı Kullanılarak Yaptığı Sınıflamalara Ait Ölçek Değerlerinin Sayı Doğrusunda Gösterilmesi (B Hali Tam Veri Matrisi)*

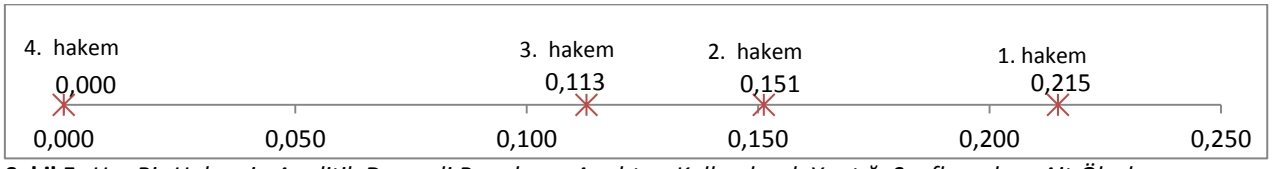
Şekil 2 ve Şekil 3 incelendiğinde bütünsel ve analitik dereceli puanlama anahtarları kullanılarak yapılan sınıflandırmaların ölçek değeri sıralarının ve aralıklarının farklılaştığı görülmektedir. Bütünsel dereceli puanlama anahtarı kullanılarak yapılan sınıflandırmada 4. Hakemin ölçek değeri en düşük, ardından 2. Hakemin, 3. Hakemin ve 1. Hakemin sınıflandırma yargılarının ölçek değerleri gelirken; analitik dereceli puanlama anahtarı kullanılarak yapılan sınıflandırmada ise 4. Hakemin ölçek değeri en düşük, ardından 3. Hakemin, 1. Hakemin ve 2. Hakemin sınıflandırma yargılarının ölçek değerleri gelmektedir. Bu durum her iki dereceli puanlama anahtarı kullanılması durumunda frekanslar matrisi incelendiğinde 4. Hakemin 264 performansı en düşük sınıflama kategorisine koyduğu en düşük ölçek değeriyle de görülmektedir. Kullanılan dereceli puanlama anahtarının dördüncü hakem dışındaki hakemlerin sınıflandırma durumlarında değişikliğe neden olduğu görülmektedir.

Sınıflama Yargıları B Hali Sayısal Çözümle Ölçekleme İle İlgili Bulgular

Aşağıdaki şekillerde (Şekil 4 ve Şekil 5) B Hali sayısal çözümle elde edilen ölçek değerlerinin sayı doğrusunda gösterimlerine yer verilmiştir.



Şekil 4. *Her Bir Hakemin Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Yaptığı Sınıflamalara Ait Ölçek Değerlerinin Sayı Doğrusunda Gösterilmesi (B Hali Sayısal Çözüm)*

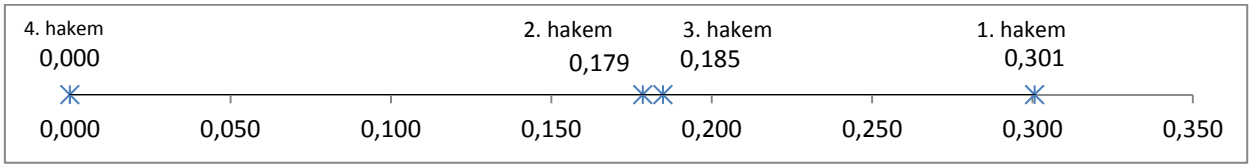


Şekil 5. *Her Bir Hakemin Analitik Dereceli Puanlama Anahtarı Kullanılarak Yaptığı Sınıflamalara Ait Ölçek Değerlerinin Sayı Doğrusunda Gösterilmesi (B Hali Sayısal Çözüm)*

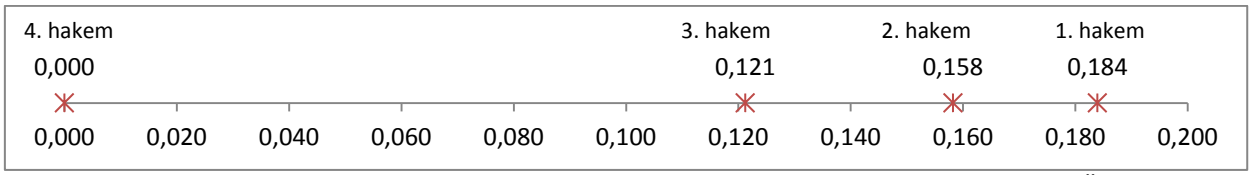
Şekil 4 ve Şekil 5 incelendiğinde bütünsel ve analitik dereceli puanlama anahtarları kullanılarak yapılan sınıflandırmaların ölçek değeri sıralarının ve aralıklarının farklılaştığı görülmektedir. Bütünsel dereceli puanlama anahtarı kullanılarak yapılan sınıflandırmada 4. Hakemin ölçek değeri en düşük, ardından 2. Hakemin, 3. Hakemin ve 1. Hakemin sınıflandırma yargılarının ölçek değerleri gelirken; analitik dereceli puanlama anahtarı kullanılarak yapılan sınıflandırmada ise 4. Hakemin ölçek değeri en düşük, ardından 3. Hakemin, 2. Hakemin ve 1. Hakemin sınıflandırma yargılarının ölçek değerleri gelmektedir. Bu durum her iki dereceli puanlama anahtarı kullanılması durumunda frekanslar matrisi incelendiğinde 4. Hakemin 264 performansı en düşük sınıflama kategorisine koyduğu en düşük ölçek değeriyle de görülmektedir. Kullanılaran dereceli puanlama anahtarının ikinci ve üçüncü hakemin yaptığı sınıflamalara ait ölçek sıralarında değişikliğe neden olduğu bulunmuştur.

Sınıflama Yargıları D Haliyle Ölçekleme İle İlgili Bulgular

Aşağıdaki şekillerde (Şekil 6 ve Şekil 7) D Hali tam veri matrisiyle çözümle elde edilen ölçek değerlerinin sayı doğrusunda gösterimlerine yer verilmiştir.



Şekil 6. *Her Bir Hakemin Bütünsel Dereceli Puanlama Anahtarı Kullanılarak Yaptığı Sınıflamalara Ait Ölçek Değerlerinin Sayı Doğrusunda Gösterilmesi (D Hali Tam Verili)*



Şekil 7. *Her Bir Hakemin Analitik Dereceli Puanlama Anahtarı Kullanılarak Yaptığı Sınıflamalara Ait Ölçek Değerlerinin Sayı Doğrusunda Gösterilmesi (D Hali Tam Verili)*

Şekil 6 ve Şekil 7 incelendiğinde bütünsel ve analitik dereceli puanlama anahtarları kullanılarak yapılan sınıflandırmaların ölçek değeri sıralarının ve aralıklarının farklılaştığı görülmektedir. Bütünsel dereceli puanlama anahtarı kullanılarak yapılan sınıflandırmada 4. Hakemin ölçek değeri en düşük, ardından 2. Hakemin, 3. Hakemin ve 1. Hakemin sınıflandırma yargılarının ölçek değerleri gelirken; analitik dereceli puanlama anahtarı kullanılarak yapılan sınıflandırmada ise 4. Hakemin ölçek değeri en düşük, ardından 3. Hakemin, 2. Hakemin ve 1. Hakemin sınıflandırma yargılarının ölçek değerleri gelmektedir. Bu durum her iki dereceli puanlama anahtarı kullanılması durumunda frekanslar matrisi incelendiğinde 4. Hakemin 264 performansı en düşük sınıflama kategorisine koyduğu en düşük ölçek değeriyle de görülmektedir. Kullanılan dereceli puanlama anahtarının dördüncü hakem ve birinci hakem dışındaki hakemlerin sınıflandırma sıralarında değişikliğe neden olduğu görülmektedir.

Sınıflama Yargıları B Ve D Halleriyle Ölçekleme Bulgularının Karşılaştırılması

Tablo 18'de bütünsel ve analitik dereceli puanlama anahtarıyla farklı yöntemlerle elde edilen ölçek değerlerinin karşılaştırılması yapılmıştır.

Tablo 18.

Bütünsel ve Analitik Dereceli Puanlama Anahtarıyla Farklı Yöntemlerle Elde Edilen Ölçek Değerlerinin Karşılaştırılması

BÜTÜNSEL DERECELİ PUANLAMA ANAHTARI				ANALİTİK DERECELİ PUANLAMA ANAHTARI			
Hakem No	B Hali Tam Veri Matrisi	B Hali Sayısal Çözüm	D Hali	Hakem No	B Hali Tam Veri Matrisi	B Hali Sayısal Çözüm	D Hali
Hakem1	0,246	0,361	0,301	Hakem1	0,175	0,215	0,184
Hakem2	0,120	0,245	0,179	Hakem2	0,186	0,151	0,158
Hakem3	0,124	0,253	0,185	Hakem3	0,145	0,113	0,121
Hakem4	0,000	0,000	0,000	Hakem4	0,000	0,000	0,000

Tablo 18 incelendiğinde bütünsel dereceli puanlama anahtarı kullanılarak dört hakemin 264 performansı problem çözme becerisi bakımından sınıflandırmalarının tercih edilen sınıflama yargılarıyla ölçekleme yöntemine göre ölçek değerlerinin sırasında bir değişikliğe neden olmadığı yalnızca ölçek aralığında bir farklılaşma oluşturduğu görülmektedir. Analitik dereceli puanlama anahtarı kullanılarak dört hakemin 264 performansı problem çözme becerisi bakımından sınıflandırmalarının tercih edilen sınıflama yargılarıyla ölçekleme yöntemine göre ölçek değerlerinin sırasında B Hali tam veri matrisinden çözüm ile diğer yöntemler arasında bir farklılaşmaya sebep olduğu, aynı zamanda aralıkları da kısmen değiştirdiği görülmektedir. B Hali tam verili matristen, B Hali sayısal çözümden ve D Hali tam verili matrisle yapılan ölçekleme işlemleri sonucunda elde edilen ölçek değerleri arasındaki tutarlığı incelemek için Spearman rho (rs) korelasyon katsayısı hesaplanmıştır. Çalışmanın bulguları, bütünsel dereceli puanlama anahtarı ile elde edilen puanlarda yapılan ölçekleme işlemleri arasında mükemmel düzeyde (rs=1,00) bir ilişki olduğunu gösterirken, bu ilişki istatistiksel olarak manidardır (p<0,01). Analitik dereceli puanlama anahtarı ile elde edilen puanlarda yapılan B Hali tam verili matristen elde edilen değerler ile diğer ölçekleme yöntemleri arasında kısmen yüksek düzeyde (rs=0,80) ve 0,05 düzeyinde manidar, 0,01 düzeyinde ise manidar olmayan bir ilişki vardır. Analitik dereceli puanlama anahtarı ile elde edilen puanlarda yapılan B Hali sayısal çözüm ve D hali ile ölçekleme işlemleri arasında ise mükemmel düzeyde (rs=1,00) bir ilişki olduğunu gösterirken, bu ilişki istatistiksel olarak manidardır (p<0,01).

Sonuç, Tartışma ve Öneriler

Bu araştırmada dereceli puanlama anahtarı türü ve kullanılan sınıflama yargılarıyla ölçekleme yönteminin ölçek değerlerinin farklılaşmasında etkili olup olmadığı incelenmeye çalışılmıştır. Araştırma sonucunda analitik ve bütünsel dereceli puanlama anahtarı kullanılarak elde edilen ölçek değerlerinin ve sırasının farklılaştığı görülmüştür. Sınıflama yargıları B Hali tam veri matrisi ile ölçekleme analitik ve bütünsel dereceli puanlama anahtarından elde edilen verilere uygulandığında kullanılan dereceli puanlama anahtarı türünün sadece dördüncü hakemin sınıflandırma sırasında değişikliğe neden olmadığı görülmektedir. Analitik ve bütünsel dereceli puanlama anahtarından elde edilen verilerin B Hali sayısal çözümle ve D Haliyle sınıflama yargılarıyla ölçeklenmesi sonucunda dördüncü ve birinci hakem dışındaki hakemlerin ölçek sıralarında değişiklik olmuştur. Bütünsel dereceli puanlama anahtarı kullanılması durumunda kullanılan ölçekleme yöntemi ölçek değerlerinin sırasında farklılaşmaya neden olmazken, analitik dereceli puanlama anahtarı kullanılması durumunda B Hali tam veri matrisiyle elde edilen ölçek sıraları ile B Hali sayısal çözümü ve D Hali tam verili matristen elde edilen ölçek sıralarının farklılık gösterdiği görülmüştür. Araştırma sonucunda bir diğer önemli sonuç ise analitik dereceli puanlama anahtarı kullanılması durumunda B Hali sayısal çözüm ile D Hali tam veri matrisinden çözümünden elde edilen ölçek sıralarının farklılaşmadığı görülürken; B Hali tam verili matristen elde edilen ölçek sıralarının farklı olması, yani B Halinin kendi içinde tutarsız ölçek sıraları vermesidir.

Alanyazın incelendiğinde sınırlı sayıda ölçekleme çalışmasının olduğu görülmektedir (Anıl ve Güler, 2006; Nartgün, 2006; Kan, 2008; Öğretmen, 2008; Güler ve Anıl, 2009; Bal, 2011; Özer ve Acar, 2011; Ekinci ve diğerleri, 2012; Öztürk ve diğerleri, 2012). Bu çalışmalar incelendiğinde iki yöntemin ya da yaklaşımın karşılaştırılmasına yönelik çalışmalar oldukça azdır (Kan, 2008; Öztürk ve diğerleri, 2011). Ayrıca alanyazında sınıflama yargılarıyla ölçekleme çalışmasına rastlanmamıştır.

İleride yapılacak araştırmalarda standart belirleme yöntemlerinin sınıflama yargılarıyla ölçekleme çalışmasında ölçek değerlerini farklılaşıp farklılaşmadığı ve farklılaşma varsa bunun nedenleri incelenebilir. Farklı ölçekleme yöntemlerinin karşılaştırmasına yönelik çalışmalar yapılabilir. Ayrıca araştırmacılar ölçekleme çalışması yaparken yöntemlerin sayıltılarını da göz önüne almalıdırlar.

References

Alharby, E. R. (2006). A *comparison between two scoring methods, holistic vs analytic, using two measurement models, the generalizability theory and the many-facet rasch measurement, within the context of performance assessment.* Unpublished Doctoral Dissertation. Pennsylvenia State University, USA.

Andrade, H. G. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4). Retrieved December 1, 2011, from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1630

Anıl, D., & Güler. N (2006). İkili karşılaştırma yöntemi ile ölçekleme çalışmasına bir örnek. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30*, 30-36.

Atılgan, H., Kan, A., & Doğan, N. (2009). *Eğitimde ölçme ve değerlendirme* (4rd edition). Ankara: Anı Yayıncılık.

Bauer, B. A. (1981). *A study of the reliabilities and cost-efficiencies of three methods of assessment for writing ability*. (ERIC Document ReproductionService No. ED 216357).

Bal, Ö. (2011). Seviye belirleme sınavı (SBS) başarısında etkili olduğu düşünülen faktörlerin sıralama yargıları kanunuyla ölçeklenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2*(2), 200-209.

Boring, R. L. (2002). *Human and computerized essay assessment: a comparative analysis of holistic, analytic and latent semantic methods*. Unpublished Doctoral Thesis, Department of Psychology, New Mexico State University, Las Cruses, New Mexico.

Brualdi, A. (1998). Implementing performance assessment in the classroom. *Practical Assessment, Research & Evaluation, 6*(2). Retrieved December 11, 2011, from http://pareonline.net/getvn.asp?v=6&n=2

Ekinci, A., Bindak, R., & Yıldırım, M.C. (2012). Ilköğretim okulu yöneticilerinin öğretmenlerin mesleki sorunlarına empatik yaklaşımlarının ikili karşılaştırmalar metodu ile incelenmesi. *Gaziantep Üniversitesi Sosyal Bilimler Dergisi*, *11*(3), 759 -776.

Follman, J. C. & Anderson, J. A. (1967). An investigation of reliability of five procedures for grading english themes. *Research in the Teaching of English, 1,* 190-200.

Güler, N., & Anıl, D. (2009). Scaling through pair-wise comparison method in required characteristics of students applying for post graduate programs. *International Journal of Human Sciences [Online]. 6*(1), 627-639.

Haladyna, T. M. (1997). Writing test ıtems to evaluate higher order thinking (1st edition). USA: Allyn & Bacon.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, *7*(25). Retrieved January 16, 2012, from http://pareonline.net/getvn.asp?v=7&n=25

Moskal, B. M. (2000). Scoring rubrics: what, when and how?. *Practical Assessment, Research & Evaluation, 7*(3). Retrieved January 15, 2012, from http://pareonline.net/getvn.asp?v=7&n=3

Nartgün, Z, (2006). Öğretmenlik meslek bilgisi derslerinin önem düzeyinin ikili karşılaştırmalarla ölçeklenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 6*(2), 161-176.

National Council of Teachers of Mathematics. (2000). *Assessment standard for school mathematics*. Reston, Va. NCTM (Available online document). Retrieved October 1, 2011, from http://standards.nctm.org

Kan, A. (2008). Yargıcı kararlarına dayalı ölçekleme yöntemlerinin karşılaştırılması üzerine ampirik bir çalışma. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 35*, 186-194.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R.M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11,* 121-137.

Linn, R.L. (1993). Educational assessment: expanded expectations and challenges. *Educational Evaluation and Policy Analysis. 15,* 1-16. doi: 10.3102/01623737015001001

Öğretmen, T. (2008). Alan tercih envanteri: ölçeklenmesi, geçerlik ve güvenirliği. *Türk Eğitim Bilimleri Dergisi*, *6*(3), 507-522.

Özer, Y. ve Acar, M. (2011). Öğretmenlik mesleği genel yeterlikleri üzerine ikili karşılaştırma yöntemiyle bir ölçekleme çalışması. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi*, *3,* 89-101.

Öztürk, N., Özdemir, S. & Gelbal, S. (2011). İki farklı ölçekleme yaklaşımından elde edilen ölçek değerleri tutarlılığının incelenmesi. *20. Ulusal Eğitim Bilimleri Kurultayı*. Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi. 8-10 Eylül 2011. Burdur.

Resnick, L. B. (1987). *Education and learning to think*. Washington, D.C.: National Academy. Retrieved March 22, 2012, from http://www.nap.edu/openbook.php?record_id=1032&page=1

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Turgut, M. F., & Baykul, Y. (1992). *Ölçekleme teknikleri* (2nd edition). Ankara: ÖSYM Yayınları.

Torgerson, W. S. (1958). *Theory and methods of scaling* (1st edition) Newyork: John Wiley & Sons Inc.

Vrasidas, C. (2000). Constructivism versus objectivism: implications for interaction, course design, and evaluation in distance education. *International Journal of Educational Telecommunications, 6*(4), 339-362.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46* (7), 41–47.

Zollman, A., & Jones, D. L. (1994). Accommodating assessment and learning: utilizing portfolios in teacher education with preservice teachers. Paper presented at the annual meeting of the *Research Council on Diagnostic and Prescritive Mathematics*, Texas, IL.