Eurasian Journal of Educational Research 91 (2021) 281-300

Eurasian Journal of Educational Research
www.ejer.com.tr

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

Seyhan SARITAS AKYOL[1], Ismail KARAKAYA[2]

ARTICLE INFO

Article History:
Received: 16 Jan. 2020
Received in revised form: 19 May 2020
Accepted: 19 May 2020
DOI: 10.14689/ejer.2021.91.13

Keywords
inter-rater reliability, MFRM, rubric, peer-assessment, self-assessment

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 in the 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.

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). Within the scope of this research, it is considered that planning a process that includes students in the evaluation process by using self, peer and teacher scores with a performance-based assessment approach to evaluate problem-solving skills is an important experience that can provide rich feedback for the student. However, in such non-objective tests, the rater's opinion may come into play. The student's score may vary from rater to rater (Ilhan, 2015). 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 (Tekin, 1991, p. 70). Inter-rater reliability is the degree of consistency of different raters giving similar scores to each student's work independently (Baykul, 2010; Crocker & Algina, 1986).

In the literature, there are studies using the methods of test reliability based on classical test theory (CTT), generalizability theory (GT) and item response theory (IRT) for calculating inter- and intra-rater reliability. According to the results obtained from the comparison studies based on these three theories, reliability calculation methods give similar results (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 (Atilgan, 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.

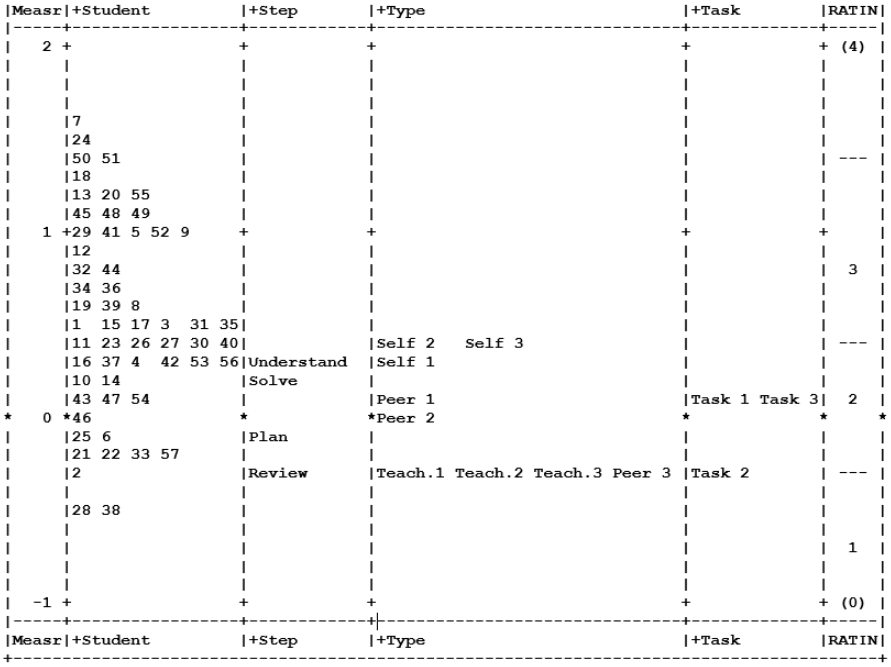
In this study, it is essential that the model data fit should be examined before performing the MFRM model. To ensure model data fit, the standardized residual values should be approximately 5% out of -2 to +2 and 1% out of -3 to +3 at most (Linacre, 2014). Accordingly, the model data fit indexes were 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.



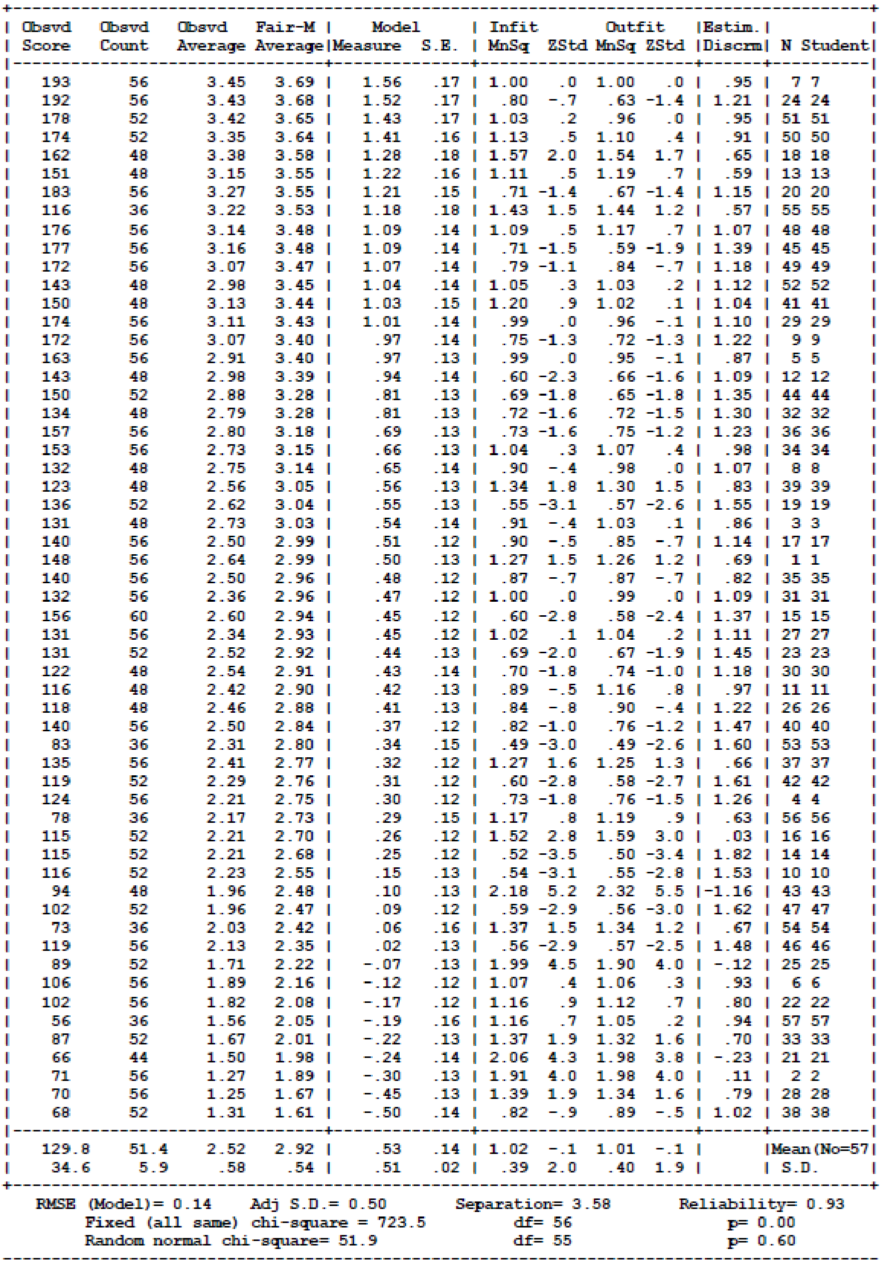
```
|Measr|+Student         |+Step       |+Type                          |+Task        |RATIN|
|-----+-----------------+------------+-------------------------------+-------------+-----|
|   2 +                 +            +                               +             + (4) |
|     |                 |            |                               |             |     |
|     |                 |            |                               |             |     |
|     |                 |            |                               |             |     |
|     |7                |            |                               |             |     |
|     |24               |            |                               |             |     |
|     |50 51            |            |                               |             | --- |
|     |18               |            |                               |             |     |
|     |13 20 55         |            |                               |             |     |
|     |45 48 49         |            |                               |             |     |
|   1 +29 41 5 52 9     +            +                               +             +     |
|     |12               |            |                               |             |     |
|     |32 44            |            |                               |             |  3  |
|     |34 36            |            |                               |             |     |
|     |19 39 8          |            |                               |             |     |
|     |1  15 17 3  31 35|            |                               |             |     |
|     |11 23 26 27 30 40|            |Self 2   Self 3                |             | --- |
|     |16 37 4  42 53 56|Understand  |Self 1                         |             |     |
|     |10 14            |Solve       |                               |             |     |
|     |43 47 54         |            |Peer 1                         |Task 1 Task 3|  2  |
*   0 *46               *            *Peer 2                         *             *     *
|     |25 6             |Plan        |                               |             |     |
|     |21 22 33 57      |            |                               |             |     |
|     |2                |Review      |Teach.1 Teach.2 Teach.3 Peer 3 |Task 2       | --- |
|     |                 |            |                               |             |     |
|     |28 38            |            |                               |             |     |
|     |                 |            |                               |             |     |
|     |                 |            |                               |             |  1  |
|     |                 |            |                               |             |     |
|     |                 |            |                               |             |     |
|  -1 +                 +            +                               +             + (0) |
|-----+-----------------+------------+-------------------------------+-------------+-----|
|Measr|+Student         |+Step       |+Type                          |+Task        |RATIN|
+----------------------------------------------------------------------------------------+
```

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.



```
+-------------------------------------------------------------------------------------------------+
|  Obsvd    Obsvd       Obsvd Fair-M |     Model      | Infit      Outfit    |Estim.|             |
|  Score    Count     Average Average|Measure   S.E.  | MnSq ZStd  MnSq ZStd |Discrm| N Student   |
|------------------------------------+----------------+----------------------+------+-------------|
|    193       56        3.45   3.69 |   1.56    .17 | 1.00   .0  1.00   .0 |   .95 |  7 7        |
|    192       56        3.43   3.68 |   1.52    .17 |  .80  -.7   .63 -1.4 |  1.21 | 24 24       |
|    178       52        3.42   3.65 |   1.43    .17 | 1.03   .2   .96   .0 |   .95 | 51 51       |
|    174       52        3.35   3.64 |   1.41    .16 | 1.13   .5  1.10   .4 |   .91 | 50 50       |
|    162       48        3.38   3.58 |   1.28    .18 | 1.57  2.0  1.54  1.7 |   .65 | 18 18       |
|    151       48        3.15   3.55 |   1.22    .16 | 1.11   .5  1.19   .7 |   .59 | 13 13       |
|    183       56        3.27   3.55 |   1.21    .15 |  .71 -1.4   .67 -1.4 |  1.15 | 20 20       |
|    116       36        3.22   3.53 |   1.18    .18 | 1.43  1.5  1.44  1.2 |   .57 | 55 55       |
|    176       56        3.14   3.48 |   1.09    .14 | 1.09   .5  1.17   .7 |  1.07 | 48 48       |
|    177       56        3.16   3.48 |   1.09    .14 |  .71 -1.5   .59 -1.9 |  1.39 | 45 45       |
|    172       56        3.07   3.47 |   1.07    .14 |  .79 -1.1   .84  -.7 |  1.18 | 49 49       |
|    143       48        2.98   3.45 |   1.04    .14 | 1.05   .3  1.03   .2 |  1.12 | 52 52       |
|    150       48        3.13   3.44 |   1.03    .15 | 1.20   .9  1.02   .1 |  1.04 | 41 41       |
|    174       56        3.11   3.43 |   1.01    .14 |  .99   .0   .96  -.1 |  1.10 | 29 29       |
|    172       56        3.07   3.40 |    .97    .14 |  .75 -1.3   .72 -1.3 |  1.22 |  9 9        |
|    163       56        2.91   3.40 |    .97    .13 |  .99   .0   .95  -.1 |   .87 |  5 5        |
|    143       48        2.98   3.39 |    .94    .14 |  .60 -2.3   .66 -1.6 |  1.09 | 12 12       |
|    150       52        2.88   3.28 |    .81    .13 |  .69 -1.8   .65 -1.8 |  1.35 | 44 44       |
|    134       48        2.79   3.28 |    .81    .13 |  .72 -1.6   .72 -1.5 |  1.30 | 32 32       |
|    157       56        2.80   3.18 |    .69    .13 |  .73 -1.6   .75 -1.2 |  1.23 | 36 36       |
|    153       56        2.73   3.15 |    .66    .13 | 1.04   .3  1.07   .4 |   .98 | 34 34       |
|    132       48        2.75   3.14 |    .65    .14 |  .90  -.4   .98   .0 |  1.07 |  8 8        |
|    123       48        2.56   3.05 |    .56    .13 | 1.34  1.8  1.30  1.5 |   .83 | 39 39       |
|    136       52        2.62   3.04 |    .55    .13 |  .55 -3.1   .57 -2.6 |  1.55 | 19 19       |
|    131       48        2.73   3.03 |    .54    .14 |  .91  -.4  1.03   .1 |   .86 |  3 3        |
|    140       56        2.50   2.99 |    .51    .12 |  .90  -.5   .85  -.7 |  1.14 | 17 17       |
|    148       56        2.64   2.99 |    .50    .13 | 1.27  1.5  1.26  1.2 |   .69 |  1 1        |
|    140       56        2.50   2.96 |    .48    .12 |  .87  -.7   .87  -.7 |   .82 | 35 35       |
|    132       56        2.36   2.96 |    .47    .12 | 1.00   .0   .99   .0 |  1.09 | 31 31       |
|    156       60        2.60   2.94 |    .45    .12 |  .60 -2.8   .58 -2.4 |  1.37 | 15 15       |
|    131       56        2.34   2.93 |    .45    .12 | 1.02   .1  1.04   .2 |  1.11 | 27 27       |
|    131       52        2.52   2.92 |    .44    .13 |  .69 -2.0   .67 -1.9 |  1.45 | 23 23       |
|    122       48        2.54   2.91 |    .43    .14 |  .70 -1.8   .74 -1.0 |  1.18 | 30 30       |
|    116       48        2.42   2.90 |    .42    .13 |  .89  -.5  1.16   .8 |   .97 | 11 11       |
|    118       48        2.46   2.88 |    .41    .13 |  .84  -.8   .90  -.4 |  1.22 | 26 26       |
|    140       56        2.50   2.84 |    .37    .12 |  .82 -1.0   .76 -1.2 |  1.47 | 40 40       |
|     83       36        2.31   2.80 |    .34    .15 |  .49 -3.0   .49 -2.6 |  1.60 | 53 53       |
|    135       56        2.41   2.77 |    .32    .12 | 1.27  1.6  1.25  1.3 |   .66 | 37 37       |
|    119       52        2.29   2.76 |    .31    .12 |  .60 -2.8   .58 -2.7 |  1.61 | 42 42       |
|    124       56        2.21   2.75 |    .30    .12 |  .73 -1.8   .76 -1.5 |  1.26 |  4 4        |
|     78       36        2.17   2.73 |    .29    .15 | 1.17   .8  1.19   .9 |   .63 | 56 56       |
|    115       52        2.21   2.70 |    .26    .12 | 1.52  2.8  1.59  3.0 |   .03 | 16 16       |
|    115       52        2.21   2.68 |    .25    .12 |  .52 -3.5   .50 -3.4 |  1.82 | 14 14       |
|    116       52        2.23   2.55 |    .15    .13 |  .54 -3.1   .55 -2.8 |  1.53 | 10 10       |
|     94       48        1.96   2.48 |    .10    .13 | 2.18  5.2  2.32  5.5 | -1.16 | 43 43       |
|    102       52        1.96   2.47 |    .09    .12 |  .59 -2.9   .56 -3.0 |  1.62 | 47 47       |
|     73       36        2.03   2.42 |    .06    .16 | 1.37  1.5  1.34  1.2 |   .67 | 54 54       |
|    119       56        2.13   2.35 |    .02    .13 |  .56 -2.9   .57 -2.5 |  1.48 | 46 46       |
|     89       52        1.71   2.22 |   -.07    .13 | 1.99  4.5  1.90  4.0 |  -.12 | 25 25       |
|    106       56        1.89   2.16 |   -.12    .12 | 1.07   .4  1.06   .3 |   .93 |  6 6        |
|    102       56        1.82   2.08 |   -.17    .12 | 1.16   .9  1.12   .7 |   .80 | 22 22       |
|     56       36        1.56   2.05 |   -.19    .16 | 1.16   .7  1.05   .2 |   .94 | 57 57       |
|     87       52        1.67   2.01 |   -.22    .13 | 1.37  1.9  1.32  1.6 |   .70 | 33 33       |
|     66       44        1.50   1.98 |   -.24    .14 | 2.06  4.3  1.98  3.8 |  -.23 | 21 21       |
|     71       56        1.27   1.89 |   -.30    .13 | 1.91  4.0  1.98  4.0 |   .11 |  2 2        |
|     70       56        1.25   1.67 |   -.45    .13 | 1.39  1.9  1.34  1.6 |   .79 | 28 28       |
|     68       52        1.31   1.61 |   -.50    .14 |  .82  -.9   .89  -.5 |  1.02 | 38 38       |
|------------------------------------+----------------+----------------------+------+-------------|
|  129.8     51.4        2.52   2.92 |    .53    .14 | 1.02  -.1  1.01  -.1 |       |Mean(No=57   |
|   34.6      5.9         .58    .54 |    .51    .02 |  .39  2.0   .40  1.9 |       | S.D.        |
+-------------------------------------------------------------------------------------------------+
   RMSE (Model)= 0.14    Adj S.D.= 0.50       Separation= 3.58        Reliability= 0.93
        Fixed (all same) chi-square = 723.5        df= 56                  p= 0.00
        Random normal chi-square= 51.9             df= 55                  p= 0.60
---------------------------------------------------------------------------------------------------
```

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 (X^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

Steps Measurement Report

	Obsvd	Fair-M	Model		Infit		Outfit	
Steps	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd
Understand	2.81	3.28	0.27	0.04	0.85	-3.2	0.84	-2.6
Solve	2.71	3.19	0.16	0.04	1.11	2.1	1.09	1.5
Plan	2.37	2.85	-0.15	0.03	0.96	-0.9	0.97	-0.4
Review	2.22	2.68	-0.28	0.03	1.11	2.2	1.09	1.7
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 (X^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.

Table 2

Rater Type Measurement Report

Rater Type	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
Self2	3.02	3.37	0.39	0.07	1.39	3.6	1.33	2.7
Self3	3.35	3.36	0.38	0.08	1.32	2.4	1.05	0.4
Self1	3.27	3.30	0.29	0.08	1.31	2.4	1.13	0.9
Peer1	3.11	3.15	0.12	0.09	0.94	-0.4	0.83	-1
Peer2	2.54	3.05	0	0.07	1.07	0.7	1.11	1
Peer3	2.68	2.71	-0.26	0.08	1.04	0.4	1	0
Teacher3	2.41	2.66	-0.30	0.04	1.07	1.2	1.14	2.1
Teacher2	1.87	2.65	-0.31	0.03	0.86	-3.2	0.87	-2.7
Teacher1	2.41	2.65	-0.31	0.04	0.89	-2.3	0.87	-2.5
RMSE=0.07	Adj S.D.=0.28		Separation=4.07		Reliability=0.94			
Fixed chi-square=196.8	*df*=8	*p*=0.00						

8. 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

Tasks	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
Task 3	2.72	3.17	0.15	0.03	1.11	2.2	1.09	1.5
Task 1	2.70	3.16	0.14	0.03	0.96	-0.9	0.92	-1.6
Task 2	2.22	2.68	-0.29	0.03	0.98	-0.5	1.01	0.1
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, according to some studies, including peer scoring and teacher scoring, peer raters scored more severely than the teacher (Karakaya, 2015; Nakamura, 2002).

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

Alharby, E.R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, the generalizability theory and the many facet Rasch measurement within the context of performance assessment* (Unpublished doctoral dissertation). The Pennsylvenia State University, Pennsylvania.

Atilgan, H. (2004). *Genellenebilirlik kurami ve cok degişkenlik kaynakli Rasch modelinin karsilaştirilmasina iliskin bir arastirma* (Unpublished doctoral dissertation). Hacettepe University, Ankara.

Baird, J.A., Hayes, M., Johnson, R., Johnson, S. & Lamprianou, I. (2013). Marker effects and examination reliability a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling. Oxford: University of Oxford for Educational Assessment. Retrieved from https://dera.ioe.ac.uk/17683/1/2013-01-21-marker-effects-and-examination-reliability.pdf.

Bal, A. P., & Doganay, A. (2010). İlkogretim besinci sinif matematik ogretiminde olcme-degerlendirme surecinde yasanan sorunlarin analizi [An analysis of problems encountered in the process of measurement and evaluation in teaching mathematics at primary school 5th grade]. *Educational Administration: Theory and Practice 16*(3), 373-398.

Baykul, Y. (2006). *İlkogretimde matematik ogretimi.* Ankara: Pegem Akademi.

Baykul, Y. (2010). *Egitimde ve psikolojide olcme: Klasik test teorisi ve uygulamasi.* İkinci Baski. Ankara: Pegem Akademi.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and granding.* Virginia USA: ASCD Alexandria.

Brown, William L., O'Gorman, K., & Du, Y. (1996). *The reliability and validity of mathematics performance assessment.* Paper presented at the annual meeting of the American Educational Research Association, NY, New York.

Crocker, L., & Algina, J. (1986). *Introduction to classical & Modern test theory.* USA: Harcourt Brace Javanovich College.

Cakici Eser, D., & Gelbal, S. (2013). Genellenebilirlik kurami ve lojistik regresyona dayali hesaplanan puanlayicilar arasi tutarligin karsilastirilmasi [Comparison of interrater agreement calculated with generalizability theory and logistic regression]. *Kastamonu Education Journal, 21*(2), 421-438. Retrieved from http://www.kefdergi.com/pdf/21_2/21_2_2.pdf

Farrokhi F., Esfandiari R., & Dalili, M. V. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal* 15: 70-77.

Farrokhi F., Esfandiari R., & Schaefer E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal, 34*(1), 79-102.

Gelbal, S., & Kelecioglu, H. (2007). Ogretmenlerin olcme ve degerlendirme yontemleri hakkindaki yeterlik algilari ve karsilastiklari sorunlar [Teachers' proficiency perceptions of about the measurement and evaluation techniques and the problems they confront]. *Hacettepe University Journal of Education, 33*(33), 135-145.

Guler, N. (2008). *Klasik test kurami genellenebilirlik kurami ve Rasch modeli üzerine bir arastirma* (Unpublished doctoral dissertation). Hacettepe University, Ankara.

Guler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research*, 55, 73-90.

Hetherman, S. C. (2004). *An application of multi-faceted Rasch measurement to monitor effectiveness of the written composition in English in the New-york City Department of Education* (Unpublished doctoral dissertation). Colombia University, Colombia.

Holster, T. A. (2012). Many-faceted Rasch analysis of student peer assessment. 76,69-86.

Ilhan, M. (2015). *Standart ve solo taksonomisine dayali rubrikler ile puanlanan acik uclu matematik sorularinda puanlayici etkilerinin cok yuzeyli Rasch olcme modeli ile incelenmesi* (Unpublished doctoral dissertation). Gaziantep University, Gaziantep.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review,* 2, 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Karakaya, I. (2015). Comparison of self, peer and instructor assessments in the portfolio assessment by using many facet Rasch model. *Journal of Education and Human Development,* 4(2), 182-192. https://doi.org/10.15640/jehd.v4n2a22

Karakaya, I., Saritas, S., & Salmaner, R. (2015). *Assessment of performance-based tasks within the context of statistics lesson with multi-faceted Rasch model.* Paper presented at the annual meeting of The International Congress on Education for the Future: Issues and Challenges (ICEFIC 2015), Ankara University, Ankara.

Kose, I. A., Usta, H. G., & Yandi, A. (2016). Sunum yapma becerilerinin cok yuzeyli Rasch analizi ile degerlendirilmesi [Evaluation of presentation skills by using many facets Rasch model]. *Abant İzzet Baysal University Journal of Faculty of Education, 16*(4).

Kuehl, G., Sofronas, K., & Lau, A. (2015). *Pre-service and in-service teachers' rubric assessments of mathematical problem solving. proceedings 2014.* Paper presented at the annual meeting of the NERA Conference Proceedings 2014. 1.

Kutlu, O., Dogan, C. D., & Karakaya, I. (2010). *Ogrenci basarisinin belirlenmesi - performansa ve portfolyoya dayalı durum belirleme.* Ankara: Pegem Akademi.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology,* 28, 563–575.

Linacre, J. M. (1993). *Generalizability theory and many facet Rasch measurement.* Paper presented at annual meeting of The American Educational Research Association. Georgia, Atlanta.

Linacre, J. M. (2014). A user's guide to FACETS Rasch-model computer programs.

Macmillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection interrater variability in large, sparse data sets. *The Journal of Experimental Education, 68*(2), 167-190.

Matsuno, S. (2006). *Self, peer, and teacher assessment in Japanese university EFL writing classrooms* (Unpublished doctoral dissertation). Temple University, Pennsylvania.

Moskal, B. M. (2000). Scoring rubrics: what, when and how? *Practical Assessment, Research and Evaluation, 7*(3), 70-80.

Nakamura, Y. (2000). Many facet Rasch based analysis of communicative language testing results. *Journal of Communication Students,* 12, 3-13.

Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies,* 44.

Organisation for Economic Co-operation and Development (OECD). (2014). PISA 2012 results: Creative problem-solving: Students' skills in tackling real-life problems.

Parlak, B., & Dogan, N. (2014). Dereceli puanlama anahtari ve puanlama anahtarindan elde edilen puanlarin uyum düzeyleri [Comparison of answer key and scoring rubric for the evaluation of the student performances]. *Hacettepe University Journal of Education 29*(2), 189-197.

Polya, G. (1973). *How to sove it? A new aspect of mathematical method.* Second edition. New Jersey: Princeton University.

Punch, K. F. (1998). *Introduction to social research: Quantitative and qualitative approaches.* Thousand Oaks, CA: Sage.

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Principles and Standards for School Mathematics (NCTM) 94*(1), 22.

Semerci, C. (2011a). Analyzing micro teaching applications with many-facet Rasch measurement model. *Education & Science,36*(161), 14-25.

Semerci, C. (2011b). Doktora yeterlikler cercevesinde ogretim uyesi, akran ve oz degerlendirmelerin Rasch olcme modeliyle analizi [Analysis of faculty member, peer and self evaluation by applying Rasch model within the scope of doctorate competencies]. *Journal of Measurement and Evaluation in Education and Psychology, 2*(2), 164-17.

Smith Jr, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement, 64*(4), 617- 639.

Sata, M. & Karakaya, I. (2020). Investigation of the use of electronic portfolios in the determination of student achievement in higher education using the many-facet Rasch measurement model. *Educational Policy Analysis and Strategic Research, 15*(1), 7-21. doi:10.29329/epasr.2020.236.1

Sata, M. (2019). *Performans degerlendirme surecinde puanlayici egitiminin puanlayici davranislari uzerindeki etkisinin incelenmesi* (Unpublished doctoral dissertation). Gazi University, Ankara.

Tekin, H. (1991). *Egitimde olcme ve degerlendirme.* Ankara: Yargi.

Tesio, L., Simone, A., Grzeda, M. T., Ponzio, M., Dati, G., Zaratin, P., Perucca, L., & Battaglia, M. A. (2015). Funding medical research projects: Taking into account referees' severity and consistency through many-faceted Rasch modeling of projects' scores. *Journal of Applied Measurement, 16*(2), 129-152.

Wang D. (2017) *Self-and peer assessment of oral presentation in advanced Chinese classrooms: An exploratory study. In Chinese as a second language assessment* (pp. 271-286). Springer, Singapore.

Yuzuak, A. V., Yuzuak, B., & Kaptan, F. (2015). Performans görevinin akran gruplar ve ogretmen yaklasimlari dogrultusunda çok-yuzeyli Rasch olcme modeli ile analizi [A many-facet Rasch measurement approach to analyze peer and teacher assessment for authentic assessment task]. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 1-11.

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

Atıf:

Saritas Akyol, S., & Karakaya, I. (2021). Investigating the consistency between students' and teachers' ratings for the assessment of problem-solving skills with many-facet rasch measurement model. *Eurasian Journal of Educational Research 91, 281-300*, DOI: 10.14689/ejer.2021.91.13

Özet

Problem Durumu: Problem çözme gibi üst düzey becerilerin ölçülmesinde kullanılan performansa dayalı durum belirleme ve açık uçlu soruların puanlanmasında, öğrencilerin şansla doğru cevaba erişme olasılığı sıfıra indirgenir. Ancak değerlendirmede kullanılan puanlama anahtarındaki beklenen performansa göre ölçütler, olabildiğince somut olmalı ve yapılan puanlamaların nesnelliği konusunda güvenirlik ve geçerlik kanıtlarına ihtiyaç duyulmaktadır. İyi planlanmış ölçme durumları, iyi hazırlanmış puanlama anahtarları, puanlayıcı sayısının artırılması puanlamanın öznelliğine karşın kullanılan önlemlerdir ve değerlendirmenin niteliğini artırmaktadır.

Değerlendirmeye öğrencilerin dahil edildiği iyi planlanmış ölçme durumları, öğrencilerin yapmış oldukları çalışmaları puanlarken geçirmiş olduğu bilişsel sürece eleştirel bir gözle yaklaşmasını sağlamaktadır. Ayrıca öğrencilerin akranlarının çalışmalarını puanlama anahtarındaki ölçütlere bağlı kalarak puanlamaları da öğrencilerin beklenen performans düzeylerine ilişkin ölçütler hakkındaki kavrayışlarını artırmaktadır. Ancak ülkemizde yapılan araştırmalara göre okulda öğretmenlerin müfredatı yetiştirmek için zaman bulamaması, bilgi eksikliği, isteksizliği, sınıfların kalabalıklığı nedenlerinden dolayı matematik programındaki ölçme değerlendirme uygulamalarını çok sağlıklı gerçekleştiremediği bilinmektedir. Sınıf içi değerlendirmelerdeki bu eksiklik, bu alanda yapılan araştırmaların sınırlılığını da beraberinde getirmektedir. Alan yazında sınıf içi ölçme değerlendirme tekniklerinde öz, akran ve öğretmen puanlamalarının güvenirlik ve geçerlik açısından incelendiği ve öğrenci puanlamalarının zaman içerisinde öğretmen puanlarıyla uyumunun çok yüzeyli Rasch ölçme modeliyle incelendiği çalışmaya rastlanılmamıştır.

Araştırmanın Amacı: Araştırmada ilköğretim altıncı sınıf öğrencilerinin problem çözme becerilerini değerlendirmek üzere; öğrenci performansları öz, akran ve üç öğretmen tarafından analitik dereceli puanlama anahtarı 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.

Araştırmanın Yöntemi: Araştırmada öğrenci performanslarının öz, akran ve öğretmen tarafından dereceli puanlama anahtarı kullanılarak puanlanması ve süreçte öz ile akran puanlamalarının öğretmen puanlamalarına göre uyumunun incelenmesi amaçlanmıştır. Bu çalışmada mevcut konu hakkında detaylı bilgi toplama ve konuyu açıklama amaçlandığından betimsel bir araştırmadır. Araştırmada çalışma grubu, 2014-2015 eğitim-öğretim yılı bahar döneminde, Ankara ili, Keçiören ilçesinde bulunan bir devlet okulunda öğrenim gören 6. sınıf düzeyindeki 75 öğrenci oluşturmaktadır. Araştırmada, öğrencilerin problem çözme becerilerinin değerlendirilmesi için üç öğretmen puanlayıcı belirlenirken gönüllülük esas alınmıştır.

Araştırmanın amacı doğrultusunda, öğrencilerin problem çözme becerilerini ölçmek üzere rutin olmayan problem durumu içeren dört tane performans görevi geliştirilmiştir. Geliştirilen görevlerin öz, akran ve öğretmenler tarafından değerlendirilmeleri için analitik dereceli puanlama anahtarları (rubrik) oluşturulmuştur. Dereceli puanlama anahtarının aşamaları Polya' nın (1973) problem çözme süreci temel alınarak; problemin anlaşılması, çözüm ile ilgili stratejinin seçilmesi, seçilen stratejinin uygulanması, çözümün değerlendirilmesi olarak dört aşamada ele alınmıştır.

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

Araştırmanın Bulguları: Öz, akran ve öğretmen puanlamaları ile problem çözme becerisine ilişkin performansı ölçülen öğrencilerin FACET analiz sonucu kalibrasyon haritasına bakıldığında, genel olarak logit cetveldeki yeri pozitif alanda olduğu için grubun genel olarak başarılı olduğu gözlenmiştir. Öğrenci yüzeyine ilişkin hesaplanan ayırma indeksi ve güvenirlik katsayı değerlerinin oldukça yüksek olduğu belirlenmiştir. Öğrenci yüzeyine ait güvenirlik katsayısı, KR-20, Cronbach alfa ve genellenebilirlik katsayısına eş değer olduğundan, çalışmada öğrencilerin problem çözme becerilerine ilişkin performanslarındaki öz, akran ve öğretmen puanlamaları, çok yüzeyli Rasch ölçme modeline göre iç tutarlılık anlamında güvenirliği sağlamaktadır. Ayırma indekslerinin yüksek olması, öğrencilerin performanslarda anlamlı olarak farklılaşması ve hesaplanan ayırma indeksi oranlarına göre

öğrencilerin performans olarak farklı beceri düzeylerine ayrılabileceğinin görülmesi, yapılan puanlamalarda puanlayıcıların merkeze yönelme etkisinin olmadığı sonucuna işarettir. Dereceli puanlama anahtarındaki aşamaların zorluk/kolaylık düzeylerini gösteren logit cetvelde "Anlama" (problemi anlama) ile "Uygulama" (seçilen çözüm stratejisini uygulama) kolay yapılan aşamalar olarak belirlenirken "Çözüm yolu" (çözüm ile ilgili stratejinin seçimi) ile "Kontrol etme" (çözümü kontrol etme ve değerlendirme) aşamaları ise zor yapılan aşamalar olarak yer almıştır. Aşama yüzeyine ait ayırma indeksinin büyük olması istenilen bir durumdur ve analizde de yüksek çıkmıştır. Bu durum puanlama anahtarındaki problem çözme aşamalarının aynı kapsamı yoklamadığına işaret olarak düşünülebilir. Puanlama anahtarının aşama yüzeyine ait ki-kare değeri de her bir aşamanın istatistiksel olarak anlamlı farklılaşması da bu durumu desteklemektedir. Diğer bir yüzey olan görev yüzeyinde görevler arasında logit değer olarak büyük bir fark bulunmazken kolaydan zora doğru sıralaması "Görev1", "Görev3", "Görev2" şeklindedir. Ki-kare değerinin farklılaşması, ayırma indeksinin ve ayırma oranının da yüksek çıkması görevlerin farklı zorluklarda olması ve yapılan puanlamalarda halo etkisi olmadığı sonucunu desteklemektedir. Puanlayıcı türü yüzeyinin bulunduğu logit cetvel haftalara göre incelendiğinde; "Öz1", "Öz2" ve "Öz3" cetvelde pozitif alanda olduğu için öğrencilerin kendi çalışmalarını puanlamalarında genel olarak cömert oldukları belirlenmiştir. Akran puanlamaları incelendiğinde puanlama katılık/cömertliklerinde haftalara göre bir farklılaşma olduğu görülmektedir. Birinci uygulamadaki akran puanlamalarını gösteren "Akran1" logit cetvelde sıfırın üzerinde kalarak cömert puanlayıcı olurken, ikinci haftaki akran puanlarını gösteren "Akran2" orta düzeyde katılık/cömertlik göstermiştir. "Akran3" ise logit cetvelde negatif alanda yer alarak son uygulamadaki akran puanlamalarının birinci ve ikinci haftaya göre puanlamada daha katı olduğu belirlenmiştir. Kalibrasyon haritasına bakıldığında genel olarak öğretmen puanlamalarının da üç uygulamada logit cetvelde negatif alanda yer aldıkları gözlenerek en katı puanlayıcılar oldukları belirlenmiştir. Uygulamanın verilerini içeren analiz sonucu her bir yüzey için hesaplanan ayırma indeksi ve istatistik değerleri hesaplanmış ve güvenirliğe ilişkin kat sayılar istenen aralıklarda bulunmuştur.

Araştırmanın Sonuçları ve Öneriler: Araştırma sonucunda genel olarak öz puanlayıcı cömert, öğretmen puanlayıcı türü ise katı puanlayıcı olarak yer almıştır. Puanlayıcı türü olarak öz ve öğretmen puanlayıcılar katılık/cömertlik açısından haftalara göre büyük farklılaşma göstermemiştir. Ancak akran puanlayıcının katılık/cömertlik durumu, birinci uygulamadan üçüncü uygulamaya öğretmen puanlayıcıların katılık/cömertliğine yaklaşmıştır. Bu noktadan hareketle, puanlayıcı türü olarak öğretmenlerin öğrencilerin gerçek performanslarına en yakın değerlendirmeleri yaptıkları varsayıldığında, akran puanlayıcıların katılığının öğretmen puanlayıcılara yaklaşması durumundan öğrencilerin akranlarını değerlendirmede bir gelişim gözlendiği çıkarımı yapılabilir. Çok yüzeyli Rasch ölçme modelinin, problem çözme becerilerinin öz, akran ve öğretmen puanları ile değerlendirilmesinde kullanılan her bir yüzey ve her yüzeyin her elemanına ilişkin ayrıntılı bilgiler sağlaması katkısı ile sınıf içinde de kullanımı elverişli bir ölçme modeli olabileceği sonucuna varılmıştır.

Anahtar Sözcükler: Puanlama güvenirliği, çok yüzeyli Rasch ölçme modeli, dereceli puanlama anahtarı, problem çözme.