The Impact of Rubrics on the Evaluation of Students' Success on the "Subject of Force and Motion"

Sümeyye AYDIN GÜRLER[1,] Oktay BAYKARA[2].

1. Elifoğlu Secondary School Science Teacher, Karkamış, Gaziantep, Turkey.
Email address: pinar_aydin_23@hotmail.com / Telephone: +905334358398
2. Department of Elemantary Education, Faculty of Education, University of Fırat, Elazığ, Turkey
Email address: obaykara@gmail.com / Telephone: +905064930799

Corresponding Author:
Sümeyye AYDIN GÜRLER, Elifoğlu Secondary School, Karkamış, Gaziantep, Turkey.
Email address: pinar_aydin_23@hotmail.com / Telephone: +905334358398

ABSTRACT: This study's aim is to investigate whether the use of a scoring rubric provides success for students and objectivity in student evaluation. Peer assessment that allows for peer evaluation, which was among the goals of this study, was also carried out. This study included 25 students in the 6th grade of a secondary school. Both during and after the teaching of the subject Force and Motion, 20 multiple-choice questions were given to the students. Students were then given an exam on the subject. When the scores of the pre-test and post-test on the subject were analyzed, the average post-test scores were higher than the average pre-test scores. Additionally, there was a positive relationship (r = 1) between the pre-test and post-test scores given by three peer raters and the researcher. The study shows that a scoring rubric ensures objectivity in evaluation.
Keywords: Scoring Rubric, Pre-test, Post-test, Objectivity.

1. INTRODUCTION

Assessment has a vital place in the teaching and learning process and it should be effectively carried out by the instructor as a gauge for monitoring progress of that process. A variety of tools has been used for assessment in today's educational practices. Classical measurement tools (short answer, long answer, multiple-choice exams, etc.), which are used to measure a student's knowledge in a limited period of time, are not enough to assess the chance of success and shortcomings of the students. They do not provide enough information for the teacher about their students (Shepard, 1989; Mumme, 1991; Romberg, 1993; Birgin, 2008). Today, due to the rapid growth of information, students are required not only to keep information in mind, but use the information they have learned in new and different situations. For example, for many vocational groups, it is desirable to be able to solve problems at work, think critically, analyze data, have oral and written communication skills, create new things, efficiently use information, and evaluate themselves (Dochy, 2001). Therefore, classical evaluation approaches which have been generally applied are inadequate to measure the properties mentioned above.

Different (or alternative) assessment tools are being developed by educators with the aforementioned objectives of these educational activities in mind. Alternative assessments expect the students to do or produce something and to have advanced thinking and problem-solving skills.

Thus, alternative assessment provides students the opportunity to evaluate each other as peers and gives the students an opportunity for both personal and group work (Kocakülah, 2009). In this new concept, in addition to measuring the success of students, keeping students and parents informed and monitoring the development of each student's individual characteristics in the learning-teaching process and the active participation of students in the evaluation process have become more prominent. Thus, the use of alternative assessment methods such as portfolios, performance, self- and peer assessment, scoring rubrics, projects, and concept maps, has become a necessity in determining a student's academic success. According to the constructivist approach based on the student-centered evaluation, scoring rubrics are a contemporary alternative assessment tool for evaluating performance and are a popular pedagogical tool. In recent years, studies of development and the use of rubrics have been increased in the literature on science teaching and learning (Luft 1999; Popham 1997; Kocakülah, 2009). The rubric provides the opportunity for the students to evaluate and interpret learning processes and evaluate each other. Used more formatively, rubrics can also help instructors get a clearer picture of the strengths and weaknesses of their class (URL-1). With the use of rubrics, students are aware of what they are doing. They can express themselves based on the teaching objectives and improves problem solving and thinking skills. Thereby, they may not have the ability to produce fresh thinking without living and applying processes suitable for learning and teaching. There are two different scoring rubrics, analytical and holistic (global), depending on the number of scales used. While an analytical scoring rubric generally provides valuable information regarding instruction and student skills, holistic scoring is often more efficient (Boston, 2002).

The evaluator should also approach the results objectively in order to do a fair and consistent evaluation. Evaluation is made not only by the instructor, but students themselves and also alternative assessments by peers. If students are aware of what and how they learn, they will learn better. Self-assessment is a kind of reflection; a student detects the development or shortcomings in the learning process by reflecting, regulates future behavior according to these data, gains courage in becoming an independent learner, and can increase motivation (Cihanoğlu, 2008). The two important events in the contemporary assessment process are self- and peer assessment (Davies, 308). Self-assessment can be described as the participation of learners in making judgments about their own learning (Boud and Falchikov, 1989). Peer assessment can be defined as the process through which groups of individuals rate their peers (Falchikov, 1995). Peer assessment has been shown to be generally more reliable than self-assessment, while there has been shown to be a high positive correlation between self- and peer assessment in literature related to peer and self-assessment (Topping, 1998; McDowell 1995). The basic problems inherent in assessment are the issues of reliability and validity. Useful information is needed to assess student performance. The rubric, thus, must provide this to the teacher in order to be clear and be based on standards and criteria (Stuhlmann, 1999).

If a student obtains the same score on any assessment regardless of when the student completed the assessment, when the response was scored, and who scored the response, it can be said that the reliability of the test is high and consistent (Boston, 2002). Otherwise, the obtained results relating student's score may be very far from the purpose of the assessment. Generally, in classroom assessment and rubric development which involve raters (or scorers), rater reliability is considered for reliability of assessments. Rater reliability that has *interrater* and *intrarater* reliability is especially important in designing a rubric. These refer to the consistency of scores given by two independent raters and consistency of scores given by the same rater at different points in time, respectively (Boston, 2002). For well-designed scoring rubrics, both interrater and intrarater reliability needs to be developed.

According to Jonsson and Svingby (2007), validation is the difficult process of searching for an answer to the question, "Does the assessment measure what it was intended to measure?" The validity of an assessment instrument must be considered in terms of content, construct (Boston, 2002; Jonsson and Svingby 2007).

1.1. Purpose of the Study

The main purpose of the present work is to develop and use a scoring rubric in assessing the understanding and acquisition of knowledge of secondary school students in 6th grade science classes studying the subject of force and motion. Since peer assessment involving peer evaluation is also among the goals of this study, three students evaluated three random peers' question papers based on a scoring rubric.

The assessment of students' work using a scoring rubric we developed was used in this study. For this, the steps were as follows:

- preparation of a teaching plan that allows the use of a scoring rubric,
- indication of the benefits of a scoring rubric in solving questions and doing sample solutions during teaching,
- development of a scoring rubric that provides a means of assessing the questions by taking the opinions of three experts and all the students at the end of each course,
- evaluation of student performance by the researcher and three peer raters in accordance with the scoring rubric and evaluation of the consistency of both assessments.
- provision of a resource for instructors on the topics of rubric development, application, and feedback.

2. METHODS

2.1. Type of Research

The answers to the force and motion questions were compared first without then with the rubric, the obtained results were examined, and the scores to these questions given by three peer raters and the researcher were compared. Analytical and holistic rubrics were developed by the researcher for this purpose. Although similar studies were conducted in different classes, this study is basic research, as the results obtained from this study will be used in future studies on this subject.

2.2. Working Group

The working group of the study consists of 6th grade students of Elifoğlu Secondary School in Elifoğlu village of the Karkamış district of Gaziantep in Turkey. Twenty-five 6th grade students took part in the research. Each student is evaluated individual using the analytical and holistic rubrics, and pre-tests and post-tests were used in the study. The evaluations were conducted by both the researcher and three peer raters who were randomly selected.

2.3. Data Collection Tools

2.3.1. Test Questions

When preparing the test questions, an initial survey was given to 105 7th grade students who had taken the subject of force and motion the previous year to find out which aspects of force and motion they found difficult. According to the results of the survey, 11 aspects were determined. Thirty-three multiple-choice questions, three for each aspect, were prepared. These questions were then applied to the same group of 105 7th graders to evaluate the reliability and validity of each question. A "matter points matrix" was prepared based on the answers to these questions and an "item difficulty index" and "item discrimination power index" for each question were calculated.

The total number of questions was decreased to 20 by eliminating 13 that had low difficulty indexes and discrimination power. An expert was consulted in making these eliminations. This final version of the test had a coefficient alpha (or KR 20) of 0.92 and an average item difficulty index of 0.27. The final test consisted of 6 questions on "velocity in our lives," 4 questions on the subject of "let's discover force," 6 questions on "forces at work," 4 questions on the subject of "weight is a force" and was given to students as pre-test and post-test before and after forming the scoring rubric.

2.3.2. Analytic and Holistic Scoring Rubrics

Based on the questions' features, analytic rubrics were created for the subjects "velocity in our lives" and "forces at work," while holistic rubrics were created for "let's discover force" and "weight is a force." Rubrics were prepared following the steps proposed by Goodrich (2001) and were only able be used for this course. The steps are briefly summarized below.

1. Listing the criteria to determine performance. The objectives of the course and criteria and students' desired skills were determined. Each criterion focused on a substantial part of the performance.

2. Deciding which type of rubric to use as a scoring strategy. While deciding the type of rubric to be used, the objective to be measured in the questions for each of the topics was taken into consideration. A chapter composed of four topics, two topics using analytical rubrics and two using holistic rubrics, was used.

3. Determining levels of performance and defining levels. At this stage, levels of performance were determined and scored from the most efficient to the least. Because in this application the performance levels for all criteria are expressed in numbers, it is believed that this will greatly reduce the raters' bias.

4. Receiving expertise. Each rubric was prepared by a researcher in consultation with three experts. The rubric was then finalized based on the opinions of the students who would participate in the practice.

2.4. Collection of Data

Three different sets of data were collected in the research. The first data set was the pre-test scores on the subject of force and motion that were assessed by the researcher without giving the rubric to the students. The second data set consists of the post-test scores and the scoring rubric that was given to the students. The data obtained from this stage were again assessed by the researcher. Before the rubrics were put in to practice, information about the rubric was given to the students by the researcher, sample rubrics prepared by the researcher were given to each student, and the objectives of the scoring tool (rubric) and how the scoring would be done were explained.

The third and final data set consists of the pre-test and post-test scores of three randomly selected students' exam papers obtained from both the researcher and the randomly selected peer raters.

2.5. Data Analysis

Data were analyzed according to the following sub-problems:

1) Is there a significant difference in students' problem-solving skills on the subject of force and motion before and after using the rubric?

2) With the students' use of the scoring rubric, is there a significant difference in students' force and motion question scores compared to the raters' scores (both the researcher and randomly selected peers)?

3) Do the three peer raters' and researcher's scores on the force and motion questions differ from pre-test to post-test?

4) Is there a meaningful relationship between the scores given to pre-test and post-test questions by the three peer raters and researcher?

3. RESULTS

The results of the research on the aforementioned sub-problems are presented below.

3.1. Is there a significant difference in students' problem-solving skills on the subject of force and motion before and after using the rubric?

Comments on statistical findings related to differences of the students' problem-solving skills on the subject of force and motion before and after using rubric are given below. Table 1 shows t-test statistical results of pre-test and post-test scores of the students based on the use of rubrics as assessed by the instructor.

Table 1

There was a significant increase in students' success in solving questions with the use of the rubric ($p < 0.05$). While the average of the scores given by the researcher to questions solved by the students without using a scoring rubric is 719.2, the average of the scores of the students using a scoring rubric is 1478. The average of the post-test total score is twice the average of pre-test total score. This result shows that the use of the scoring rubric has had a significant impact on the students' problem-solving success.

3.2. With the students' use of the scoring rubric, is there a significant difference in students' force and motion question scores compared to the raters' scores (both the researcher and randomly selected peers)?

Students' pre-test and post-test scores, separated by rater, with the use of a scoring rubric are shown in Table 2.

Table 2

In the pre-test and post-test, students' success in solving problems is not significantly different according to the raters. The total average scores in pre-test given by the three peer raters and researcher are 146.80 and 142.75, respectively. In other words, the total average scores of both researcher and peer raters are very close to each other ($t = 1.1$; $p = 0.386 > 0.05$). Likewise, a similar post-test relationship had been shown. The total average scores in the post-test given by three peer raters (249.74) and the researcher (250.00) are almost the same ($t = 2$; $p = 0.862 > 0.05$).

3.3. Do the three peer raters' and researcher's scores on the force and motion questions differ from pre-test to post-test?

The variation of students' test scores, separated by rater, is shown in Table 3.

Table 3

As shown in Table 3, there is no significant difference with respect to raters among the pre-test and post-test scores of the three randomly selected students ($p>0.05$).

3.4. Is there a meaningful relationship between the scores given to pre-test and post-test questions by the three peer raters and researcher?

In this section of the study, the aim was for the researcher and three peers to analyze the exam papers of the students in accordance with the developed scoring rubric and look at the consistency across raters. For this, correlation between peer and researcher assessments of the total pre-test and post-test scores obtained by the students was analyzed. The results of the analysis have been shown in Tables 4 and 5.

Table 4

As can be seen in Table 4, there is a significant and highly positive relationship between the researcher's and three peers' assessment using the rubric (r = 1). Accordingly, the rubric is said to be a reliable assessment tool since the scores given by students are close to those given by the researcher.

Table 5

As it is seen in Table 5, there is a positive and significant correlation at a high level between the assessments of the researcher and three peers (r = 1). The close values of the correlation coefficient show that scores obtained by the students are different from the those obtained by the people making the assessment.

4. DISCUSSION AND CONCLUSION

In the literature there appears to be a lack of studies on the improvement and use of a scoring rubric to evaluate student performance and make peer assessment of problem solving in a selected science topic. This research aims to study the effects of the use of a rubric designed to measure student achievement. The relationship between the scores given by peers and instructors in evaluating student's work has been also studied.

When students' pre-test and post-test scores on the subject of force and motion were analyzed, the average of post-test scores was higher than the average of pre-test scores. This could mean that students have prior knowledge about the evaluation criteria and know what is expected of them, which has a positive effect on their performance. In addition, there seems to be a highly positive relationship between the pre-test and post-tests scores given by three peer raters and the researcher. These similar results show that whoever evaluates the students' worksheet using a scoring rubric will obtain similar results. In short, this shows that scoring rubrics ensure objectivity in evaluation.

The development of scoring rubrics and their use in problem solving is very important and useful for student assessment. For example, in this study, while many students did not write the velocity formula in the topic "the velocity in our lives" in the pre-test, many of them wrote the physical quantities correctly in the post-test in which scoring rubric was applied. Similar results have been obtained in other subjects. The results of this study provide in-depth information about the use of scoring rubrics as tools to increase student achievement in science courses.

Peer assessment gives students the opportunity for feedback on their performance and provides peer tutoring and also internal evaluation by using external evaluation (Bostock, 2001). In this study, students are given the opportunity to be a more realistic and to provide peer tutoring and detailed feedback on their performance. However, studies where students are personally involved in

the assessment are scarce. There are benefits to student participation in the evaluation process in terms of meaningful learning, gains in self-confidence, and improvement of social interaction and communication skills, both in society and the classroom.

By the end of the study, the researcher asked the students questions such as "what was your feeling about knowing how the evaluation would be done while taking the exam" and the following results emerged from the students' answers. Informing students that they will evaluate their own exam paper according to a scoring rubric is useful for students and there are several benefits, such as increased consistency of scoring and self-confidence. This can be clearly seen when comparing the pre- and post-test scores. Furthermore, hesitations about the fairness of an evaluation have been eliminated by using scoring rubrics and this has increased the students' confidence in teacher evaluation.

In the answers given by students to the question "what are the advantages and disadvantages for you of a problem given with a scoring rubric," most of the students said that a scoring rubric is effective in helping them understand not only the result but the subject itself and also that evaluating the process created a positive impact on them and that a scoring rubric is a guide in solving problems.

Based on these findings and conclusions, teachers especially should be given more details about the structure and scoring of rubrics, and the objective of the study should be better explained so that improvements could be made in evaluating exam papers and interpreting results. Because the preparation of a scoring rubric is difficult and takes time, instructors generally want to use traditional assessment tools. However, the outcomes of assessment can make it difficult to get reliable results and to make comments. It is hope that there will be an increase in the number of teachers who use a rubric process in operating their courses and that teachers will include students to process of rubric assessment. In addition, scoring rubrics can be used not only in the evaluation of the exam papers but in assessing student performance during teaching, as was done in this study. In this way, this study aims to contribute to the science education literature. Scoring rubrics can also provide alternative assessment not only for teachers but for peers as they participate in the assessment process.

Acknowledgement

This work was supported by the TUBITAK (BİDEB-2211 Direct Domestic Ph.D. Scholarship Program)

5. REFERENCES

Birgin, O. (2008). Alternatif bir değerlendirme yöntemi olarak portfolyo değerlendirme uygulamasına ilişkin öğrenci görüşleri. *Türk Eğitim Bilimleri Dergisi*, 6 (1), (pp.1-24).

Bostock, S. (2001). Student peer assessment. Retrieved February 20, 2015 from http://78.158.56.101/archive/palatine/files/994.pdf

Boston, C.(2002). *Understanding scoring rubrics: a guide for teachers*, ed. United States of America,

Boud, D., Falchikov, N. (1989). Quantitative Studies of Self-Assessment in Higher Education: a Critical Analysis of Findings. *Higher Education*, 18(5), (pp. 529-549).

Cihanoğlu, M.(2008). *Alternatif Değerlendirme Yaklaşımlarından Öz ve Akran Değerlendirmenin İşbirlikli Öğrenme Ortamlarında Akademik Başarı, Tutum ve Kalıcılığa Etkileri,* Doktora Tezi. Dokuz Eylül Üniversitesi, Eğitim Bilimleri Enstitüsü, İzmir.

Davies, P. (2002). Using Student Reflective Self-Assessment for Awarding Degree Classifications, *Innovations in Education and Teaching İnternational,* 39(4), (pp.307-319). DOI: 10.1080/13558000210161034

Dochy, F. (2001). A New assessment era: different needs, new challenges, *Research Dialogue in Learning and İnstruction,*10 (1), (pp.11-20).

Falchikov, N. (1995). Peer feedback marking: developing peer assessment, *Innovations in Education and Training International,* 32(2), (pp.175–187).

Goodrich, A. H. (2000). Using Rubrics to Promote Thinking and Learning, *Educational Leadership,* 57 (5), (pp.13-18).

Jonsson, A., Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences, *Educational Research Review,* 2, (pp.130-144).

Kocakülah, M. S. (2009). Development and Application of a Rubric for Evaluating Students' Performance on Newton's Laws of Motion, *Science Educational Technology,* 19, (pp.146-164). DOI: 10.1007/s10956-009-9188-9

Luft, J.A. (1999). Rubrics: design and use in science teacher education. *Journal of Science Teacher Education* 10(2), (pp.107–121).

McDowell, L. (1995). The impact of innovative assessment on student learning, *Innovations in Education and Training International*, 32, (pp.302–13).

Mumme, J. (1991). *Portfolio assessment in mathematics*. Santa Barbara: California Mathematics Project, University of California.

Popham, W.J. (1997). What's wrong- and what's right-with rubrics. *Educational Leadership,* 55(2), (pp.72–75).

Romberg, T. A. (1993). How one comes to know models and theories of the learning of mathematics. In M. Niss (Ed). *Investigations Into Assessment In Mathematics Education,* (pp. 97-111). Netherlands: Kluver Academic Publishers.

Shepard, L. A. (1989). Why we need better assessment?. *Educational Leadership*, 46 (7), (pp.4-9).

Stuhlmann, J., Danieli C., Dellinger, A., Denny, R. K. and Powers, T. (1999). ''A generalizability study of the effects of training on teachers' abilities to rate children's writing using rubrics," *Journal of Reading Psychology*, 20 (2), (pp. 27-107). DOI: 10.1080/027027199278439

Topping, K. J. (1998) Peer assessment between students in colleges and universities, *Review of Educational Research*, 68, (pp. 249–76).

URL-1, Grading and Performance Rubrics. Retrieved February 9, 2014 from http://www.cmu.edu/teaching/designteach/teach/rubrics.html

TABLES

Table 1: T-test results showing the pre-test and post-test scores of the students based on the use of scoring rubric.

Measurement	N	$\overline{X}$	S	sd	t	p
Pre-test total (Researcher)	25	719.2	23.51	24	8.7	.000
Post-test total (Researcher)	25	1478	10.82			

Table 2: T-test results of the change of students' problem-solving success in pre-test and post-test, separated by rater.

	Measurement (Rubric)	N	$\overline{X}$	S	sd	t	p
Pre-test total	Three peer raters	3	146.8	33.26	4	1.1	.386
	Researcher	3	142.75	34.85			
Post-test total	Three peer raters	3	249.74	20.73	4	2	.862
	Researcher	3	250	21.29			

Table 3: T-test results of students' success of pre-test and post-test, separated by rater

QUESTIONS	SCORER	N	$\overline{X}$	S	sd	t	P
Pre-test of Student Number 4	Peer rater 1	3	41	38.95	4	.72	.479
	Researcher	3	43	39.58			
Pre-test of Student Number 4	Peer rater 2	3	40.5	38.96	4	1.07	.298
	Researcher	3	43	39.58			
Pre-test of Student Number 4	Peer rater 3	3	49	35.6	4	2.85	.010
	Researcher	3	43	39.58			
Pre-test of Student Number 18	Peer rater 1	3	14.5	30.68	4	.48	.634
	Researcher	3	15.25	31.09			
Pre-test of Student Number 18	Peer rater 2	3	19	31.56	4	1.24	.228
	Researcher	3	15.25	31.09			
Pre-test of Student Number 18	Peer rater 3	3	21.5	31.16	4	2.91	.009
	Researcher	3	15.25	31.09			
Pre-test of Student Number 20	Peer rater 1	3	83.25	22.72	4	1.09	.287
	Researcher	3	84.5	20.25			
Pre-test of Student Number 20	Peer rater 2	3	83.5	24.12	4	.55	.585
	Researcher	3	84.5	20.25			
Pre-test of Student Number 20	Peer rater 3	3	86	20.55	4	1.03	.316
	Researcher	3	84.5	20.25			
Post-test of Student Number 4	Peer rater 1	3	92	14.72	4	1.21	.239
	Researcher	3	95.25	12.51			
Post-test of Student Number 4	Peer rater 2	3	94.5	12.55	4	.21	.833
	Researcher	3	95.25	12.51			
Post-test of Student Number 4	Peer rater 3	3	96.5	9.19	4	.66	.514
	Researcher	3	95.25	12.51			
Post-test of Student Number 18	Peer rater 1	3	59.5	40.42	4	.33	.742
	Researcher	3	58.75	43.64			
Post-test of Student Number 18	Peer rater 2	3	59.25	42.89	4	.38	.705
	Researcher	3	58.75	43.64			
Post-test of Student Number 18	Peer rater 3	3	59.25	40.17	4	.22	.827
	Researcher	3	58.75	43.64			
Post-test of Student Number 20	Peer rater 1	3	95	11.58	4	1	.330
	Researcher	3	96	11.07			
Post-test of Student Number 20	Peer rater 2	3	96	11.07	4	-	-
	Researcher	3	96	11.07			
Post-test of Student Number 20	Peer rater 3	3	97.25	9.38	4	1.5	.135
	Researcher	3	96	11.07			

Table 4: The correlation of pre-test total scores between the researcher and three peers

	Measurement (Rubric)	Peer raters correlation	Researcher Correlation
Peer raters Correlation	Pearson Correlation	1	1.000(**)
	Sig. (2 tailed)	.	.014
	N	3	3
Researcher Correlation	Pearson Correlation	1.000(**)	1
	Sig. (2 tailed)	.014	.
	N	3	3

** Correlation is significant at the 0.01 level (2 tailed).

Table 5: The correlation of post-test total scores between the researcher and three peers

	Measurement (Rubric)	Evaluator Correlation	Researcher Correlation
Evaluator Correlation	Pearson Correlation	1	1.000(**)
	Sig. (2 tailed)	.	.016
	N	3	3
Researcher Correlation	Pearson Correlation	1.000(**)	1
	Sig. (2 tailed)	.016	.
	N	3	3

** Correlation is significant at the 0.01 level (2 tailed).