Investigating the Consistency between Students’ and Teachers’ Ratings for the Assessment of Problem-solving Skills with Many-facet Rasch Measurement Model*

Seyhan SARITAS AKYOL¹, Ismail KARAKAYA²

ABSTRACT

Purpose: To assess students’ problem-solving skills, this study aims to investigate the consistency between self- and peer-ratings in consideration of the teachers’ ratings in the process.

Method: This study was a descriptive study which examines the mathematical problem-solving skills with the MFRM model concerning self-, peer- and teachers’ ratings. The study group consisted of 57 sixth grade students studying in a secondary school in Ankara. The data collection procedure was as follows: i) the students were trained in how to use rubric during the first week, ii) they practiced the rubric and a performance task in the second week and, iii) three performance tasks were applied following consecutively. These tasks included non-routine problem situations and two analytical rubrics were developed. For data analysis, student, steps, rater type, and task were determined as facets and rater was defined as dummy facet, and reliability statistics related to each facet were estimated.

Findings: Ratings of performance tasks obtained from three-week data collection had high reliability coefficients according to MFRM modeling. The findings showed that self-, peer- and teachers’ ratings vary in terms of generosity/severity according to the weeks given the rater types. Generally, self-raters were the most generous raters, whereas teachers were the most severe raters. In addition, generosity/severity of peer-raters gets closer to generosity/severity of teachers from the first task to the third one.

Implications for Research and Practice: This research strengthens the possibility that peer-rating can provide reliable rating through appropriate training and practices.

© 2021 Ani Publishing Ltd. All rights reserved

Introduction

* This study was partly presented at the at the 27th International Congress on Educational Sciences (ICES/UEBK-2018) in Antalya, 18-22 April, 2018.

¹ Gazi University, TURKEY, e-mail: seyhansaritasakyol@gmail.com ORCID: 0000-0002-6172-207X

² Gazi University, TURKEY, e-mail: ikarakaya@gazi.edu.tr ORCID: 0000-0003-4308-6919
According to OECD (2014), some significant findings regarding students' performance in problem-solving could be juxtaposed as follows: In many countries, more than 10% of students are not capable of solving basic problems. While the average of the students in the high performance group is approximately 11.4% in OECD countries, it is determined that the average of students who do not reach the minimum performance level is approximately 21.4%. Half of the average students in OECD countries are not capable of solving problems that are slightly more difficult than basic problems. Within the framework of the countries participating in PISA 2012, the results show that the problem-solving skills, as one of the main objectives of mathematics programs and which is associated with different disciplines, have not yet been achieved at the desired level. In Turkey, it has been realized that students experience substantial fails in transferring their school learning into their daily lives while solving the problems they encounter (Gelbal & Kelecioglu, 2007).

Since the cognitive development of children is fast in the period of elementary school, a major development of problem-solving skills can be provided with appropriate approaches to children in this period. (Baykul, 2006, p. 61). Therefore, in educational activities, it is necessary to use teaching methods and techniques that shed light on the development of high-level cognitive skills, including problem-solving, as well as measurement and evaluation approaches that measure these skills more effectively (Kutlu, Dogan, & Karakaya, 2010, p. 15-16). Considering that the main purpose of assessments conducted in primary education is mainly to monitor the students (Baykul, 2006, p. 87), performance-based assessment studies that are not independent of the learning process and provide rich feedback to the student are of pivotal importance. However, due to lack of knowledge, the reluctance of teachers, the limited time allocated for the implementation of the curriculum, crowded classes, lack of resources and equipment, performance-based assessment practices in mathematics program cannot be performed in a suitable way (Bal & Doganay, 2010; Gelbal & Kelecioglu, 2007).

Since the cognitive development of children is fast in the period of elementary school, a major development of problem-solving skills can be provided with appropriate approaches to children in this period. (Baykul, 2006, p. 61). Therefore, in educational activities, it is necessary to use teaching methods and techniques that shed light on the development of high-level cognitive skills, including problem-solving, as well as measurement and evaluation approaches that measure these skills more effectively (Kutlu et al., 2010, pp. 15-16). Considering that the main purpose of assessments conducted in primary education is mainly to monitor the students (Baykul, 2006, p. 87), performance-based assessment studies that are not independent of the learning process and provide rich feedback to the student are of pivotal importance. However, due to lack of knowledge, the reluctance of teachers, the limited time allocated for the implementation of the curriculum, crowded classes, lack of resources and equipment, performance-based assessment practices in mathematics program cannot be performed in a suitable way (Bal & Doganay, 2010; Gelbal & Kelecioglu, 2007).
Another advantage of performance-based assessment is that the student's chance of finding the correct answer is eliminated. In this way, measurement errors are eliminated from a chance factor to provide a more reliable result for the student's ability level (Guler, 2014). However, when evaluating open-ended activities and performance-based studies used to measure students' high-level skills, such as problem-solving, it is often a problem to make as an objective assessment as possible. The lack of objective scoring of open-ended activities and performance tasks is also a reliability problem (Kutlu et al., 2010, p. 49; Romagnano, 2001). One of the difficulties is to decide how to score students, and the other difficulty is to ensure the reliability of the measurement (Cakici Eser & Gelbal, 2013). The number of raters can be increased to increase the reliability of the measurement process for non-objective scoring in open-ended activity applications and performance evaluations (Cakici Eser & Gelbal, 2013; Ebel, 1951 cited in Ilhan, 2015).

In the literature, that the use of well-prepared scoring key improves the quality of evaluation and contributes to more objective determination is accepted by many educators and researchers (Jonsson & Svingby, 2007; Karakaya, Saritas, & Salmaner, 2015; Kutlu et al. 2010; Parlak & Dogan, 2014). Defining each criterion in the rubric according to the desired performance level increases the probability of independent raters to give the same score (Moskal, 2000). Therefore, scoring reliability gains importance. Scoring reliability is the consistency between the scores obtained by scoring the measurement tool at different times by the same rater or by different raters (Guler, 2008; Macmillan, 2000). However, in CTT, estimating the size of different sources of variance requires multiple analyzes and; however, the error caused by the interaction between different sources of variance is not calculated. In addition, the estimation methods used in CTT are group dependent. In reliability studies based on GT, it is possible to examine the error variances that arise from raters, different sources of variability, and their interactions (Guler, 2008). However, the Rasch model test reliability, which is based on IRT, is higher than the reliability given by GT. That is because the Rasch model does not include error variance in the item and rater variance (Linacre, 1993). An important advantage of Rasch methods is that it attempts to evaluate objectively based on rater judgments (Hetherman, 2004). In this model, the abilities of individuals are estimated independently from the characteristics...
of certain item distributions and scores given by certain raters to performance (Smith Jr & Kulikowich, 2004). It provides information about the status of performances in unexpected situations for each element of each source of variability (Alharby, 2006). In addition, the MFRM (Many-facet Rasch measurement) model uses a common logit scale for the values of facets by providing a linear inter-facet connection for each source of variability.

In the literature, there are studies that compare the results obtained from MFRM modeling with CTT or GT (Atiğan, 2004; Brown, O’Gorman, & Du, 1996; Guler, 2008; Smith Jr & Kulikowich, 2004; Sata, 2019). In the study, the method that will be determined to calculate the reliability estimates can be selected according to conditions of the sources of variability. With the MFRM model, it is possible to take into account all sources of variability that may affect achievement scores (Baird, Hayes, Johnson, Johnson, & Lamprianou, 2013). There are studies that use the MFRM model to determine the psychometric properties of performance measurement in the assessment of high-level skills (Guler, 2014; Hetherman, 2004; Ilhan, 2015; Nakamura, 2002; Sata & Karakaya, 2020; Semerci, 2011a). In another study, the MFRM model was used in decision-making processes and reliable rater selection for the evaluation of projects (Tesio et al., 2015). In some studies that use the MFRM model, three types were identified as self, peer, and teacher as raters (Farrokhi, Esfandiari, & Dalili, 2011; Farrokhi, Esfandiari, & Schaefer, 2012; Karakaya, 2015; Semerci, 2011b), and in another study (Kose, Usta, & Yandi, 2016) two types of rater scoring consisting of peer and teacher were evaluated. However, this study has different characteristics than other studies in terms of ensuring that the consistency of student scores as self and peers are monitored in the process according to teacher scores and using the statistically powerful model MFRM model in the analysis of the study. This research aimed to investigate the consistency between self- and peer-ratings in consideration of teachers' ratings in the process. To assess the problem-solving skills of sixth grade students in primary school, student performances were rated by themselves, peers and teachers through the use of the rubric. For the purpose of this study, the following questions and sub-questions were investigated using the MFRM model.

What are the key points in the calibration report for the analysis of the self, peer and teacher ratings of the students' problem-solving skills with the MFRM model in the process that includes the first, second and third performance task applications?

In the analysis including self, peer and teacher ratings of three performance tasks;

1. What are the levels of achievement for the students?
2. What are the levels of difficulty/easiness for the problem-solving steps?
3. What are the levels of difficulty/easiness for the tasks?
4. What are the levels of severity/generosity for the rater types?
5. What are the statistics in the measurement report of the achievement of the students?
6. What are the statistics in the measurement report for the problem-solving steps?
7. What are the statistics in the measurement report for the rater type?
8. What are the statistics in the measurement report for the tasks?

**Method**

**Research Design**

In this research, students’ performance was rated by students, peers and teachers using rubric and this research aimed to examine the consistency between self and peer ratings in consideration of teacher ratings. This study is a descriptive study because it is aimed to gather detailed information about the research topic and explain this topic. Descriptive research tries to describe and explain what events, objects, institutions, groups and various areas are (Punch, 1998).

**Study Group**

The study group consisted of students from the 6th-grade students and three teachers from a state secondary school in Ankara in the spring term of 2014-2015 academic year. In the six-week implementation phase of this study, the integrity of the process is important because the development of student ratings in consideration of teacher scores is within the scope of this research. Therefore, students who did not participate in any of the performance task practices were excluded from the analysis. Thus, while the number of the students was 75, the data of 57 students who participated in three performance tasks were used in the analyzes.

**Research Instruments and Procedures**

In the preparation stage of performance tasks, four performance tasks including non-routine problem situations were prepared for the 6th-grade students, and expert opinions on the effectiveness of the tasks were obtained. To prepare the rubrics to evaluate performance tasks, firstly, the goal was determined according to the expected performance. Then, steps of problem-solving were examined, and levels were written according to expected performance. According to Polya (1973), the problem-solving process consists of understanding the problem, devising a plan, carrying out the plan, and looking back. It was tried to determine the qualifications corresponding to the performance levels of the defined problem-solving steps as concrete as possible so that they do not change from rater to rater. To collect evidence for the psychometric properties of the tools, content validity ratio developed by Lawshe (1975) was used. Critical value for Lawshe’s content validity ratio was 0.75 for nine expert opinions. The content validation indexes of the tools ranged from .78 to .89, so they were found to be higher than .75.

The data collection procedure took six-weeks. In the first week, the students were trained in how to use the rubric and they practiced the rubric and a performance task
in the second week. The other three performance tasks were applied in the following three weeks consecutively. The students first rated their own performances and then the papers were collected and their names were closed. Then, the performances were distributed randomly to the students and each student rated the performance of a peer. Students were given 15-20 minutes for tasks, 10 minutes for self-assessment, and 10 minutes for peer-assessment. In the sixth week, semi-structured questions were asked to gather the students' opinions in written texts.

**Data Analysis**

For the data analysis, student, steps, rater type, and task were determined as facets, and rater was defined as dummy facet. Reliability statistics related to each facet were estimated. In the analysis, generosity/severity of the self, peer and teacher raters were coded separately for each task to be monitored week by week. FACETS 3.71.4 (Linacre, 2014) package program was used for data analysis. As a result of the FACETS program analysis, iteration report, the map of facets, logit measurement estimates, standard error and fit statistics were obtained for each facet identified (Linacre, 1993). The number of iterations required to estimate well from the data depends on how well the data fit the Rasch model.

In this study, it is essential that the model data fit should be examined before performing the MFRM model. To ensure model data fit, the standardized residual values should be approximately 5% out of -2 to +2 and 1% out of -3 to +3 at most (Linacre, 2014). Accordingly, the model data fit indexes were explained for the analysis. In this research, 100 of the standardized residual values of the 2928 measurements in the analysis (approximately 3.4% of measurements) are out of -2 to +2, and 21 of the standardized residual values of the 2928 measurements (approximately 0.7% of measurements) are out of -3 to +3. Therefore, the data used in this study showed model fit for FACET analysis.

**Results**

In the analysis including self, peer and teacher ratings of three performance tasks:

1. What are the levels of achievement for the students?
2. What are the levels of difficulty/easiness for the problem-solving steps?
3. What are the levels of difficulty/easiness for the tasks?
4. What are the levels of severity/generosity for the rater types?

The first four sub-questions of study were explained according to Figure 1. Figure 1 shows the calibration map of the student's self, peer and teacher ratings using the MFRM model. The calibration map includes logit scales between + and - for student, step, rater type, task and rater. On the logit scale, which shows the performance of the individual's problem-solving skills, student abilities decrease from top to bottom. On the second facet, which illustrates the problem-solving steps, difficulties of the steps increase from top to bottom. On the third facet, which illustrates the severity of the raters, severity increases from top to bottom. The raters were coded in consonance with
the tasks implemented to see the differences clearly between the self, peer and teacher raters in the evaluation process on the facet of the rater type. The students' self-ratings in the first, second and third applications were coded as “Self1,” “Self2,” and “Self3,” respectively. Similarly, ratings of peer and teacher were coded as “Peer1,” “Peer2,” “Peer3,” and “Teacher1,” “Teacher2,” “Teacher3”. On the fourth facet, the task was ordered from top to bottom and from the easiest task to the most difficult task. Performance tasks were coded as “Task1,” “Task2,” and “Task3” according to the week applied. Finally, the fifth facet, which was the rater facet, was identified as a dummy facet, and the elements inside were equal to zero.

<table>
<thead>
<tr>
<th>Mosaiq+Student</th>
<th>+Type</th>
<th>+Task</th>
<th>RATING</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 +</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150 51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 20 55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45 48 49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 29 41 5 52 9</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32 44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34 36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 39 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 15 17 3 31 35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 23 26 37 39 40</td>
<td>Self 2 Self 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 37 4 42 53 56</td>
<td>Understand Self 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 14</td>
<td>Solve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43 54</td>
<td>Peer 1</td>
<td>Task 1 Task 3</td>
<td>2</td>
</tr>
<tr>
<td>0 46</td>
<td>* Peer 2</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>26 6</td>
<td>Plan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 57</td>
<td>Review</td>
<td>Teach.1 Teach.2 Teach.3 Peer 3 Task 2</td>
<td>---</td>
</tr>
<tr>
<td>28 38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1 +</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

**Figure 1. The Calibration Map**

Figure 1 showed that the students with the highest performance were the students 7 and 24, and the students with the lowest performance were the students 28 and 38. In the analysis which included the data of the three performance tasks, it could be said that the group was generally successful since the students' place on the logit scale was in a positive area. For the second facet, “Understand” (understanding the problem) and “Solve” (carrying out the plan) were found as easy steps while “Plan” (devising a plan) and “Review” (looking back) were found as difficult steps on the logit scale illustrating levels of difficulty/easiness for the problem-solving steps. For the third facet, “Self1,” “Self2,” and “Self3” were in the positive area so it is seen that students were generally generous in rating their own studies on the logit scale illustrating the levels of severity/generosity for the rater types. When peer ratings were examined, it is seen that differences were observed according to the weeks. “Peer1” representing
peer-ratings for the first performance task was the generous rater by staying above zero on the logit scale, while “Peer2” representing peer-ratings for the second performance task showed moderate severity/generosity. “Peer3” representing peer-ratings for the third performance task was in the negative area on the logit scale, and therefore it was found that the peer-ratings in the last week were more severity than the first and second weeks. The calibration map in Figure 1 explains that the teachers’ ratings were in the negative area on the logit scale for three weeks, and they were the most severity raters.

5. What are the statistics in the measurement report of the achievement of the students?

According to the analysis result, the statistics in the measurement report of the students’ achievement are given in Figure 2. As shown in Figure 2, the logit values of 57 students were between 1.56 and -0.50. The RMSE value of students’ achievement measurement was found as 0.14. For the separation index (G), this formula (4G+1)/3 indicates the number of levels of elements on the facet for each source of variability (Hetherman, 2004). The separation index was 3.58 and the reliability coefficient was 0.93. Reliability of the separation index can be interpreted as internal consistency because it is equivalent to KR-20 and Cronbach’s alpha in CTT or the generalizability coefficient in GT (Nakamura, 2000). The fixed effect hypothesis was tested by chi-square ($X^2=723.5$, $df=56$, $p=0.00$) and the null hypothesis was rejected. That is, there was a significant difference between the students’ performances. It was revealed that the student performances could be divided into approximately five levels. It was found that there was no central tendency effect on ratings. If the fit values are examined, it was observed that approximately 96% of the students, namely 55 students, were within the acceptable range in terms of infit and outfit indices. As a result of the analysis, it can be said that the performance shown in the implementation is not at the expected level for the two individuals who are out of the compliance and non-compliance value ranges. Also, it can be seen that two students’ performance in the tasks was not at the expected level.
<table>
<thead>
<tr>
<th>Obs#</th>
<th>Score</th>
<th>Obs#</th>
<th>Score</th>
<th>Pair-Mean</th>
<th>S.E.</th>
<th>Model</th>
<th>Init</th>
<th>Outlift</th>
<th>Statistic</th>
<th>N</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>193</td>
<td>3.45</td>
<td>56</td>
<td>3.69</td>
<td>3.56</td>
<td>1.56</td>
<td>1.00</td>
<td>0.00</td>
<td>1.10</td>
<td>0.90</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>3.45</td>
<td>56</td>
<td>3.69</td>
<td>3.56</td>
<td>1.52</td>
<td>0.80</td>
<td>-0.8</td>
<td>0.63</td>
<td>1.11</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>178</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>0.64</td>
<td>-0.2</td>
<td>0.45</td>
<td>0.96</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.13</td>
<td>0.10</td>
<td>0.51</td>
<td>0.95</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>162</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>0.57</td>
<td>2.00</td>
<td>1.54</td>
<td>0.77</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>151</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>0.71</td>
<td>-1.4</td>
<td>0.41</td>
<td>0.98</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>177</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>172</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>142</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>136</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>3.42</td>
<td>56</td>
<td>3.84</td>
<td>3.63</td>
<td>1.61</td>
<td>1.11</td>
<td>0.19</td>
<td>0.75</td>
<td>0.99</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Student Measurement Report
6. What are the statistics in the measurement report for the problem-solving steps?

As shown in Table 1, the logit values of “Understand”, “Plan,” “Solve,” and “Review” were between 0.27 and -0.28. The RMSE showing the difficulty/easiness levels of the problem-solving steps was found as 0.04. The separation index was 6.16 and the reliability coefficient was 0.97. The fixed effect hypothesis was tested by chi-square \( (\chi^2=155.9, df=3, p=0.00) \) and the null hypothesis was rejected. Thus, there was a significant difference between the level of difficulty/easiness of the steps in the three-week process to measure students’ problem-solving skills. For an efficient measurement in Rasch analysis, the infit and outfit values are expected to be between 0.5 and 1.5 (Linacre, 2014). Table 1 shows that the infit and outfit values were between 0.84 and 1.11, and therefore the model fitted the data well.

Table 1

<table>
<thead>
<tr>
<th>Steps</th>
<th>Obsvd Average</th>
<th>Fair-M Average</th>
<th>Model Measure S.E.</th>
<th>Infit MnSq ZStd</th>
<th>Outfit MnSq ZStd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understand</td>
<td>2.81</td>
<td>3.28</td>
<td>0.27 0.04</td>
<td>0.85 -3.2</td>
<td>0.84 -2.6</td>
</tr>
<tr>
<td>Solve</td>
<td>2.71</td>
<td>3.19</td>
<td>0.16 0.04</td>
<td>1.11 2.1</td>
<td>1.09 1.5</td>
</tr>
<tr>
<td>Plan</td>
<td>2.37</td>
<td>2.85</td>
<td>-0.15 0.03</td>
<td>0.96 -0.9</td>
<td>0.97 -0.4</td>
</tr>
<tr>
<td>Review</td>
<td>2.22</td>
<td>2.68</td>
<td>-0.28 0.03</td>
<td>1.11 2.2</td>
<td>1.09 1.7</td>
</tr>
</tbody>
</table>

RMSE=0.04 Adj S.D.=0.22 Separation=6.16 Reliability=0.97

Fixed chi-square=155.9 df=3 p=0.00

7. What are the statistics in the measurement report for the rater type?

The statistics related to levels of the severity/generosity for the rater type are given in Table 2. The logit values of the self-raters were 0.29, 0.39 and, 0.38, respectively. The logit values of peer-raters were 0.12, 0 and, -0.26, respectively. The logit values of teacher ratings were -0.30, -0.31 and, -0.31, respectively. The RMSE indicating severity/generosity of the rater type was found as 0.07. The separation index was 4.07 and the reliability coefficient showing the reliability of the scoring severity/generosity rank of the rater types was 0.94 sufficiently high. The fixed effect hypothesis was tested by chi-square \( (\chi^2=196.8, df=8, p=0.00) \), and the null hypothesis was rejected. Thus, there was a significant difference between level of severity/generosity of the rater type in the three-week process to measure students’ problem-solving skills. Significant chi-square, high separation index, high separation rate and reliability indicate that there is no halo effect in rating tasks. As can be seen in Table 2, the infit and outfit values were between 0.83 and 1.39 and it explains that the data fitted well to the model. Fit statistics also showed that the raters rated consistently.
What are the statistics in the measurement report for the tasks?

In the data analysis related to the three performance tasks, statistics and fit indexes regarding the difficulty levels of the tasks are given in Table 3. In Table 3, the logit values representing the difficulty levels of the “Task1,” “Task2,” and “Task3” which constituted the facet of the task, were found to be 0.14, -0.29 and 0.15, respectively. The RMSE showing the rating severity/generosity of the rater type was found to be 0.03. The index of separation was found to be 6.41. A high index of separation was also desirable for the tasks. The reliability coefficient, which shows how reliable the ranking of the difficulty levels of the performance tasks is, was 0.98, and it is sufficiently high. The fixed effect hypothesis was tested by chi-square ($X^2=138.9$, df=2, $p=0.00$) and the null hypothesis was rejected. Thus, there was a significant difference between the levels of difficulty/easiness of the tasks in the three-week process to measure students’ problem-solving skills. Table 3 shows that the infit and outfit values were between 0.92 and 1.11, and it indicates that the data fitted the model well.

### Table 3

#### Task Measurement Report

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Obsvd Average</th>
<th>Fair-M Average</th>
<th>Model Measure S.E.</th>
<th>Infit MnSq ZStd</th>
<th>Outfit MnSq ZStd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 3</td>
<td>2.72</td>
<td>3.17</td>
<td>0.15 0.03</td>
<td>1.11 2.2</td>
<td>1.09 1.5</td>
</tr>
<tr>
<td>Task 1</td>
<td>2.70</td>
<td>3.16</td>
<td>0.14 0.03</td>
<td>0.96 -0.9</td>
<td>0.92 -1.6</td>
</tr>
<tr>
<td>Task 2</td>
<td>2.22</td>
<td>2.68</td>
<td>-0.29 0.03</td>
<td>0.98 -0.5</td>
<td>1.01 0.1</td>
</tr>
</tbody>
</table>

RMSE=0.03 Adj S.D.=0.20 Separation=6.41 Reliability=0.98

Fixed chi-square=138.9 $df=2$ $p=0.00$
Discussion, Conclusion and Recommendations

Students, peers and teachers rated the first, second and third performance tasks to measure the problem-solving skills of the students using rubrics. According to the MFRM model, the model data fit for FACET analysis was examined, and it was concluded that the data fitted the model. It was found that the separation index and reliability coefficient values calculated by the FACET analysis of 57 students whose performance on problem-solving skills was measured were quite high. As the reliability coefficient of the facet of the student is known to be equivalent to the KR-20, Cronbach's alpha, and generalizability coefficient, it was concluded that the internal consistency reliability of ratings of the students, peers and, teachers on the performances of students' problem-solving skills was ensured according to the MFRM model. The high index of separation values, significant chi-square values, the difference of students' performances and the fact that students might be divided into different skill levels according to the calculated index of separation ratios indicated that the rater has no central tendency effect on ratings.

In the analysis of the data, the students were overall successful on the performance of problem-solving. It was considered that the success of the groups might also have an effect on the generosity of student raters in their ratings. In the analysis of the problem-solving steps, it was seen that the easiest steps were “understanding the problem” and “carrying out the plan” while the most difficult steps were “devising a plan” and “looking back”. Given that the step of “look back” was at the bottom of the logit scale indicates that the students' performance of checking and evaluating the solution was less successful. The separation index of the step facet was high in the analysis, which is desirable. It can be concluded that the problem-solving steps in the rubric do not measure the same cognitive component. The significant chi-square value for the step facet in the rubric supports this issue.

There is no considerable difference between the logit values of the tasks on the facet of the task, and the order from easy to difficult tasks was “Task1,” “Task3,” “Task2”. The chi-square value indicates that the tasks differed statistically. The significant chi-square value, high separation index and separation ratio support the conclusion that the tasks had different difficulties, and there was no halo effect in the rating. The most generous raters regarding the severity/generosity of the rater type were self-raters. According to the results of different studies in the literature (Farrokhi et al., 2011; Farrokhi et al., 2012; Karakaya 2015; Karakaya et al., 2015), self-evaluation scores were more generous than teacher scores. It can be said that students tend to give high scores to their own studies. However, when Matsuno (2006) compared self-, peer- and teachers' ratings in the study, it was found that self-ratings were severe, peer-ratings were generous, and teacher’s ratings were neither severe nor generous.

In this study, the first, second and third performance tasks of peer raters scored more severe than self-raters and more generous than teacher raters. This is similar to the results of the study conducted by Farrokhi et al. (2012). Yuzuak, Yuzuak, and Kaptan (2015) also concluded that peers scored more generously than teachers, and peer scores were consistent in consideration of the general analysis results. However,
Another noteworthy point in this study was the approach to teacher severity by decreasing the logit measurements of peer scorings from the first performance task to the third task. Considering the ratings of the teacher raters, it is seen that they rated approximately with the same severity when they were examined week by week. Besides that, there was no considerable difference in weeks in terms of severity/generosity of self-raters. In the analysis, it can be concluded from the logit values in the process that peer ratings came closer to teacher ratings week by week in terms of the severity/generosity. It is seen that the student performance was rated more accurately by teachers, and the severity of peer-raters is similar to teacher ratings. It shows that there was an improvement in the peer-ratings. It should not be forgotten that this research is limited with the measurement tool, skill, time and study group used. It can be interpreted that the assessment becomes more objective when the student experience in scoring increases. To test this idea, there is a need for structured studies with larger study groups at a longer programmed time.

In the assessment, rubric indicates to the teacher and the student what is necessary and what is important in the assessment. This is suitable for both critical decisions and learning evaluations (Jonsson & Svingby, 2007). Rubrics are also suitable reference points for monitoring students' own work (Brookhart, 2013, pp. 104-105). In reliability studies where students and teachers rated the same performance, it is seen that inter-rater reliability was ensured. Thus, the use of both student and teacher ratings was suggested as a practical method, especially in classroom assessments (Holster, 2012; Karakaya 2015; Karakaya et al., 2015; Yuzuak et al., 2015). Rubrics have the potential to make it easier for students to better understand concepts and skills in their subsequent tasks through effective feedback (Kuehl, Sofronas, & Lau, 2015). Additionally, it is thought that the students will have a better understanding of the performance levels expected when they perform self- and peer-assessment in the classroom, and this may affect the complex and time-consuming problem-solving skills positively.

However, the reliability coefficient of the rater type facet was high, and the chi-square value of the facet indicates that the ratings of self, peer and teacher raters differed significantly. It is not desirable that the chi-square value was significant for the raters. This finding may stem from the use of self and peer raters on the same facet as well as teacher raters. Also, in the literature, there are some suggestions about the findings in which the evaluations differ significantly. In his study Wang (2017) found that the practice and discussion in the training session had a positive effect on the accuracy of students' self- and peer assessment. Also, in another study, it was found that the rater training had a significant impact on rater severity, rater generosity, differing rater severity, and differing rater generosity (Sata, 2020). Karakaya (2015) underlined the importance of self and peer evaluations, rater training, and the use of rubrics. In this study, the raters were trained at first, and then general feedback was given to these raters in the following weeks. The findings indicated that rater types showed differences. These difference maybe because of the time limitation, lack of
familiarity with self and peer assessment, and the fact that problem-solving skills are a complex skill.

Considering all the results of this research study, it can be concluded that the MFRM model is an appropriate measurement model to be used in the classroom since it gives detailed information about each facet and each element used in the assessment of problem-solving skills with self, peer and teacher ratings. As further studies, it is recommended that more time can be allocated for self and peer assessment practices, and more students can take part in these practices. In this study, it was found that the most difficult step in the problem-solving steps was to “look back”. Thus, to ensure the development of controlling skills, evaluation performance and critical perspective towards the studies, there should be more self and peer assessment practices in the classroom. Also, mixed methods can be employed as a research design to explain the study results better in terms of the cause and effect. Practically speaking, the changes in student performance can be monitored through the involvement of students in the scoring process in the evaluation. In the evaluation of problem-solving skills, the study can be carried out by using the MFRM model but by selecting different facets for the analysis. Similar reliability studies can be carried out in the evaluation of other high-level cognitive skills. Also, scoring behaviors can be examined concerning the difficulty of tasks in self and peer-scoring.

References


**Problem Çözme Becerilerinin Değerlendirilmesinde Öğrenci ve Öğretmen Puanlarının Çok Yüzeyli Rasch Modeliyle İncelenmesi**

**Atıf:**

**Özet**

*Problem Durumu:* Problem çözme gibi üst düzey becerilerin ölçülmek istendiği performansa dayalı durum belirleme ve açık uçlu soruların puanlanmasında, öğrencilerin şansla doğru cevaba erişme olasılığı sıfıra indirgenir. Ancak değerlendirme sırasında kullanılan puanlama anahtarındaki beklenen performansına göre ölçütler, olabildiğince somut olmalı ve yapılan puanlamaların neselliği konusunda güvendiğini ve geçeri kantları ihtiyaç duymaktadır. İyi planlanmış ölçme durumları, iyi hazırlanmış puanlama anahtarları, puanlayıcı sayısının artırılması puanlamaların özelliğine karşı kullanlarının önlemleridir ve değerlendirme niteliğini artırmaktadır.

Değerlendirmeye öğrencilerin dahil edildiği iyi planlanmış ölçme durumları, öğrencilerin yapmış oldukları çalışmaları puanlarna geçirmiş olduğu bilisel süreçle eleştirel bir gözle yaklaşılmasını sağlamaktadır. Ayrıca öğrencilerin akranlarının çalışmalarını puanlama anahtarındaki ölçütlerle bağlantı katarak puanlamaları da öğrencilerin beklenen performans düzeylerine ilişkin ölçütler hakkındaki kavramlarını artırmaktadır. Ancak ülkemizde yapılan araştırmalarla göre okulun öğretmenlerin mutfedat yetiştirme için zaman bulamaması, bilgi eksikliği, istekliziğinin, sınıfların kalabalığının zorlamaları nedeniyle matematik programındaki ölçme değerlendirme uygulamaları çok sağlıklı gerçekleştiremediği bilinmektedir. Sınıf içi değerlendirme tekniklerinde öz, akran ve öğretmen puanlamalarının güvenilirlik ve geçerlik açısından incelendiği ve öğrencilerin puanlamalarının zaman içerisinde öğretmen puanlarıyla uymumunun çok yüzeyli Rasch ölçme modeliyle incelendiği çalışmaya rastlanmamıştır.
Araştırmanın Amacı: Araştırımda ilköğretim altıncı sınıf öğrencilerinin problem çözme becerilerini değerlendirmek üzere; öğrenci performanslarını öz, akran ve üç öğretmen tarafından analitik dereceli puanlama anahatları kullanılarak puanlanmış ve süreç içerisinde öz ile akran puanlamalarının öğretmen puanlamalarına göre uyumunun izlenmesi amaçlanmıştır. Çalışmanın analizinde istatistiksel olarak güçlü bir model olan çok yüzeyli Rasch ölçme modeli kullanılmıştır.


Uygulama süreci olarak; ilk hafta öğrenciler ölçüme araçları ve süreç hakkında bilgi verilmiş, ikinci hafta deneme uygulanması yapılmış, izleyen üç hafta analizde kullanılan üç performans görevi uygulanmıştır. Belirtilen 75 öğrenciyi üç uygulamanın tamamına katılan 18 kişi analizden çıkarılmıştır. Kalan 57 öğrencinin üç performans görevindeki performansları, analitik dereceli puanlama anahatları ile öz, akran ve üç öğretmen tarafından puanlanmıştır. Analiz için uygulamalar ait veriler yüzey olarak; öğrenci, aşama, görev, puanlayıcı türü ve yapay (dummy) yüzey olarak puanlayıcı yüzeyi tanımlanmış, her yüzeye ilişkin güvenilik istatistikleri kestirilmiştir.


Anahtar Sözcükler: Puanlama güvenilirliği, çok yếu yeli Rasch ölçü modeli, dereceli puanlama anahtarı, problem çözme.