Educational Policy Analysis and Strategic Research

Volume 15, Issue 1 March 2020

epasr.penpublishing.net

ISSN: 1949-4270 (Print) 1949-4289 (Online)

Investigation of the Use of Electronic Portfolios in the Determination of Student Achievement in Higher Education Using the Many-Facet Rasch Measurement Model

Mehmet Sata & Ismail Karakaya

To cite this article

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

Published Online	March 24, 2020
Article Views	17 single - 24 cumulative
Article Download	24 single - 34 cumulative
DOI	https://doi.org/10.29329/epasr.2020.236.1

Educational Policy Analysis and Strategic Research, V 15, N 1, 2020

Investigation of the Use of Electronic Portfolios in the Determination of Student Achievement in Higher Education Using the Many-Facet Rasch Measurement Model

Mehmet ŞATA[1]

Ağrı İbrahim Çeçen University

İsmail KARAKAYA[2]

Gazi University

Abstract

This study aimed to determine the rater behavior in the evaluation process of student electronic portfolios used to measure student achievement in higher education, and thus to evaluate the usability of the electronic portfolio system. Considering that rater behavior adversely affects both validity and reliability in determining the performance of individuals, it is important to identify the effect of this factor and evaluate the related results in line with this effect. The data of the study were collected from the students enrolled in an English language teaching program at Gazi University Gazi Education Faculty within the scope of the measurement and assessment course in the fall semester of 2017-2018. An analytic rubric developed by the researchers was used in the evaluation of the student electronic portfolios. The study included two participants groups consisting of three raters and 61 students (11 male, 50 female). In the analysis of the data, the many-facet Rasch measurement model was used as an analysis method since it was appropriate for the nature of the current data set. When the findings of the study were examined, it was found that one or more rater behaviors interfered with the performance of the individual in the use of non-objective measurement tools, and consequently negatively affected the validity and reliability of the measurements. In conclusion, it can be stated that the individual's performance related to electronic portfolios in higher education is generally affected by the rater behavior in the evaluation process independent of the measurement tool. In addition, it has been confirmed that electronic portfolios can be used to determine individual performance in higher education.

Keywords: Electronic portfolios, rater behavior, higher education, many-facet Rasch, validity.

DOI: 10.29329/epasr.2020.236.1

[1]Dr., Faculty of Education, Ağrı İbrahim Çeçen University University, Turkey, ORCID: 0000-0003-2683-4997

Correspondence: mehmetwsata@gmail.com

[2] Prof.Dr., Faculty of Education, Gazi University, Turkey, ORCID: 0000-0003-4308-6919, Email:ikarakaya2002@gmail.com

Introduction

In today's world, beyond acquiring knowledge, it has become necessary to develop high-level mental skills, such as decision-making, critical thinking, and problem-solving (Barak, Ben-Chaim, & Zoller, 2007). Gaining or developing these skills has become the focal point of curricula (Boddy, Watson & Aubusson, 2003; Riedler &Eryaman, 2016; Watts, Jofili & Bezerra, 1997). High-level mental skills take a long time to acquire and vary from individual to individual. They are process-oriented and measured by complementary measurement tools (portfolio, electronic portfolio [e-portfolio], performance tasks, etc.). Kutlu, Doğan and Karakaya (2014) emphasized that high-level mental skills referred to the whole of the cognitive, affective and psychomotor characteristics of an individual in the process of exhibiting her/his abilities. It was also suggested that the use of multiple-choice tests was not appropriate for the measurement of high-level cognitive skills, and that such test formats were more suitable for measuring knowledge and lower-level cognitive skills (Ebel, 1965; Kutlu et al., 2014). Thus, it is necessary to use novel test/assessment tools to measure high-level cognitive skills.

One of the novel measurement tools that help students develop their high-level cognitive skills and effectively reflects the development of students is an e-portfolio (Egan, 2012; Jenson, 2011). Unlike traditional measurement tools, e-portfolios are both process- and outcome-oriented, and they have a wide range of use in education, from primary to higher levels (Barker, 2005). Jenson (2011) stated that when e-portfolios were used in education, they helped students develop their high-level cognitive skills. However, despite the advantages of using e-portfolios in education, there are also certain disadvantages. When the literature is examined, it is seen that e-portfolios take a long time to prepare, require technological competence, present difficulty in standardization of scoring, and have low objectivity in assessment compared to the traditional measurement tools (Bahar, Nartgün, Durmuş & Bıçak, 2006; Chang, Tseng, Chou & Chen, 2011; Hung, 2012).

One of the objectives of performance assessment is to determine the competency of an individual in relation to the measured performance through accurate and reliable scoring (Johnson, Penny & Gordon, 2008). In other words, if an individual receives the same or similar scores upon completing different performance tasks or being scored by different raters, the objectivity of the assessment is considered to be high. The scores assigned in the evaluation of individual performance are attributed to reliability while the inferences made using these scores are associated with validity (Johnson, Penny & Gordon, 2008). Therefore, achieving a high level of objectivity in the performance assessment process would also increase both reliability and validity. In the literature, to ensure the objectivity of performance assessment, rubrics (holistic or analytic) (Haladyna, 1997; Kutlu et al., 2014; Oosterhof, 2003), multiple raters (Gronlund, 1977; Kubiszyn & Borich, 2013), and rater training (Bernardin & Buckley, 1981; Haladyna, 1997; Lumley & McNamara, 1995) have been recommended.

The current study used both a rubric and more than one rater for a more objective measurement during the assessment process of the student e-portfolios.

Although the use of contemporary methods helps improve objectivity in the assessment of individual performance, it cannot achieve complete objectivity as in traditional measurement tools. In the performance assessment process, one or more rater behaviors often interfere with scoring (Haladyna, 1997). Since rater behaviors that interfere with individual performance are attributed to the variance that is unrelated to the structure of the measure, they pose a direct threat to the validity (Jonsson & Svingby, 2007; Messick, 1996). Therefore, determining the rater behaviors that have a negative effect on objectivity in performance assessment is important for the validity of the decisions undertaken. The current study aimed to determine the possible rater behaviors in the assessment of students in higher education and to evaluate the usability of e-portfolios.

Method

Participants

The study included two participant groups: raters and students. The student group consisted of 61 individuals (male = 11, female = 50) that were enrolled in the English language teaching program at Gazi University Gazi Education Faculty and took the measurement and assessment course in the fall semester of the 2017-2018 academic year. The raters were three academicians enrolled in a doctoral program in the same faculty.

Measurement Tool

In this study, an analytic rubric developed by the researchers was used to evaluate the e-portfolios of students in higher education. When determining the criteria that constituted the rubric, the characteristics that should be possessed by an e-portfolio were taken into consideration. The identified criteria were presented to three experts, and as a result of their feedback, the final version of the criteria list was added to the measurement tool and prepared for implementation. After the expert feedback, the criteria of the relevant measurement tool were determined as follows: design/layout, originality, diversity of student work, time, self-reflection, amount of student work, and performance tasks. When evaluating the student e-portfolios, each criterion was scored using four-point grading (from 1, extremely poor to 4, very good). Field experts examining the relevant measurement tool stated that the weighting of the criteria should be considered differently. Similarly, in performance assessment studies, it is stated that the criteria or items should be weighted differently (Kondo-Brown, 2002) depending on the nature of the structure to be measured. However, there are also researchers that chose to perform equal weighting (Farrokhi, Esfandiari & Schaefer, 2012).

After the analytic rubric was prepared and applied, the process of collecting evidence was initiated to determine the validity and reliability of the related measurements. Factor analysis was used to obtain evidence of the validity of the measurements and McDonald's (1999) ω coefficient for

reliability. Before conducting an exploratory factor analysis (EFA), the assumptions of the relevant analysis should be tested. Therefore, first, it was checked and determined that the minimum number of samples was sufficient (at least five persons per variable), there were no missing or extreme values in the data set, there was a linear relationship between the criteria of the measuring instrument, and all the variables were normally distributed. After all the assumptions were met, the data set was examined in terms of factorability, and it was found to have a factorable structure (Kaiser-Meyer-Olkin value: .836 and Barlett sphericity test: statistically significant at $\chi^2(fd) = 233.337(21)$, $p = 0.000$). According to the results of the EFA, the measurement tool was found to represent a unidimensional structure (explained variance: 53.73%, factor loads: 0.652, 0.682, 0.824, 0.517, 0.762, 0.848, and 0.791 for Criterion 1 to 7, respectively). After collecting the validity evidence of the measurements obtained using the developed tool, McDonald's ω was used to evaluate the reliability of the measurements. Analyses using Mplus (version 7) revealed that McDonald's ω coefficient was .891 (95% confidence interval: .840 - .920). According to this result, it can be stated that the measurements obtained from the developed rubric for the scoring of the student e-portfolios were reliable, and there was also evidence of the validity of the inferences based on these measurements.

Data Collection

The data of the study were collected from the selected students by gathering their work throughout the measurement and assessment course using an e-portfolio system. After the students uploaded their work related to the topics covered by the curriculum into the e-portfolio system every week, the lecturer examined the students' work and gave individual feedback. The content and quantity of work that each student was expected to include in their e-portfolios throughout the semester were determined. It was explained to the students that the diversity of work they undertook was also important (e.g., video, written materials or visual materials). Then, the performance/ability of each student to prepare the e-portfolio file was scored using the developed rubric. The scores of each rater were transferred to an electronic spreadsheet program (Microsoft Excel) to obtain the data set.

Data Analysis

In the present study, a fully crossed design was used, in which all raters scored all student e-portfolio files. Data analysis was performed using the many-facet Rasch measurement model in FACETS (version 3.70.1) package program (Linacre, 2012). In this study, there were three facets: raters (R), criteria (C) and students (S). When analyzing the data, the recommendations provided by Myford and Wolfe (2003, 2004) were taken into consideration; thus, first group-level and then individual-level statistics were obtained. When the literature was examined, it is found that there are many rater errors/behaviors affecting the performance assessment process (Royal & Hecker, 2016). For example, Royal and Hecker (2016) provided a list of 30 different rater behaviors and noted that some of these behaviors were more common than others. The most frequent rater behaviors in the literature were rater strictness/leniency, halo effect, central tendency, differential strictness and

differantial leniency (Farrokhi, Esfandiari & Vaez Dalili, 2011; Myford & Wolfe, 2003, 2004). In the current study, the above-mentioned four behaviors were selected as the rater behaviors to be examined, and statistical indicators were obtained at both group and individual levels.

Results

Since the many-facet Rasch measurement model belongs to the Rasch family, it must meet the assumptions of the Rasch models (Eckes, 2015, s.124; Farrokhi, Esfandiari & Schaefer, 2012; Farrokhi, Esfandiari & Vaez Dalili, 2011), namely unidimensionality, local independence, minimum interval measurement, presence of ranking, and model-data fit. In order to determine whether the developed rubric was unidimensional, EFA was conducted. When the EFA results were examined, it was found that the measurement tool had a single-factor structure, the variance explained was 53.73%, and the factor loads of the criteria ranged from 0.517 to 0.848. The G^2 statistic developed by Chen and Thissen (1997) was used to test the local independence of the criteria in the scoring scale. According to this statistic, the estimated LD $\chi 2$ values between each pair of criteria should be below 10 and the marginal fit $\chi 2$ values should be close to zero as an indicator of local independence (Chen & Thissen 1997). The results of the local independence test for each pair of criteria are given in Table 1.

Table 1. Marginal Fit (χ^2) and LD χ^2 Values for the Partial Credit Model

Criteria	Marginal X^2	1	2	3	4	5	6
1	0.1						
2	0.1	2.1					
3	0.3	-0.6	1.2				
4	0.1	-0.1	-1.3	0.3			
5	0.4	-0.3	-0.4	2.1	1.0		
6	0.8	0.6	2.2	0.6	0.6	0.6	
7	0.3	0.8	-0.0	1.6	-0.5	1.6	1.2

The LD $\chi 2$ values of the criteria boundaries were below 10 and the marginal fit chi-square values were generally close to 0 (Table 1), suggesting that the assumption of local independence was generally provided. It was also determined that the developed rubric had minimum equal-interval and ranked grading (from 1 to 4), the related assumptions were considered to be met. Finally, the standardized residual values were examined and tested for the model-data fit. For a good model-data fit, it is suggested that the number of standardized residual values outside the ± 2 range should not exceed 5% of the total number of observations (Linacre, 2017). In the current study, the total number of observations was 61 x 7 x 3 = 1281, and the number of observations outside the range of ± 2 was 50 (3.90%). After ensuring that all the assumptions of the many-facet Rasch measurement model were satisfied, data analysis was initiated.

Strictness and Leniency Behavior

The first rater behavior that was examined in this study was strictness and leniency. For this purpose, the measurement report related to the rater facet was utilized, and it is presented in Table 2.

Table 2. Measurement Report Obtained for the Rater Facet of the Rasch Model

Rater	Logit measure	Standard error of measurement	Infit	Outfit	t-value
R2	-.32	.08	1.05	1.15	4.00
R1	-.01	.08	1.04	1.06	0.13
R3	+.33	.08	0.88	0.89	4.13
Mean	.00	.08	0.99	1.04	
Standard deviation	.33	.00	0.09	0.14	
Model, Sample: RMSE = .08 Standard deviation = .32					
Separation ratio = 3.88 Separation index = 5.50 Reliability of the separation index = .94					
Model, Fixed (all same) chi-square = 32 sd = 2 p = .00					
Model, Random (normal) chi-square = 1.9 sd = 1 p = .17					
Observed inter-rater agreement: 65.7%					
Expected inter-rater agreement: 46.3%					
Kappa statistic of inter-rater reliability: .36					

$t_{critical}$ (0.05, 2) = 4.303; RMSE = Root Mean Square Standard Error

In this study, the strictest rater was R3 (logit = 0.33), and the rater showing the highest leniency was R2 (logit = -.32) (Table 2). The infit and outfit values of the raters appeared to be acceptable (range .5 to 1.5), with the value of each rater being close to the expected value (1). The high values of separation ratio, separation index and separation index reliability indicate that the raters differed in their scoring of the students' performance. Similarly, the fixed-effects chi-square value was significant, suggesting that the raters exhibited different behaviors when scoring. The other evidence of the different scoring behaviors of the raters was the kappa statistic calculated using the inter-rater agreement values. It is reported that a kappa value below. 40 indicates poor agreement (McHugh, 2012; Sim & Wright, 2005). After determining the raters' differences in scoring through the statistical indicators at the group level, it is necessary to identify the rater or raters that cause this difference (Çetin & İlhan, 2017; İlhan, s.133, 2015; Myford & Wolfe, 2004). For this purpose, the t-statistic was calculated for each rater. When the calculated t values and the table t($t_{critical}$) values were compared, the t values were not significant. This result means that although the raters had different behaviors in evaluating the student e-portfolios, this did not have a significant effect on the overall assessment of the student performance.

Halo Effect

Another rater behavior that is highly likely to occur in performance assessment is the halo effect. In order to determine this effect, the measurement report of the criterion (or item) facet is examined as a statistical indicator at the group level (Çetin & İlhan, 2017). In this study, during this analysis, it was observed that the separation ratio was 5.13, the separation index was 7.74, the reliability of separation index was .97, and the fixed-effects chi-square value was statistically significant (χ^2 = 171.9, sd = 6, $p < .01$). According to these results, it can be stated that the difficulty levels of the criteria differed, and there was no halo effect on the scoring. In order to determine whether the raters displayed halo behavior, the infit and outfit values of the raters, which are statistical

indicators at the individual level, were examined. The measurement report of the criteria revealed that the differences in the logit values between the difficulty levels of the criteria was greater (.46 – (-1.53) = 1.99). If the fit value of a rater significantly differs from 1, that rater is considered to display halo behavior (Myford & Wolfe, 2004). In the current study, there was no rater with a fit value that significantly differed from 1; i.e., there was no halo effect on the scoring of any of the raters (Table 2). Finally, to confirm that the raters do not exhibit the halo behavior, the many-facet Rasch analysis should be repeated by equalizing the difficulties of the criteria of the measurement tool. After this process, if there is a rater that has a perfect fit value (equal to 1), it is accepted that this rater shows halo behavior (Linacre, 2012). In the current study, when the many-facet Rasch analysis was repeated with criteria having equal difficulties, it was determined that none of the raters had a perfect fit value.

Central Tendency

Another rater behavior that is likely to occur in the performance assessment process is the central tendency effect. In order to determine this behavior, the measurement report and category statistics related to the student and criterion facet were analyzed as group-level statistical indicators, and the results are presented in Table 3.

Table 3. Category statistics

Category	Frequency	Percentage	Cumulative percentage	Outfit
1	88	7	7	1.4
2	407	32	39	1.1
3	540	42	81	0.9
4	246	19	100	0.9

The raters mostly used categories 2 and 3 in their scoring, which may be an indicator of their central tendency or the students' moderate-level competence (Table 3). Since the category statistics are not sufficient to determine the central tendency behavior of the raters, the measurement results related to the student facet should also be examined. The results of the measurement report on the student facet are shown in Table 4. As it is not appropriate/possible to present the values for all students, only three students with the highest logit values and three students with the lowest logit values are given here as examples.

Table 4. Measurement Report Obtained for the Student Facet (The Data from Six Students are Presented as Examples)

Student	Logit measure	Standard error of measurement	Infit	Outfit
S10	-3.07	.43	0.78	0.83
S49	-3.07	.43	0.85	0.84
S52	-2.90	.41	1.25	1.23
S39	3.42	.45	1.24	1.10
S3	3.63	.47	1.25	1.03
S12	3.63	.47	0.80	0.69
Mean	.68	.37	1.00	1.04
Standard deviation	1.55	.03	0.40	0.54
Model, Sample: RMSE = .37 Standard deviation = 1.50				
Separation ratio =4.05 Separation index = 5.74 Reliability of separation index = .94				
Model, Fixed (all same) chi-square = 878.5 sd = 60 p = .00				
Model, Random (normal) chi-square = 56.5 sd = 59 p = .57				

The values of separation ratio, separation index and reliability of separation index were high, indicating that the students could be distinguished in a valid and reliable way according to their different competence levels and that there was no central tendency effect (Table 4). As the group-level statistical indicators for the determination of central tendency behavior, the infit and outfit values of the criterion facet were also examined. The measurement report of the criterion facet is given in Table 5.

Table 5. Measurement Report Obtained for the Criterion Facet

Criteria	Logit measure	Standard error of measurement	Infit	Outfit
C4	-1.53	.13	1.74	1.90
C6	-0.15	.12	0.75	0.74
C7	0.25	.12	0.78	0.80
C2	0.28	.12	1.06	1.07
C5	0.33	.12	0.87	0.89
C3	0.36	.12	0.72	0.72
C1	0.46	.12	1.13	1.14
Mean	.00	.12	1.01	1.04
Standard deviation	.70	.00	0.36	0.41
Model, Sample: RMSE = .12 Standard deviation = .64				
Separation ratio = 5.55 Separation index = 7.74 Reliability of separation index = .97				
Model, Fixed (all same) chi-square = 171.9 sd = 6 p = .00				
Model, Random (normal) chi-square = 5.8 sd = 5 p = .33				

Both infit and outfit values of the criteria ranged from 0.72 to 1.90 (Table 5). According to the results, C4 did not have acceptable fit values (0.50 to 1.50). When the category statistics were examined, it was determined that the raters clustered in the first category of the relevant criterion. After determining the central tendency behavior at the group level, the category statistics were also examined for each rater to identify the rater or raters that exhibited this behavior at the individual level, and it was determined that R2 displayed such behavior. According to the results, none of the raters exhibited the central tendency behavior in the scoring of any other criteria (except C4). In this context, it can be stated that the categories of the developed rubric provided a valid and reliable measurement. Another indicator for the validity and reliability of the rubric categories in

differentiating student performance is category probability curves, which are presented in Figure 1 for the data obtained from the current study.

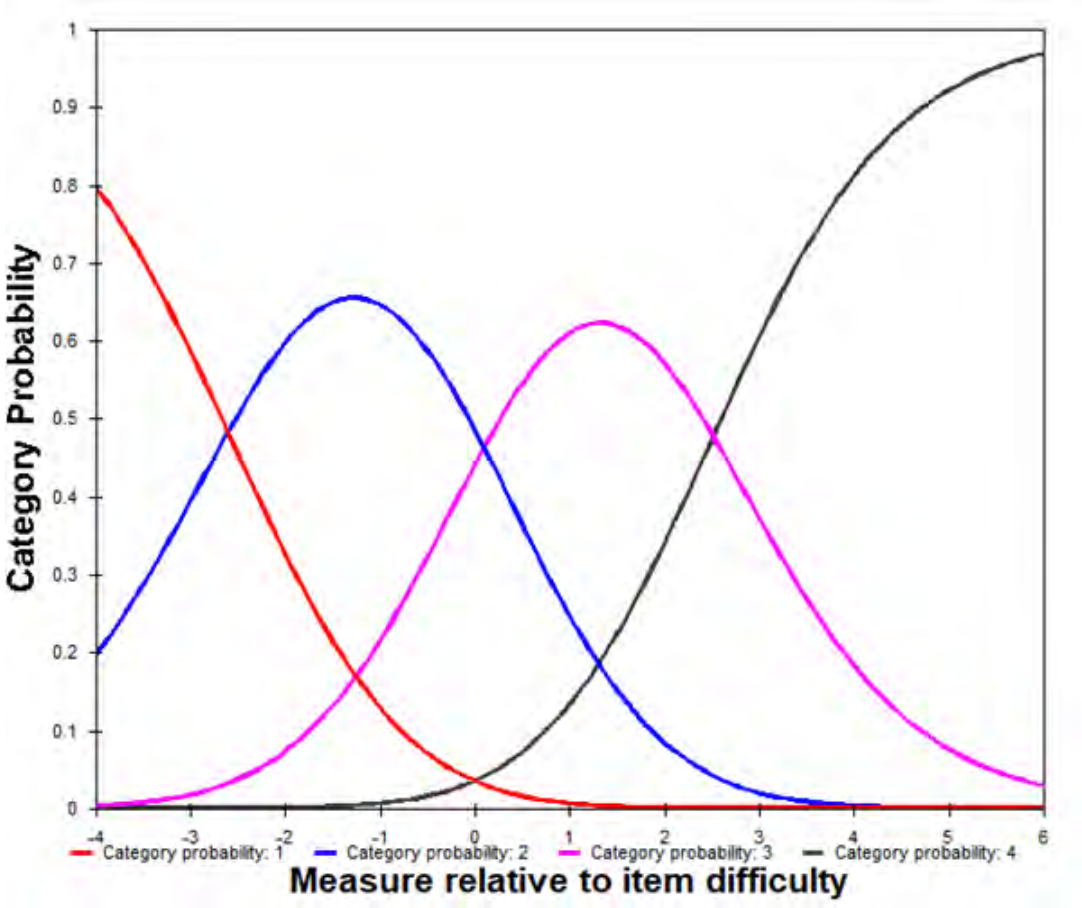

Figure 1. Category probability curves of the developed rubric.

The students with low competence levels were more likely to be in lower categories and they were less likely to be in upper categories (Figure 1). Similarly, it was more probable for the students with high ability levels to be in upper categories, and there was less probability for their presence in lower categories. These results indicate that the categories of the developed rubric were functional in distinguishing student performance.

Differantial Strictness and Leniency

Differantial strictness and leniency behavior refers to a rater's tendency to give higher scores to some students (for example, successful students) or lower grades to others (such as those with low grades or who misbehave) when assessing/scoring the student performance (Linacre, 2017; Myford & Wolfe, 2004). In this study, whether the raters evaluated all the students and all the criteria in a similar manner was investigated by analyzing the statistical indicators at the group and individual levels. First, the rater-student interactions were examined to determine how the raters behaved toward all students. When examining the bias report, a t-value is given for each rater x student interaction. According to Linacre (2017), the t-values outside the ± 2 range indicate a statistically significant interaction/bias.

When the results obtained from the current study were examined, it was found that there were seven interactions outside the ± 2 range, which are presented in Table 6.

Table 6. Statistically Significant Interactions Between the Raters and Students

Rater	Student	Observed Score	Expected Score	Bias (logit)	Standard Error	t-value
R2	S55	15.00	20.19	-1.94	.63	-3.10
R2	S58	16.00	20.52	-1.67	.62	-2.71
R2	S60	20.00	24.07	-1.62	.61	-2.66
R3	S36	12.00	15.46	-1.40	.66	-2.14
R2	S33	21.00	17.52	1.28	.62	2.09
R2	S25	23.00	19.20	1.45	.64	2.26
R2	S32	26.00	22.47	1.70	.81	2.08
Fixed chi-square = 158.4 sd = 183 p = 0.91						

The non-significance of the chi-square value given in Table 6 suggests that the raters did not show any bias behavior in the scoring process. However, when the statistical indicators at the individual level were examined, it was found that some interactions were significant; in other words, the raters behaved in a biased manner during the assessment of the student performance. For example, it was observed that R2 gave some students higher scores and other students lower scores than expected. As a result, of the three raters, one (R2) displayed bias for and against students. In this case, the validity and reliability of R2 in the assessment of student performance was considered to be lower compared to the other two raters.

In order to determine how the raters behaved according to the criteria in the process of assessing the student performance, the rater-criterion interactions were examined. The results revealed a statistically significant interaction between all the investigated cases (seven criteria x three raters = 21 interactions). It was observed that R3 displayed a more strict behavior in relation to the C4 criterion (observed score = 183, expected score = 193, t-value = -2.15). Based on the overall results, it was determined that the raters did not show any bias when assessing the student performance according to the given criteria.

Discussion, Conclusion and Recommendations

The use of e-portfolios in higher education helps students discover new things (Campbell & Schmidt, 2005), select and use effective materials in their future professions (Shaidullina et al., 2014), and make plans regarding their future (Tubaishat, 2015). It is stated that in higher education, e-portfolios are useful for monitoring students' development process (Ada, Tanberkan-Suna, Elkonca & Karakaya, 2016). However, despite the benefits of using e-portfolios in higher education, they also have certain limitations. One of these limitations is that objectivity is more difficult to assess in the process of determining the state of the student e-portfolios compared to traditional measurement tools (such as multiple-choice tests). When evaluating the e-portfolio files of the students, one of the reasons why objectivity cannot be fully ensured is that the assessment undertaken varies from rater to rater. Thus, taking into account the effect of the raters in the process of determining the student

performance will contribute to the validity and reliability of the measurements and inferences based on these measurements. Therefore, in this study, the effects of rater behaviors in the assessment of the student e-portfolios were examined in order to contribute to the validity and reliability of the inferences made in relation to the student performance.

In this study, the most frequent rater behaviors (strictness/leniency, central tendency, halo effect, and differantial rater strictness/leniency) were examined (Farrokhi, Esfandiari & Vaez Dalili, 2011; Myford & Wolfe, 2003, 2004). When the findings of the study were analyzed, it was found that in the assessment of the student e-portfolios, the raters showed strictness/leniency, centered tendency and bias behaviors; however, the halo effect was not observed. This result suggests that one or more of the rater effects/behaviors in performance evaluation interfered with the scoring of the student e-portfolios. In the current study, despite the use of both multiple raters (Gronlund, 1977, s.85; Kubiszyn & Borich, 2013, s.170) and an analytic rubric (Dunbar, Brooks & Miller, 2006; Ebel & Frisbie, 1991, s. 194; Kutlu vd., 2014, s.51; Oosterhof, 2003, s.81) to improve objectivity in performance assessment, several rater behaviors emerged during the evaluation of the student e-portfolios. Haladyna (1997, p. 137) emphasized the difficulty of maintaining consistency between raters even when rubrics were used.

In order to determine whether the rater strictness/leniency behavior had a significant effect on the assessment of the student performance, the t-values of each rater were calculated and analyzed at the statistical significance level of 0.05. According to this analysis, the t-value of none of the raters was statistically significant. This finding indicates that although the raters scored differently in terms of strictness/leniency, they did not have a significant effect on the overall assessment of the student performance.

Concerning the random strictness/leniency behavior, all raters exhibited a similar approach to the criteria in the developed analytic rubric. In other words, the rater x criterion interactions (21 in total) were not statistically significant at both group and individual levels. However, it was determined that the raters did not display the similar strictness/leniency behavior toward all students in the process of evaluating their e-portfolios, and exhibited differantial strictness and leniency behavior in favor of or against some students. Rater behaviors threaten the validity of direct measurements because they are attributed to the variance unrelated to the structure of the measurement tool (Abu Kassim, 2011; Brennan, Gao & Colton, 1995; Congdon & McQueen, 2000; Farrokhi vd., 2011). Therefore, determining the rater behaviors interfering with the performance assessment will contribute to the validity of the measurements and inferences based on these measurements. Accordingly, in the current study, identifying the rater behaviors that interfered with scoring during the evaluation process of the student e-portfolios contributed to the validity of the measurements and the decisions made based on these measurements. Considering that R2 exhibited more of the investigated rater behaviors (all

behaviors except halo) than the other two raters, it may be helpful not to include her/his scoring in the assessment of the student portfolios to increase the validity of the measurements.

According to the results of the present study, it is recommended that rater behaviors should be examined in order to ensure the validity of measurements and inferences based on these measurements during the assessment of individual performance. To increase the validity of the measurements, it is suggested that at the group level, the significant behaviors of the raters should be excluded from individual assessment, and at the individual level, if there are a sufficient number of raters, the rater displaying the significant behaviors should not be included in scoring. Considering that rater training has an effect on rater behaviors and contributes to the reliability and validity of the measurements, it should be offered to raters participating in scoring for increased validity and reliability in the performance-based assessment process. Lastly, continuous monitoring of the rater behaviors that occur in the process of evaluating the e-portfolio files of the students in higher education and providing training to minimize these behaviors can contribute to the reliability and validity of the measurements and the inferences based on these measurements.

References

Abu Kassim, N. L. (2011). Judging behaviour and rater errors: an application of the many-facet rasch model. *GEMA Online Journal of Language Studies*, 11(3), 179-197.

Ada, S., Tanberkan-Suna, H., Elkonca, F., & Karakaya, İ. (2016). Views of academicians, school administrators, and teachers regarding the use of e-portfolios in transition from elementary education to secondary education. *Educational Sciences: Theory and Practice*, 16(2), 375-397.

Bahar, M., Nartgün, Z., Durmuş, S., & Bıçak, B.(2006). *Geleneksel- Alternatif : Ölçme ve Değerlendirme Öğretmen El Kitabı*, Ankara: Pegem A Yayıncılık.

Barak, M., Ben-Chaim, D. & Zoller, U. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. ResSciEduc, 37, 353–369.

Barker, K. C. (2005). *E-portfolio for the assessment of learning*. Retrieved from http://www.futured.com/documents/FuturEdePortfolioforAssessmentWhitePaper.pdf

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6(2), 205-212.

Boddy, N., Watson, K., & Aubusson, P. (2003). A trial of the five Es: A referent model for constructivist teaching and learning. *Research in Science Education*, 33, 27–42.

Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of work key listening and writing tests. *Educational and Psychological Measurement*, 55(2), 157-176.

Campbell, M. I., & Schmidt, K. J. (2005). Polaris: An undergraduate online portfolio system that encourages personal reflection and career planning. *International Journal of Engineering*

Education, 21(5), 931–942.

Chang, C. C., Tseng, K. H., Chou, P. N., & Chen, Y. H. (2011). Reliability and validity of Web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers & Education*, 57(1), 1306-1316. doi:10.1016/j.compedu.2011.01.014

Chen, W.H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163-178.

Çetin, B., & İlhan, M. (2017). Standart ve SOLO Taksonomisine Dayalı Rubrikler ile Puanlanan Açık Uçlu Matematik Sorularında Puanlayıcı Katılığı ve Cömertliğinin İncelenmesi. *Eğitim ve Bilim*, 42(189), 217-247.

Dunbar, N.E., Brooks, C.F., & Miller, T.K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, 31(2), 115-128.

Ebel, R. L. (1965). *Measuring educational achievement*. New Jersey: Prentice- Hall Press

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. New Jersey: Prentice Hall Press.

Eckes, T. (2015). *Introduction to many-facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang Edition.

Egan, J. P. (2012). E-portfolio formative and summative assessment: Reflections and lessons learned. In *Proceedings of Informing Science & IT Education Conference* (InSITE).

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), 79-101.

Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment.1 *World Applied Sciences Journal*, 15(11), 76-83.

Gronlund, N. E. (1977). *Constructing achievement test*. New Jersey: Prentice-Hall Press.

Haladyna, T. M. (1997). *Writing test items in order to evaluate higher order thinking*. Needham Heights: Allyn & Bacon

Hung, S. T. A. (2012). A wash back study on e-portfolio assessment in an English as a foreign language teacher preparation program. *Computer Assisted Language Learning*, 25(1), 21–36.

İlhan, M. (2015). *Standart ve SOLO Taksonomisine Dayalı Rubrikler ile Puanlanan Açık Uçlu Matematik Sorularında Puanlayıcı Etkilerinin Çok Yüzeyl Rasch Modeli ile İncelenmesi*. Doktora Tezi,

Gaziantep.

Jenson, J. D. (2011). Promoting self-regulation and critical reflection through writing students' use of electronic portfolio. *International Journal of e-Portfolio*, 1(1), 49–60.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford Press.

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3-31.

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement*. New Jersey: John Wiley & Sons Incorporated.

Kutlu, Ö., Doğan, C.D. & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi/ performansa ve portfolyoya dayalı durum belirleme*. Pegem Akademi Yayınları: Ankara.

Linacre, J. M. (2012). *FACETS* (Version 3.70.1) [Computer Software]. Chicago, IL: MESA Press.

Linacre, J. M. (2017). *FACETS* (Version 3.80.0) [Computer Software]. Chicago, IL: MESA Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71.

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Lawrence Erlbaum.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276-282.

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Education Statistics.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.

Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement,* 5(2), 189-227.

Oosterhof, A. (2003). *Developing and using classroom assessments*. New Jersey: Merrill-Prentice Hall Press.

Riedler, M. & Eryaman M.Y. (2016). Complexity, Diversity and Ambiguity in Teaching and Teacher Education: Practical Wisdom, Pedagogical Fitness and Tact of Teaching. *International Journal of Progressive Education. 12*(3): 172-186

Royal, K. D., & Hecker, K. G. (2016). Rater errors in clinical performance assessments. *Journal of*

veterinary medical education, 43(1), 5-8.

Shaidullina, A. R., Fassakhova, G. R., Valeyeva, G. K., Khasanova, G. B., Komelina, V. A., & Ivanova, T. L. (2014). A comparative research on levels of students' formation skills of their career advancement portfolio in secondary and higher education systems. *Asian Social Science*, 11(1), 375–379.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3), 257-268.

Tubaishat, A. (2015). Can e-portfolio improve students' readiness to find an IT career?. *Issues in Informing Science and Information Technology,* 12, 192–203.

Watts, M., Jofili, Z., & Bezerra, R. (1997). A case for critical constructivism and critical thinking in science education. *Research in Science Education*, 27(2), 309–322.