The Effect of the Analytical Rubrics on the Objectivity in Physiotherapy Practical Examination

Analitik Dereceli Puanlama Anahtarlarının Fizyoterapi Eğitiminde Kullanılan Pratik Sınavlarda Nesnel Puanlanmaya Etkisi

Celal Deha DOĞAN,[a]
Hayri Baran YOSMAOĞLU[b]

[a]Department of Measurement and Evaluation,
Ankara University Faculty of Educational Science,
[b]Department of Physiotherapy and Rehabilitation,
Başkent University
Faculty of Health Science, Ankara

Geliş Tarihi/*Received:* 25.02.2014
Kabul Tarihi/*Accepted:* 01.09.2014

This study was presented at 4th National Physiotherapy and Rehabilitation Congress in Denizli, Turkey, 9-11 May 2013.

Yazışma Adresi/*Correspondence:*
Hayri Baran YOSMAOĞLU
Başkent University
Faculty of Health Science,
Department of Physiotherapy and Rehabilitation, Ankara,
TÜRKİYE/TURKEY
hayribaran@baskent.edu.tr

doi: 10.5336/sportsci.2014-39517

ABSTRACT Objective: The aim of the study was to find out if the scoring tools increase the objectivity of instructors during the assessment of practice performance in physiotherapy education. **Material and Methods:** Participant group of this study consist of 14 senior class students. Each student was asked to make a physiotherapy assessment and therapy planning for different musculoskeletal disorders in the practice exam. Half of the students were scored using scoring rubric and the rest of those were scored using graded category rating scales by the same instructors. Cohen's kappa coefficient and Spearmans Brown correlation coefficient have been calculated to find out the consistency between the ratings of the instructors. **Results:** Cohen Kappa analyzes show that the consistency between the raters was not statistically significant when raters used graded category rating scale. It was respectively stronger and statistically significant when they use scoring rubric (kappa= 0.47). Spearman Brown Correlation Coefficient results showed that there was positive, middle strong correlation (r: 0.51) between the scores when the raters use graded category rating scale. It was found positive, and strong correlation (r: 0.70) between the scores when the raters use scoring rubrics. **Conclusion:** The graded category rating scale may not be an effective solution for addressing the subjectivity of the raters. However, scoring rubrics can increase the objectivity of scorers during the assessment of clinical skills in physiotherapy education. If physiotherapy educators prepare appropriate rubrics while assessing the student performance, this may contribute to the more accurate determination of each students achievement.

Key Words: Scoring rubric; analytic rubric; graded category rating scale; inter-rater agreement; assessment objectivity

ÖZET Amaç: Bu çalışmanın amacı fizyoterapi eğitiminde pratik sınavlarda kullanılan farklı puanlama araçlarının, değerlendiricilerin nesnel puan vermelerinde ne düzeyde etkili olduğunu belirlemektir. **Gereç ve Yöntemler:** Çalışma grubu 14 üniversite son sınıf öğrencisinden oluşmaktadır. Pratik sınavda her öğrenciden kas iskelet sistemine ilişkin farklı sorunlar için fizyoterapi değerlendirmesi yapması ve bir tedavi program oluşturması istenmiştir. Öğrencilerin yarısının performansı dereceli puanlama anahtarı diğer yarısının performansı ise dereceleme ölçeği kullanılarak aynı öğretim üyeleri tarafından puanlanmıştır. Puanlayıcılar arasındaki uyumun belirlenmesi için Cohen in Kappa katsayısı ve Spearman Brown Sıra Sayıları Korelasyon Katsayısı hesaplanmıştır. **Bulgular:** Cohen Kappa analizi sonuçları, puanlayıcıların dereceleme ölçeği kullandıkları durumda puanlayıcılar arasındaki uyumun çok düşük ve istatistiksel olarak anlamlı olmadığını ortaya koymuştur. Puanlayıcılar dereceli puanlama anahtarı kullandıkları durumda ise, puanlaycılar arasındaki uyum daha yüksek ve istatistiksel olarak anlamlı bulunmuştur (kappa=0,47). Spearman Brown Sıra Sayıları Korelasyon Katsayısı sonuçları ise puanlayıcılar arasındaki uyum düzeyinin dereceleme ölçeği kullanıldığında pozitif ve orta güçlükte (r=0,51); dereceli puanlama anahtarı kullanıldığında ise pozitif ve yüksek güçlükte (r=0,70) olduğunu göstermiştir. **Sonuç:** Bu sonuçlara göre dereceleme ölçeğinin puanlayıcıların nesnelliğine önemli düzeyde katkı sağlamadığı öte yandan; dereceli puanlama anahtarının klinik becerilerin ölçülmesinde nesnelliği artırdığı belirtilebilir. Eğer eğitimciler pratik sınavların puanlanması sürecinde iyi hazırlanmış dereceli puanlama anahtarı kullanırlarsa, öğrenci başarısına ilişkin daha doğru ve güvenilir belirlemeler yapabilirler.

Anahtar Kelimeler: Dereceli puanlama anahtarı; analitik puanlama anahtarı; dereceleme ölçeği; puanlayıcı güvenirliği; değerlendirme nesnelliği

Turkiye Klinikleri J Sports Sci 2015;7(1):9-15

Practical exams are used to determine the success of students in the fields of medicine, physiotherapy, nursing and sport sciences. These exams basically aim to determine whether the student can make use of theknowledge in simulated real-life situations. The main problem faced in such exams can be defined as the risk of subjective assessment in determining the performance, or the possibility of the scorers grading the same level of performances differently. This, in turn, causes errors which interfere with the measurement of results and hence reduces the reliability of the measurement.[1]

In assessing students' performance during the practical exams, the two tools that can be used are the "graded category ratingscale" and the "scoring rubric". These are grading tools which include criteria relating to student performance and a particular number of success levels for each criterion.[2] Unlike graded category rating scales, scoring rubrics include performance definitions that help the raters to match the performance of each student with the appropriate success level. The possible advantages of these tools are stated clearly in the literature; however, no literature is available about their usage for practical performance evaluation in physiotherapy. Morover, although practical examinations is an extremely important for evaluation of the students' clinical performance, there is no study investigating the inter-rater realibility of the exams or the objectivity of the raters. This study aimed to find out if these assessment tools increase the objectivity of the scorer during the practical performance examination in physiotherapy education. The research questions addressed in this investigation are as follows:

1. When using scoring rubrics, to what extent were the ratings of the physiotherapy educators consistent in judging the performance levels of students in the practical examination?

2. When using graded category rating scales, to what extent were the ratings of the physiotherapy educators consistent in judging the performance levels of students in the practical examination?

MATERIAL AND METHODS

RESEARCH MODEL

This is a correlation research which aims to find out the relation between two or more variables without any intervention.

PARTICIPANT GROUP

The participant group for this study consists of 14 senior class students and two physiotherapy instructors, each with ten years experience and approximately the same proficiency in physiotherapy education. The study was conducted in 2012-2013 academic year at a foundation university stated in Turkey. Out of 14 participant subjects 8 of them are male and 7 of them are female. Because assessment of each student took approximately 45 minutes, participant subjects limited to 14 students. Students were randomly divided into two groups (Group A and Group B) and a practical examination was administered. Each student in both of the groups was asked to make a physiotherapy assessment and create a treatment programme for different musculoskeletal disorders. Two different assessment tools were prepared before the assessment: the scoring rubric and a graded category rating scale. Firstly, the performance of the students in Group A was scored by the two instructors using the graded category rating scale. Then, the performance of the students in Group B was scored by the same instructors using the scoring rubric. Each instructor scored the performance without knowing how the other instructor scored the students. The same instructors assessed the students in Group A and Group B in order to eliminate the effects caused by the experience and proficiency differences between the raters. The research process was demonstrated in Figure 1.

INSTRUMENTS

The practical examination which was administered during this study aimed to assess students' ability to make the correct physiotherapy assessment. The practical examamination lasted 45 minutes for each student. During the practical examination students were asked to make a physiotherapy assessment

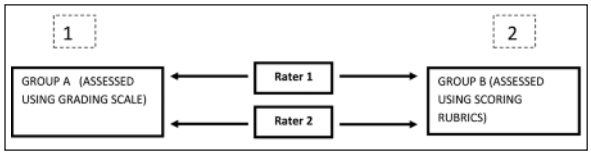

FIGURE 1: The schema that shows the research process.

and create a treatment programme for different musculoskeletal disorders. Students' performance was assessed considering six criteria. Below are details about the criteria;

Knowledge about the disease: The raters assess if the student knows symptoms and prognosis of the disease.

Pain assessment: The raters assess to what degree the student does pain assessment.

Strenght assessment: The raters assess to what gegree the student does strength assessment.

Flexibility assessment: The raters assess to what gegree the student does flexibility assessment.

ROM assessment: The raters assesss to what gegree the student does ROM assessment.

Communication with the patient: The raters assess if the student use appropriate language and does clear explanations.

Because there are 6 criteria and 5 performance level the maximum score that the students can get from the practical examination is 30. The students who get score below 18 considered as unsuccessful. On the other hand the scores between 19 and 25 considered as average and 26-30 considered as very successful.

In order to asssess students' performance, a scoring rubric and a graded category rating scale including the criteria given above was developed by the researchers as follows.

Scoring Rubric: The scoring rubric includes six criteria for assessing students' practical performance. The criteria included in the rubric are knowledge about the disease, painassessment, strength assessment, range of motion (ROM) assessment, flexibility assessment and the assessment of communication with the patient. The rubric has 5 performance levels rated from *poor* to *excellent*. Each criterion includes a performance description for each performance level, which helps the scorers to assess the students objectively. The scoring rubric that was used in this research was given in Table 1.

Graded Category Rating Scale: The graded category rating scale has six criteria for assessing students' practical performance. The criteria in the grading scale are knowledge about the disease, pain assessment, strength assessment, range of motion (ROM) assessment, flexibility assessment and the assessment of communication with the patient. The scale has 5 performance levels rated from *poor* to *excellent*; however, it has no performance descriptions for the criteria. The graded category rating scale that was used in this research was given in Table 2.

DATA ANALYSIS

SPSS 13.0 was used to analyse the data. Two different statistical analyses were performed to assess consistency. Firstly, Cohen's kappa coefficient has been calculated to find out the consistency between the scorers. Cohen's kappa coefficient is a statistical measure of inter-rater agreement or *inter-annotator agreement* for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. The two residents randomly assign their ratings; however, they would sometimes agree just by chance. Kappa gives us a numerical rating of the degree to which this occurs. The calculation is based on the difference between how much agreement is actually present ("observed" agreement) compared to how much agreement would be expected to be present by chance alone.[3] SPSS 13.0 was used to analyse the data.

Secondly, Spearmans Brown correlation coefficient has been calculated to find out the relation between the ratings of the instructors. Spearman's rank correlation coefficient is used as a measure of linear relationship between two sets of ranked data, i.e. it measures how tightly the ranked data clusters around a straight line. Spearman's rank correlation coefficient, like all other correlation coefficient, will take a value between -1 and +1. A

TABLE 1: The scoring rubric.

Criteria	Performance Levels				
	1 (Poor)	2 (Insufficent)	3 (Sufficent)	4 (Good)	5 (Excellent)
Knowledge About The Disease	The student does not have knowledge about the disease	The student defines the disease but does not know the symptoms and prognosis of the disease	The student knows the symptoms and prognosis of the disease but has some important insufficiency or mistakes	The student knows the symptoms and prognosis of the disease but has little mistakes or defects	The student knows symptoms and prognosis of the disease perfectly
Pain Assessment	The student does not know that he/she has to do pain assessment.	The student knows that he/she has to do pain assessment but does not know how to do	The student knows that he/she has to do pain assessment and knows how to do but he/she does some important mistakes while doing it.	The student knows that he/she has to do pain assessment and knows how to do but he/she does little mistakes while doing it.	The student knows that he/she has to do pain assessment and does it perfectly
Strength Assessment	The student does not know that he/she has to do strength assessment	The student knows that he/she has to do strength assessment but does not know how to do	The student knows that he/she has to do strength assessment and knows how to do but he/she does some important mistakes while doing it.	The student knows that he/she has to do strength assessment and knows how to do but he/she does little mistakes while doing it.	The student knows that he/she has to do strength assessment and does it perfectly
Flexibility Assessment	The student does not know that he/she has to do flexibility assessment	The student knows that he/she has to do flexibility assessment but does not know how to do	The student knows that he/she has to do flexibility assessment and knows how to do but he/she does some important mistakes while doing it.	The student knows that he/she has to do flexibility assessment and knows how to do but he/she does little mistakes while doing it.	The student knows that he/she has to do flexibility assessment and does it perfectly
ROM Assessment	The student does not know that he/she has to do ROM assessment	The student knows that he/she has to do ROM assessment but does not know how to do	The student knows that he/she has to do ROM assessment and knows how to do but he/she does some important mistakes while doing it.	The student knows that he/she has to do ROM assessment and knows how to do but he/she does little mistakes while doing it	The student knows that he/she has to do ROM assessment and does it perfectly
Comminication with the patient	The self confidence of the student is low and does not communicate with the patient	The self confidence of the student is low and he/she communicates with patient insufficiently	The self confidence of the student is high but he/she uses a inappropriate language to the patient and doesn't make clear explanations.	The self confidence of the student is high and he/she uses a appropriate language but doesn't make clear explanations.	The self confidence of the student is high, he/she uses a appropriate language and make clear explanations.

TABLE 2: The graded rating scale

Criteria	Performance Levels				
	1 (Poor)	2 (Insufficent)	3 (Sufficent)	4 (Good)	5 (Excellent)
Knowledge About The Disease					
Pain Assessment					
Strength Assessment					
Flexibility Assessment					
ROM Assessment					
Communications with the patient					

positive correlation is one in which the ranks of both variables increase together. A negative correlation is one in which the ranks of one variable increase as the ranks of the other variable decrease. A correlation of +1 or -1 will arise if the relationship between the two variables is exactly linear. A correlation close to zero means there is no linear relationship between the ranks. In this study Spearman Brown Correlation Coefficient was used to rainforce the statistical accuracy of the analyses.

TABLE 3: Consistency between the raters when they use graded - raiting scale or and scoring rubric.

	Kappa Measure of Agreement Value	Asymp. Std. Error(a)	Approx. T(b)	Appox. Sig
Graded-raiting scale	0.05	0.081	0.629	0.529
Scoring rubric	**0.47**	0.089	5.812	**0.000***

Key *: p<0.05 for Cohen Kappa Test The first column (Kappa Measure of Agreement Value) shows the strength of the consistency (agreement) between the raters. Kappa is a measure lies on a -1 to 1 scale, where 1 is perfect agreement 0 is exactly what would be expected by chance and negative values indicate agreement less than chance. The last column in Table 3 that is another important value should be interpreted (Approx. Sig) tests if estimated kappa is not due to chance. It does not test the strength of the agreement. If it is significant it means that the consistencey between the raters is not due to chance.

RESULTS

Cohen's kappa analysis showed that the consistency between the raters was very low when they used graded category rating scale and was not statistically significant (kappa=0.05, p>0.01). On the other hand, the consistency between raters was respectively stronger and statistically significant when they used scoring rubric (kappa=0.47 p<0.01; see Table 3). So when scoring rubric is used agreement between the raters is higher than the one when graded category rating scale is used. In other words results of Kappa statistics show that the raters do more objective assessment when they use scoring rubric rather than graded category rating scale. Figure 2 shows the comparison of the Cohen kappa results when instructors used the graded category rating scale and the scoring rubric.

Moreover Spearman-Brown correlation coefficient has been calculated to strengthen the kappa results. Spearman-Brown correlation coefficient results showed that there was a positive, middle–strong correlation (r=0.51, p<0.001) between the scores when the raters used the graded category rating scale. A positive and strong correlation (r=0.70, p<0.01) was found between the scores when the raters used the scoring rubric. The consistency between the raters was higher when they used the scoring rubric than when they used the graded category rating scale. Figure 3 shows the comparison of Spearman-Brown correlation coefficient results when instructors used the graded category rating scale and scoring rubric.

DISCUSSION

The most important finding of the study was that graded category rating scale did not counteract the

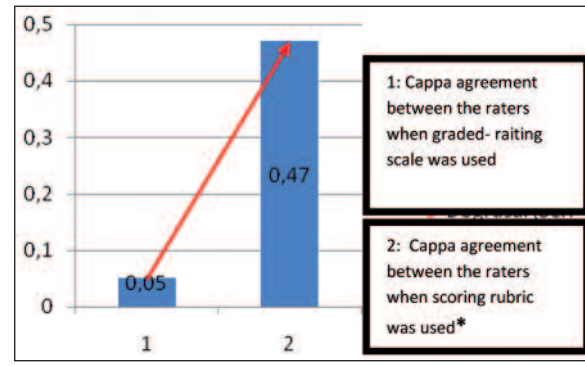

FIGURE 2: Cohen Kappa Agreement between the raters when graded-raiting scale and scoring rubrics were used.

Key *: p<0.05 for Cohen Kappa Test.

The figure 2 shows that the agreement between the raters increases when they use scoring rubric rather than graded category rating scale. The first bar shows the Kappa agreement between the raters when they use the graded category rating scale. The second bar shows the Kappa agreement between the raters when the use the scoring rubric.

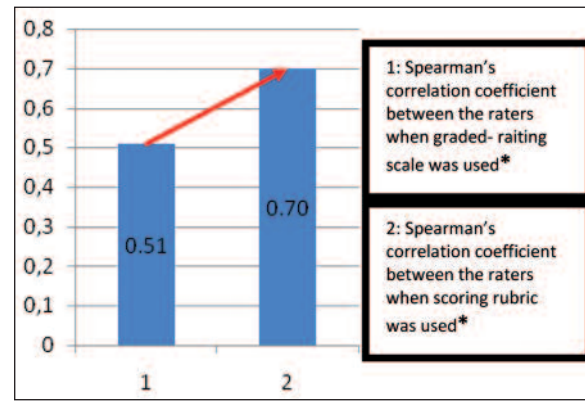

FIGURE 3: Spearmans's Correlation coefficient between the raters when graded- raiting scale and scoring rubric were used.

Key: * p<0.05 for Spearman Brown Correlation Test

The first bar shows the Spearman's correlation coefficient between the raters when they use graded category raiting scale. The Second bar also shows the same correlation coefficient between the raters when they use scoring rubrics. As it is seen in figure 3, the agreement between the raters increases when they use scoring rubric. Therefore, the results of Spearman's correlation coefficient prove the results of Kappa statistics in general.

subjectivity of the raters sufficiently; however, when the scoring rubric was used, the congruity between observers was much higher. This finding suggests that when scoring rubric is used in evaluating proficiency in the practical examination, the probability of different scorers assigning the same score to the same level of performances is higher. This is important for minimising the problem of the raters' making subjective judgments during the assessment. In literature, the advantages of the scoring rubric are well explained. The most important advantage of scoring rubrics is that they make the students' performance evaluation more objective. Since the scoring rubrics include separate criteria for student performance and a particular number of success levels for each criterion, they make it easy for the graders to designate grades for the separate performances of each student.[2] The components essential to the rubric include concise performance criteria, graded category rating scales and a description of the expected performance at each level.[4-8] As a result of this, different graders can make assessments within the same span of points for the same criteria. Another contribution the scoring rubric demonstrated was that it enabled students to get effective feedback. Students can see what point to they need to reach according to each criterion. Hence, students know what is expected from them. This, in turn, provides a better study atmosphere for students while it gives the instructor the opportunity for more objective grading.[2,9,10] According to Rucker and Thompson, feedback is most effective when given quickly after task completion, and students are more concerned about getting information back quickly than about getting helpful feedback.[11] Rubrics foster rapid review and facilitate communication because feedback can be provided quickly, fairly, efficiently and individually. With these characteristics, the use of rubrics provides, on the one hand, an effective solution to the subjectivity/objectivity clinical grading dilemma; on the other hand, it is also an effective tool for providing students with accurate feedback.[12]

Although the advantages of scoring rubrics are obvious, studies investigating the use of rubrics in terms of assessing students' clinical practice skills are limited. The study conducted by Lunney and Sammarco, which focuses on the inter-rater reliability of a three-part instrument used to evaluate nursing students who were in an online baccalaureate nursing programme, showed that the correlations of the raters' scores were positive and statistically significant.[13] Isaacson and Stacy stated that using a scoring rubric for clinical evaluation decreases the discrepancies in how the learning objectives are interpreted by instructor and students.[12] Moreover, since rubrics provide a clear scoring guide by identifying achievement levels of criteria, they may be beneficial in preventing educators from making biased judgements.[12,14] In our study, it was found that using rubrics for assessing the clinical skills in physiotherapy education increased the objectivity of scoring.

A rubrics thought to be a fair, equitable and consistent scoring guide for measuring student achievement. However, rubrics that are not developed properly have some negative features.[8] Goodrich, and Kutlu et al. have highlighted the aspects of rubrics that are widely investigated in literature.[9,15] First, there should be concordance between the behaviours to be measured and criteria. If scoring rubrics have some criteria that it is not intended should be measured, the validity of the tool will be compromised. Similarly, if a criterion related to a feature to be measured is not presented in the scoring rubric, then the validity of the grading scale would again be compromised. For this reason, in developing scoring rubrics, the consistency between criteria and the characteristics that it is intended to measure should, first of all, be established. Second, subjective statements should not be used. This is because subjective statements such as "a little", or "partially" may compromise the integrity of the scoring. Hence, subjective statements should be avoided as much as possible and concrete statements should be used instead. Another important point is that the criteria in the rubrics should not aim to assess more than one feature. Criteria measuring similar characteristics or more than one characteristic will harm the validity of the assessment. Thus, the criteria in the scoring rubrics should be independent and should not coincide or overlap with each other.

LIMITATIONS

Since the evaluation of each student lasts for 45 minutes, limited numbers of students were included in the study. A similar study could be carried out with a larger group of students. In this study, the rubric was planned for use in the assessment of clinical skills in musculoskeletal diseases. Similar studies investigating the effect of rubrics on the objectivity of scoring could be repeated in other subcategories of physiotherapy.

CONCLUSION

The graded category rating scale may not be an effective solution for addressing the subjectivity of the raters. However, scoring rubrics can increase the objectivity of scorers during the assessment of clinical skills in physiotherapy education. If educators prepare appropriate rubrics while assessing the student performance, this may contribute to the more accurate determination of each student's achievement. It is strongly advised to the researchers to test the interrater agreement using analysis based on generalizability theory for the future studies. The level of agreement among three or more raters may give more reliable results. Similar studies should also be done with more than two raters.

Acknowledgement

Authors thank to Aydan Aytar PT PhD and Özlem Yürük PT PhD for contribution of this study.

REFERENCES

1. Roid GH, Haladyna TM. Chapter 5. Technology for Test-Item Writing. 1st ed. New York: Academic Press; 1982.p.68-84.
2. Haladyna TM. Chapter 5. Writing Test Items to Evaluate Higher Order Thinking. 2nd ed. Boston: Viacom Campany; 1997 p.123-50.
3. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-3.
4. McDonald M, Sudbury MA. The Nurse Educator's Guide to Assessing Learning Outcomes. 2nd ed. Burlington, MA: Jones and Bartlett Publishers; 2007.p.1-387.
5. Spence L. Discerning writing assessment: insights into an analytical rubric. Language Arts 2010;87(5):337-50.
6. Suskie L. Assessing Student Learning: A Common Sense Guide. 2nd ed. San Francisco, CA: Jossey-Bass; 2009.p.1-384.
7. Truemper CM. Using scoring rubrics to facilitate assessment and evaluation of graduate-level nursing students. J Nurs Educ 2004; 43(12):562-4.
8. Shipman D, Roa M, Hooten J, Wang ZJ. Using the analytic rubric as an evaluation tool in nursing education: the positive and the negative. Nurse Educ Today 2012;32(3):246-9.
9. Goodrich AH. Teaching with rubrics: the good, the bad, and the ugly. College Teaching 2005; 53(1):27-31.
10. Moskal BM. [Scoring rubric: what, when and how?] Pratical Assessment, Research & Evaluation 2000;7(3).
11. Rucker ML, Thomson S. Assessing student learning outcomes: An investigation of the relationship among feedback measures. CollegeStudent Journal 2003;37(3):400-4.
12. Isaacson JJ, Stacy AS. Rubrics for clinical evaluation: objectifying the subjective experience. Nurse Educ Pract 2009;9(2):134-40.
13. Lunney M, Sammarco A. Scoring rubric for grading students' participation in online discussions. Comput Inform Nurs 2009;27(1):26-31;quiz 32-3.
14. Gantt LT. Using the Clark Simulation Evaluation Rubric with associate degree and baccalaureate nursing students. Nurs Educ Perspect 2010;31(2):101-5.
15. Kutlu O, Dogan D, Karakaya I. Öğrenci Başarısının Belirlenmesi: Performansa ve Portfolyoya Dayalı Durum Belirleme. 3. Baskı. Ankara: Pegem Akademi Yayıncılık; 2010. p.51-88.