Journal of Education and Human Development
June 2015, Vol. 4, No. 2, pp. 182-192
ISSN: 2334-296X (Print), 2334-2978 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/jehd.v4n2a22
URL: http://dx.doi.org/10.15640/jehd.v4n2a22

Comparison of Self, Peer and Instructor Assessments in the Portfolio Assessment by Using Many Facet Rasch Model

İsmail Karakaya[1]

Abstract

The purpose of this study was to compare self, peer and instructor assessments during the assessment of portfolios prepared by prospective teachers. The many facet Rasch Model was used for this purpose. Four facets were determined for the application of the many facet Rasch Model. These facets were students' achievement of criteria; prospective teachers' gender; severity/leniency of self, peer and instructor assessments; and criteria used in the portfolio assessment. The participants of the study were 74 prospective teachers who were third-year students in University and who were taking the Measurement and Evaluation course and preparing a portfolio within the scope of the course. The rubric, which was prepared by the researcher and composed of 7 criteria, was used as the data collection instrument in the research. As a result of the analysis, it was seen that the students with high achievement levels had better portfolio performances, and that the female students' performances were better than those of the male students. When self, peer and instructor assessments were compared, it was concluded that self-assessments were the most lenient while peer assessments were the severest, and that there was a statistically significant difference among the raters.

Keywords: Portfolio Assessment, Peer Assessment, Self-Assessment, Interrater reliability, Many Facet Rasch Model

1. Introduction

Portfolios have been widely used in recent years so as to determine student performances. Portfolios can be used for both the determination of student and teacher performances and the evaluation of curricula in all educational institutions from pre-school education to higher education. In Turkey, portfolios have become increasingly important in elementary education especially since 2005 because of various reasons such as he renewal of elementary and secondary curricula and the emphasis put on performance assessment in the new curricula. As is the case in elementary schools, portfolios are popular measurement tools at universities as well which are effectively used to determine student performances especially in teacher education programmes (Krause 1996; Anderson & DeMeulle 1998; Fredrick, McMahon & Shaw 2000; Gadbury-Amyot 2003). Kutlu, Doğan and Karakaya (2008) defined portfolios as a file in which learner studies are collected systematically for a specific purpose while Vavrus (1990) defined portfolios as the systematic collection of learner studies to enable teachers to monitor and control learners' knowledge, skills and aptitudes in a specific field. Portfolios can be considered as teaching and assessment tools which reveal and contribute to the change in students' characteristics such as attitudes, interests, motivation, knowledge, skills and aptitudes. Therefore, portfolios can be useful for both classroom teachers and counsellors, and can be effectively used in guidance services (Kutlu, Doğan & Karakaya, 2008). It is known that in elementary and secondary education portfolios make great contribution to the monitoring of student development, and to the acquisition of various skills such as the ability to effectively carry out self-assessment.

[1] Assoc. Prof., Gazi University, Gazi Education Faculty, Department of Measurement and Evaluation in Education, 06500 Teknikokullar- Ankara / Turkey. Email: ikarakaya@gazi.edu.tr.

Klenowski (2000) stated that, according to the result of the studies conducted with students on portfolios, portfolios have a positive effect on the development of students' a) teaching, presentation and questioning skills, b) self-assessment skills, and c) self-learning skills. In addition to these contributions, portfolios can develop students' self-esteem and organizational skills, support cooperative learning methods (Dollase 1996; Krause, 1996; Mokhtari & Yellin, 1996), and contribute to the improvement of curricula, the development of instructors' teaching skills and the maintenance of records of student development (Anderson & DeMeulle, 1998). When students actively participate in portfolio applications and prepare portfolios, they can have answers to various questions like "How do we make students prepare portfolios? How do students present portfolios in the classroom? How do we relate portfolios with the teaching process? How do students' self-reflection related to their studies in portfolios? How do students assess themselves and their peers?" This process can help students carry out portfolio applications smoothly in the classroom. These outcomes of portfolio applications can be closely related to the fact that students actively participate in both learning and assessment processes. When students participate in learning and assessment processes, they are required to organize their studies themselves and reflection on school activities and learning outcomes (Maclellan 2004). While portfolios are being prepared, assessed and presented, students can assess both their own studies and their peers' studies individually or in groups during the process. This enables them to actively participate in the learning process as well as assessment activities, and promotes the integration of teaching and assessment activities. Thus, self and peer assessments can be considered as learning materials (Lindblom-Ylänne, Pihlajamäki & Kotkas 2006). Self and peer assessments contribute to the development of students' objective self-assessment, critical thinking, academic achievement by revealing their strengths and weaknesses, and self-learning (Pierce 2003). Self and peer assessments can be used for two different purposes of product-oriented assessment and process-oriented assessment (Topping 2003). However, students can be biased when self and peer assessments are product-oriented, in other words based on grading, while more effective results are achieved when these assessments are carried out within the process so as to promote learning (Bound, 1995; Sulijsman, 2002).

Self and peer assessments can be regarded as useful applications which contribute to students' achievement of learning objectives (Orshmond, Merry & Reiling, 1997). Studies on validity and reliability of self and peer assessments show that self-assessment can result in higher grades than teacher assessment. Nevertheless, successful students can give lower grades for their performances than their teachers do while less successful students can give higher grades for their performances than their teachers do (Falchikov & Bound 1989; Lejk & Wyvill 2001). When peer assessment is compared with teacher assessment and self-assessment, it can be said that students assess their peers in a more biased and severer way than they assess themselves (Farrokhi, Esfandiari & Schaefer 2012). In order that portfolios are more effective on both kind of assessment and learning skills of students, and that the desired objectives are achieved, rating scales should be prepared and used appropriately, and raters should be trained properly. Supovitz, Macgowan and Slattery (1997) stated that a high correlation is required among the scores of raters so that these scores are highly valid and reliable. Interrater reliability can be measured though various methods in performance-based assessments such as portfolio assessment. This study was aimed at comparing self, peer and instructor assessments in terms of interrater reliability during the portfolio assessment by using the Many Facet Rasch Model. Interrater reliability can be determined through various methods such as Fleiss' Kappa, Cohen's Kappa, Pearson Product-Moment Correlation Coefficient, Spearman's Rank Correlation Coefficient, Cronbach's Alpha, and Intraclass Correlation Coefficient (Jonsson & Svingby 2007; Multon 2010). No method can be considered the best to measure interrater reliability. Each method has some strengths and weaknesses. What is important is to use the most appropriate method in accordance with the purpose of the study. The Many Facet Rasch Model can be used when there is more than one rater; it is required to determine rater severity/leniency; it is required to determine results for each participant separately; and it is required to determine rater performances for each item separately (Multon 2010). Thus, in this study the Many Facet Rasch Model was used with the aim of comparing self, peer and instructor assessments in the course of the assessment of portfolios prepared by students

2. Methods

2.1. Participants

The participants of the study were students from the Undergraduate Programme for Classroom Teacher Education in the Department of Elementary Education, Faculty of Education at Ondokuz Mayıs University in the fall

semester of the academic year 2011-2012, and who were third-year students taking the Measurement and Evaluation course, and preparing portfolios within the scope of the course. The participants were composed of 74 students in total divided into two classrooms, of whom 16 are male and 58 are female.

2.2. Collected Data

In the first two weeks of the semester, the participants were informed about portfolios in general, portfolio preparation in the process, and portfolio assessment. A rubric was used for assessing any study in student portfolios. The validity and reliability analyses of the rubric were carried out through the data obtained from the students who were attending the programmes in the Department of Mathematics Education and the Department of Turkish Language Education at Ondokuz Mayıs University in the spring semester of the academic year 2010-2011. The content validity of the measurement tool was tried to be ensured through the opinions of two measurement and evaluation experts. Within the scope of the Measurement and Evaluation course, preparing portfolios, the students developed various materials such as achievement tests and performance tasks, assessed studies weekly and took examinations about the content. During material development and weekly assessments, the students were informed about how they should assess themselves and what they should take into account during peer assessment. The students assessed their portfolios and their peers' portfolios at the end of the semester by he rubric which was developed by the researcher. During peer assessment, so as to minimize the number of systematic errors, it was made sure that each portfolio was assessed by a prospective teacher from the other classroom, and that personal details on the first page of portfolios were removed. The rubric which was used in the portfolio assessment included such criteria as organization, authenticity, critical thinking, punctuality, number and variety of studies, self-assessment and completeness of studies in terms of content.

2.3. Data Analysis

The FACETS is an ideal programme to measure the interrater reliability of studies which include more than one rater like students' performances, portfolios, performance tasks and open-ended questions (Linacre, 2012). The MINIFAC programme which is the student version of the FACETS 3.70 programme was used in this study since multiple raters were included in portfolio applications and their scores were compared. Four facets were determined for the application of the many facet Rasch Model. These facets were students' achievement of criteria; students' gender; severity/leniency of self, peer and instructor assessments; and criteria used in the portfolio assessment. During data analysis, the minimum value of 0.6 and the maximum value of 1.4 were used as the infit and outfit indices. In order that the data used in the analysis is consistent with the model, it is required that less than approximately 5% of the data should be larger than or equal to 2 in absolute value, or less than 1% of the data should be larger than or equal to 3 (Linacre, 2003). Thus, it was determined whether the standard scores (z scores) of the scores given by students hemselves, by peers and by the instructor according to the seven criteria were larger than or equal to +/- 3. six students ve teachers with extreme values were removed from the study, and the analyses were conducted with 74 students.

3. Findings

Prior to the analysis, the students were ranked in descending order of success according to the means of academic achievement. In other words, the prospective teacher in the first rank was the most successful while the one in the 74th rank was the least successful. The main purpose was to compare the students' means of academic achievement and portfolio performances. As it can be seen Figure 1, four facets were determined in the data analysis. These were, respectively, students, gender, kinds of raters and rubric criteria used in the portfolio assessment.

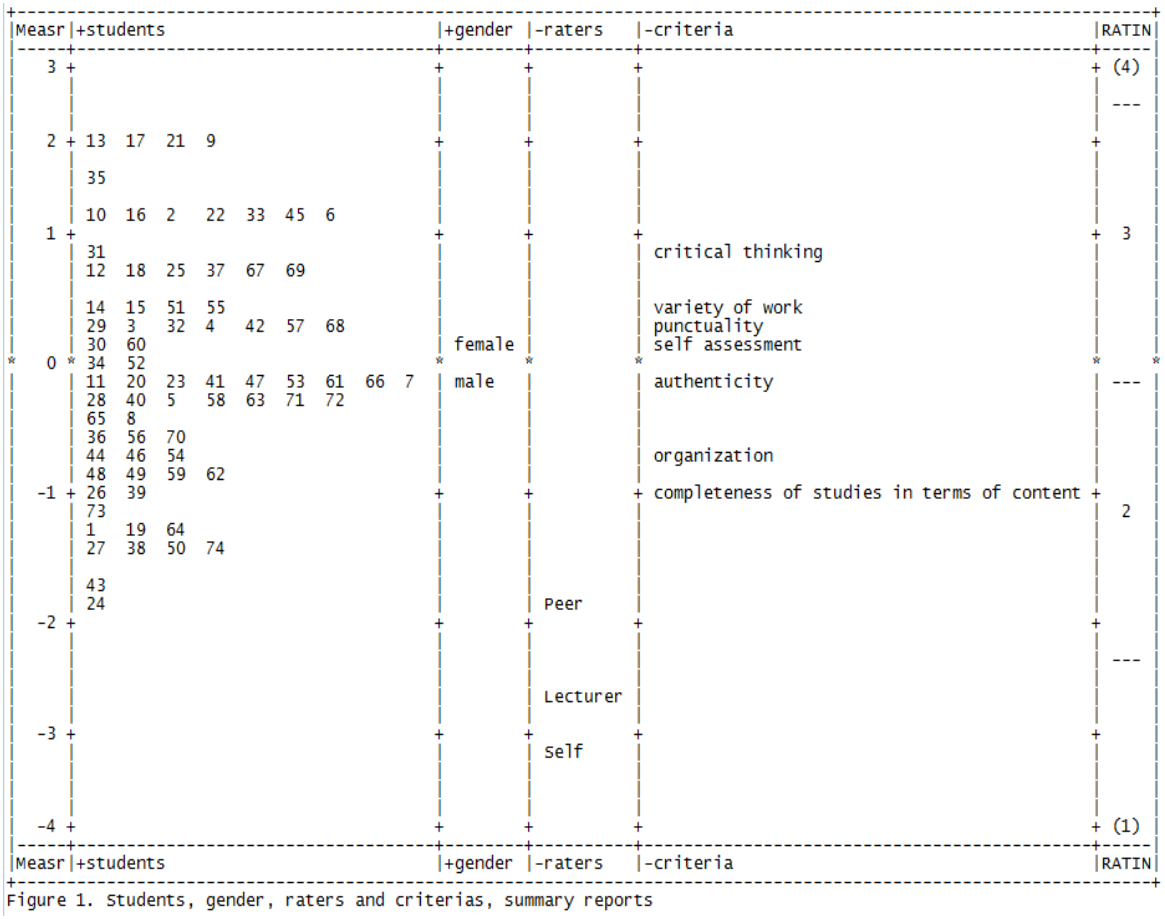

Figure 1. Students, gender, raters and criterias, summary reports

Figure 1: Students, Gender, Kind of Raters and Criterias, Summary Reports

When the first column for students is examined, it is seen that 13, 17, 21, 9, 35, 10, 16, 2, 22 and 33, who rank high according to the means of academic achievement, are the most successful students according to self, peer and instructor assessments. Among the least successful students, on the other hand, are 73, 1, 19, 64, 27, 38, 50, 74, 43 and 24. It is interesting that the prospective teacher, who is at the top of the ranking according to the means of academic achievement, is among the least successful students according to self, peer and instructor assessments. As for gender of students, it can be said that female students are more successful than male students. When the third column for raters is examined, it is seen that peer assessment ranks first, instructor assessment ranks second and self-assessment ranks third. In other words, self-assessments of the students are the most lenient while peer assessments are the severest. The students gave themselves high grades, and they gave their peers lower grades. It can be said that instructor assessments and self-assessments are closer than instructor assessments and peer assessments are. When the students' achievements are compared according to the criteria used in the portfolio assessment, it is seen that they are most successful in terms of completeness of studies (content of studies) and organization, and least successful in terms of critical thinking, variety of studies, punctuality and self-assessment.

3.1. Analyses of Students' Gender

Gender of students was determined as the second facet. As can be seen Table 1, The RMSE value of the gender variable is 0.09.

Table 1: Students' Gender Measurement Reports

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Model Measure	S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	N Gender
4728	1344	3.52	3.58	.14	.05	1.00	.0	.97	-.6	2 Female
713	210	3.40	3.50	-.14	.12	1.04	.4	1.10	.9	1 Male
2720.5	777.0	3.46	3.54	.00	.09	1.02	2	1.03	2	Mean (Count:2)
2007.5	567.0	.06	.04	14	.04	.02	2	.07	8	S.D (population)
2839.0	801.9	.09	.06	.19	.05	.03	3	.09	1.1	S.D. (Sample)
Model, Populn: RMSE .09 Adj (True) S.D. .10 Separation 1.01 Strata 1.69 Reliability .51										
Model, Sample: RMSE .09 Adj (True) S.D. .17 Separation 1.75 Strata 2.66 Reliability .75										
Model, Fixed (all same) chi-square: 4.1 d.f.: 1 significance (probability): .04										

The reliability coefficient is 0.75. The separation index is 0.75. The chi-squared test was used to determine whether the students' performances according to the rubric criteria differed significantly in terms of gender. The result ($\chi 2 = 4.1$, sd = 1, p = 0.00) shows that the students' performance according to the rubric criteria differs significantly in terms of gender. When the infit and outfit values are examined, it is seen that the values of female and male students range between 0.6 (minimum) and 1.4 (maximum) the expected values.

3.2. Students' Achievement According to Rubric Criteria

The detailed information about the students' achievement according to the criteria used in the portfolio assessment is shown in Table 2. The students' performances were ranked in descending order of success. According to the data, 13, 17, 21, 9 and 35 are the most successful according to the criteria in general. 24, 43, 50, 74 and 38, on the other hand, are the least successful. The most and the least successful students are all composed of only female students. The RMSE value refers to the standard error of the collected data excluding the students with extreme values. When the RMSE value is 0.45, the standard error can be considered low. The reliability coefficient obtained from the Rasch analysis refers to the reliability of the assessment of students' performances according to the rubric criteria. The reliability coefficient of 0.76 shows that the reliability of the portfolio assessment is moderate. With the separation index of 1.76 and the reliability coefficient of 0.76, the fixed effect chi-squared test is used to determine whether the students' performances according to the rubric criteria differ significantly. It can be said that the students' performances according to the rubric criteria ($\chi 2 = 285.7$, sd = 73, p = 0.00) differ significantly. As for the infit values, none of the students have an infit value below the minimum value of 0.6, and 2, 6, 34 and 43 have infit values above the maximum value of 1.4.

Table 2: Students's Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Model Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu Students
82	21	3.90	3.91	2.03	.75	1.03	.2	1.09	.3	9
82	21	3.90	3.91	2.03	.75	.99	.1	.74	.0	13
82	21	3.90	3.91	2.03	.75	.91	.0	1.25	.5	17
82	21	3.90	3.91	2.03	.75	1.10	.3	1.16	.4	21
81	21	3.86	3.87	1.56	.63	.80	-.2	.57	-.5	35
80	21	3.81	3.82	1.21	.56	1.45	1.0	.99	.1	2
80	21	3.81	3.82	1.21	.56	1.48	1.1	1.11	.3	6
80	21	3.81	3.82	1.21	.56	.80	-.3	.66	-.5	10
80	21	3.81	3.82	1.21	.56	.83	-.3	.67	-.4	22
80	21	3.81	3.82	1.21	.56	.83	-.2	.68	-.4	33
80	21	3.81	3.82	1.21	.56	1.09	.3	1.58	1.0	45
79	21	3.76	3.82	1.19	.51	1.26	.7	1.67	1.3	16
79	21	3.76	3.77	.92	.51	1.06	2	.97	.0	31
78	21	3.71	3.72	.67	.48	.96	.0	.83	-.2	12
78	21	3.71	3.72	.67	.48	.88	-.2	.72	-.6	18
78	21	3.71	3.72	.67	.48	.90	-.1	.75	-.5	25
78	21	3.71	3.72	.67	.48	1.02	.1	.98	.1	37
78	21	3.71	3.72	.67	.48	.99	.0	.88	-.1	67
78	21	3.71	3.72	.67	.48	.79	-.5	.77	-.4	69
77	21	3.67	3.67	.45	.46	1.18	.6	1.64	1.5	14
77	21	3.67	3.67	.45	.46	.96	.0	.94	.0	15
77	21	3.67	3.67	.45	.46	.98	.0	.89	-.1	51
77	21	3.67	3.67	.45	.46	1.23	.7	1.35	.9	55
76	21	3.62	3.62	.25	.44	.89	-.2	.76	-.6	3
76	21	3.62	3.62	.25	.44	1.51	1.5	1.44	1.2	4
76	21	3.62	3.62	.25	.44	1.47	1.4	1.35	1.0	29
76	21	3.62	3.62	.25	.44	.96	.0	.82	-.4	32
76	21	3.62	3.62	.25	.44	.89	-.2	.74	-.7	42
76	21	3.62	3.62	.25	.44	.79	-.6	.77	-.6	57
76	21	3.62	3.62	.25	.44	.80	-.5	.90	-.1	68
74	21	3.52	3.59	.17	.41	.60	-1.4	.59	-1.4	30
74	21	3.52	3.59	17	.41	1.13	.5	1.21	.7	60
75	21	3.57	3.56	.07	.42	1.49	1.5	1.39	1.1	34
73	21	3.48	3.55	.00	.40	.75	-.8	.72	-.9	52
74	21	3.52	3.51	-.10	.41	1.11	.4	1.00	.0	7
74	21	3.52	3.51	-.10	.41	1.37	1.2	1.29	.9	11
74	21	3.52	3.51	-.10	.41	.75	-.8	.69	-1.0	20
74	21	3.52	3.51	-.10	.41	1.01	.1	.89	-.2	23
74	21	3.52	3.51	-.10	.41	.77	-.7	.73	-.8	41

Table 2: (Continued)

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Model Measure	S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Nu Students
74	21	3.52	3.51	-.10	.41	.83	-.5	.75	-.7	47
74	21	3.52	3.51	-.10	.41	.95	.0	.90	-.2	53
74	21	3.52	3.51	-.10	.41	.85	-.4	.83	-.4	61
72	21	3.43	3.50	-.15	.39	1.17	.6	1.15	.7	66
73	21	3.48	3.46	-.27	.40	1.17	.6	1.52	.5	5
73	21	3.48	3.46	-.27	.40	1.78	2.2	.86	1.5	28
73	21	3.48	3.46	-.27	.40	.76	-.7	1.14	-.3	40
73	21	3.48	3.46	-.27	.40	1.07	.3	1.01	.5	63
73	21	3.48	3.46	-.27	.40	1.00	.1	1.31	.1	71
73	21	3.48	3.46	-.27	.40	1.21	.7	1.31	1.0	72
71	21	3.38	3.45	-.30	.38	1.44	1.4	1.31	1.0	58
72	21	3.43	3.41	-.42	.39	.89	-.2	.84	-.4	8
70	21	3.33	3.40	-.44	.38	1.07	.3	1.04	.2	65
71	21	3.38	3.36	-.57	.38	.77	-.7	.78	-.7	36
71	21	3.38	3.36	-.57	.38	.61	-1.4	.60	-1.5	56
69	21	3.29	3.35	-.58	.37	1.31	1.0	1.50	1.5	70
70	21	3.33	3.31	-.71	.38	1.40	1.3	1.37	1.2	44
70	21	3.33	3.31	-.71	.38	.80	-.6	.80	-.6	46
68	21	3.24	3.31	-.72	.36	.90	-.2	.87	-.3	54
69	21	3.29	3.26	-.85	.37	.84	-.4	.84	-.4	48
69	21	3.29	3.26	-.85	.37	.79	.6	.77	-.7	49
69	21	3.29	3.26	-.85	.37	1.00	0	.95	.0	59
69	21	3.29	3.26	-.85	.37	.65	-1.2	.67	-1.2	62
68	21	3.24	3.21	-.99	.36	.93	-.1	.97	.0	26
68	21	3.24	3.21	-.99	.36	1.21	.7	1.18	.6	39
67	21	3.19	3.16	-1.12	.36	.82	-.5	.81	-.5	73
66	21	3.14	3.11	-1.25	.36	.73	-.8	.73	-.8	1
66	21	3.14	3.11	-1.25	.36	.75	-.7	.76	-.7	19
63	21	3.00	3.07	-1.34	.35	.86	-.3	.88	-.3	64
65	21	3.10	3.06	-1.37	.35	1.14	.5	1.13	.5	27
65	21	3.10	3.06	-1.37	.35	.99	.0	.97	.0	38
65	21	3.10	3.06	-1.37	.35	.87	-.3	.88	-.3	74
64	21	3.05	3.01	-1.49	.35	.36	-2.7	.38	-2.6	50
62	21	2.95	2.91	-1.73	.34	1.67	1.9	1.58	1.7	43
61	21	2.90	2.87	-1.85	.34	1.38	1.2	1.41	1.3	24

RMSE (Model) .45 Adj (True) S.D. .81 Separation 1.80 Strata 2.73 Reliability .76
Fixed (all same) chi-square: 285.7 d.f.: 73 significance (probability): .00
Random (normal) chi-square: 57.8 d.f.: 72 significance (probability): .89

The expected outfit values range between 0.6 and 1.4 as well (Wright & Linacre 1994). The expected values of 0.7 minimum and 1.3 maximum can be accepted for infit and outfit values in some cases (Multon 2010). In that respect, it can be said that 35, 30 and 50 have outfit values below 0.6 and 16, 14, 4, 70, 28 and 43 have outfit values above 1.4. When infit and outfit values are considered, it can be said that only 12 students' infit and outfit values of the rubric criteria used in portfolio assessment of do not range between the expected quality-control values, and they do not perform appropriately.

3.3. Analysis of Raters

Table 3. demonstrates results of the analysis related to the students' performances according to the rubric criteria in portfolio assessment.

Table 3: Raters' Measurement Reports

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Model Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N Raters
1718	518	3.32	3.33	-1.90	.08	1.22	3.02	1.19	2.7	3 Peer
1835	518	3.54	3.58	-2.66	.09	.71	-5.0	.68	-4.5	1 Lecturer
1888	518	3.64	3.69	-3.08	.09	1.05	.7	1.08	.8	2 Self
1813.7	518.0	3.50	3.53	-2.55	.09	.99	-.3	.98	-.3	Mean (Count:3)
71.0	.0	.14	.15	.49	.01	.21	3.5	.22	3.1	S.D (population)
87.0	.0	.17	.18	.60	.01	.26	4.3	.27	3.8	S.D. (Sample)

RMSE (Model) .09 Adj (True) S.D. .59 Separation 6.96 Strata 9.61 Reliability .98
Fixed (all same) chi-square: 103.8 d.f.: 2 significance (probability): .00
Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

Results of the assessment in term of rubric criteria used in portfolio assessment are shown in Table 3. At that stage, 74 students, who were the participants of the study, assessed themselves. 74 students carried out peer assessment as well as self-assessment. The raters were divided into three groups including students, peers and the course instructor. Self and peer assessments were analyzed in groups, not individually. Thus, three kinds of raters including students, peers and the instructor were used during the analysis. The raters were ranked from the severest to the most lenient. The most lenient raters were students which were followed by the instructor, and the severest raters were peers. The RMSE provides data on measurement errors. Since the RMSE value of rater severity/leniency is 0.09, the standard error is low. In addition, the adjusted standard error is below the critical value of 1.00. Since the reliability coefficient related to the raters' scoring behaviours is 0.97, the scores given by raters in different groups are highly reliable. The separation index is 5.65 and the chi-squared test is used to determine whether the raters' scoring behaviours differ significantly. The analysis result shows that self, peer and instructor assessments differ significantly ($\chi 2 = 103.8$, $sd = 2$, $p = 0.00$). When the infit and outfit statistics of rater performances in self, peer and instructor assessments are examined, it is seen that these statistics ranged between the expected values. This means that the scores of the three different groups were at different levels in the assessment of portfolios prepared by the students according to the rubric criteria.

3.4. Analysis of Criteria Used in Portfolio Assessment

Table 4. shows the analysis results related to the criteria used in the rubric prepared for self, peer and instructor assessments of the students' portfolios.

Table 4: Item/Criterion Statistics of the Rubric used in Portfolio Assessment

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Model Measure	S.E	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N Criteria
719	222	3.24	3.24	.89	.11	.93	-.7	.99	.0	(6) critical thinking
754	222	3.40	3.41	.41	.12	1.30	2.9	1.28	2.5	(3) variety of work
767	222	3.45	3.48	.22	.12	.92	-.8	-.88	-1.1	(4) punctuality
770	222	3.47	3.49	.17	.12	.99	.0	1.00	.0	(5) self assessment
785	222	3.54	3.56	-.07	.13	.85	-1.5	.91	-.7	(2) authenticity
817	222	3.68	3.72	-.67	.15	.94	-.5	.92	-.5	(1) organization
829	222	3.73	3.77	-.95	.16	1.11	1.0	.92	-.4	(7) completeness of studies in terms of content
777.3	222	3.50	3.53	.00	.13	1.01	.0	.98	-.1	Mean (Count:7)
34.7	.0	.16	.17	.58	.01	.14	1.4	.13	1.1	S.D (population)
37.4	.0	.17	.18	.63	.01	.15	1.5	.14	1.2	S.D. (Sample)

RMSE (Model) .13 Adj (True) S.D. .62 Separation 4.70 Strata 6.60 Reliability .96
Fixed (all same) chi-square: 131.1 d.f.: 6 significance (probability): .00
Random (normal) chi-square: 5.7 d.f.: 5 significance (probability): .33

As is seen in Table 4, the students are less successful in meeting the criteria of critical thinking, variety of studies, punctuality, and effective and accurate self-assessment. The students are more successful in meeting the criteria of completeness of studies, and organization of the dossier and the studies. Since the standard error of the criterion analysis (RMSE) is 0.13, the standard error related to the assessment quality is low. The adjusted standard deviation value determined based on this standard error is measured to be below 1.00. The reliability coefficient of the rubric criteria used in the assessment of the students' portfolios is 0.96. This finding shows that the reliability of the criteria used in determining student performances in portfolio assessment is high. It can be said that the rubric criteria are suitable to determine the quality of student studies, and each criteria measures a different feature since the separation index of the criterion analysis is 4.70, the reliability coefficient is 0.96, and the results of the fixed effect chi-squared test are statistically significant. In terms of the infit and outfit values of the rubric criteria, none of the criteria exceeds the critical value. This finding means that the rubric criteria can be used in portfolio assessment.

4. Discussions and Conclusions

The purpose of the study was to investigate self, peer and instructor assessments through the assessment of portfolios used in teacher education by using the Many Facet Rasch Model. For this purpose, the portfolios prepared by the students within the scope of the Measurement and Evaluation course were assessed by the rubrics prepared by the researcher. Instructor, self and peer assessments were carried out during the assessment of portfolios. The research was also aimed comparing self and peer assessments with instructor assessment. In this context, students, gender, raters and assessment criteria were determined as the facets of the study. At the stage of data collection, the students were ranked in descending order of success according to the means of academic achievement. While the student code 1 was used for the most successful student, the student code 74 was used for the least successful student. When the students' performances according to the criteria of portfolio assessment and their rankings according to the means of academic achievement were examined, it was seen that students who were ranking high could perform poorly in terms of achieving the rubric criteria. In other words, the student, who ranked first according to the means of academic achievement, was among the least successful students in the portfolio assessment. On the other hand, it can be said that in general there is a partial correlation between the academic achievements and the rankings of portfolio assessment. According to the rater assessments, female students were more successful than male students. This finding is in line with the findings of many researches in the literature (Wainer and Steinberg 1992; Leonard and Jiang 1999). According to the rater assessments, the students assessed themselves in a more lenient way than peers and the instructor do. This finding is in line with the findings of some researches in the literature (Boud and Falchikov 1989; Lejk & Wyvill 2001; Topping 2003; Farrokhi, Esfandiari & Dalili 2011; Farrokhi, Esfandiari & Schaefer 2012). As for peer and instructor assessments, it can be said that peer assessments were severer than instructor assessments. This finding is contrary to some researches in the literature (Farrokhi, Esfandiari & Dalili 2011; Farrokhi, Esfandiari & Schaefer 2012). These researches reveal that instructor assessments are severer than peer assessments. It can be said that in this research either the instructor assessed students more severely or the students assessed their peers more severely.

The chi-squared test was used to determine whether the raters' scoring behaviors differed significantly in self, peer and instructor assessments, and it was revealed that there was a significant difference. This result shows that self, peer and instructor assessments differed statistically. This can adversely affect the reliability of the assessments. Since peer assessments were the severest and self-assessments were the most lenient, the students can be provided with trainings related to self and peer assessment. This is because, as Eckes (2009) stated, the interrater reliability increases when raters are trained. As for the analysis results related to the rubric criteria used to assess the students' performances in the portfolio preparation process, the students were most successful in terms of completeness of studies, and organization of file or folder and the studies. The students were least successful in terms of critical thinking, variety of studies and self-assessment, respectively. In addition, according to the analysis results, the rubric criteria used in portfolio assessment differed statistically. Since the students assessed themselves most leniently according to the analysis of students' performances and the analysis of raters, it can be said that the students could not acquire self-assessment skills adequately. The Rasch Model provides a reliability result which is equal to the Cronbach Alpha coefficient. In other words, the Rasch Model shows statistically how reliable it is in separating the students' performances according to quality, criteria difficulty, and rater severity/leniency (Akın & Baştürk, 2012).

In this research, the reliability coefficient is 0.76 for ranking of the students' by academic achievement; 0.98 for rater severity/leniency in self, peer and instructor assessments; and 0.75 for comparison of the students' performance by gender.

Particularly the fact that the reliability coefficient of self, peer and instructor assessments is high can be commented that it provides significant information about self and peer assessment in portfolio applications at universities. Although the reliability coefficient of scores given by different raters is high, self, peer and instructor assessments differ significantly. Against this background, it can be recommended that students are provided with trainings in self and peer assessment through rubrics, or more importance can be attached to such applications in classes. Sluijsman and Moerkerke (1999) stated that trainings in self and peer assessment can contribute to the acquisition of high level performances. Various suggestions can be offered directly or indirectly based on the research findings. In performance-based studies in higher education, research can be made on the interrater reliability of self and peer assessments of students. In addition, various studies on self and peer assessment can be conducted by use of Many Facets Rasch Model by taking into account various variables such as academic achievement and gender of prospective teachers/students in terms of portfolios and performance-based activities such as performance tasks especially in elementary or secondary schools.

References

Akın, Ö & Baştürk, R. (2012). Keman eğitiminde temel becerilerin Rasch ölçme modeli ile değerlendirilmesi, Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 31, 175-187.

Anderson, R. S., & DeMeulle, L. (1998). Portfolio use in twenty-four teacher education programs. Teacher Education Quarterly, 25 (1), 23-31.

Boud, D. (1995). Enhancing learning through self-assessment. London: Kogan Page.

Dollase, R.H. (1996). The Vermont experiment in state-mandated portfolio program approval. Journal of Teacher Education, 47 (2), 85-100.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Falchikov, N. & Boud, D. (1989). Student self-assessment in higher education: A Meta-analysis, Review of Educational Research, 59 (3), 395–430.

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 (Innovation and Pedagogy for Lifelong Learning): 70-77.

Farrokhi, F., Esfandiari, R. & Schaefer, E. (2012). A Many-Facet Rasch Measurement of differential rater severity/leniency in sel assessment, peer assessment, and teacher assessment. Journal of Basic and Applied Scientific Research, 2 (9), 8786-8798.

Fredrick, L., McMahon, R., & Shaw, E.L. (2000). Preservice teacher portfolios as autobiographies. Education, 120 (4), 634-640.

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

Klenowski, V. (2000). Portfolios: promoting teaching. Assessment in Education, 7 (2), 215-236.

Krause, S. (1996). Portfolios in teacher education: Effects of instruction on preservice teachers' early comprehension of the portfolio process. Journal of Teacher Education, 47 (2), 130-138.

Kutlu, Ö., Doğan, D. & Karakaya, İ. (2010). Performansa ve portfolyoya dayalı durum belirleme. Ankara: Pegem Akademi.

Leonard, D. K. and Jiang, J. (1999) Gender bias and the college predictors of the SATs: A cry of Despair, Research in Higher Education, 40, 375-407.

Lejk, M. & Wyvill, M. (2001). The Effect of the inclusion of self-assessment with peer-assessment of contributions to a group project: A Quantitative study of secret and agreed assessments, Assessment and Evaluation in Higher Education, 26 (6), 551–61.

Lindblom-Ylänne, S., Pihlajamäki, H. & Kotkas, T. (2006). Self, peer and teacher assessment of student essays. Active Learning in Higher Education, 7 (1), 51-62.

Maclellan, E. (2004) 'How convincing is alternative assessment for use in higher education? Assessment and Evaluation in Higher Education, 29 (3), 311–21.

Mokhtari, D., & Yellin, D. (1996). Portfolio assessment in teacher education: Impact on preservice teachers' knowledge and attitudes. Journal of Teacher Education, 47 (4), 245-248.

Multon, K. (2010). Interrater reliability. In N. Salkind (Ed.), Encyclopedia of research design. (pp. 627-629). Thousand Oaks, CA: SAGE Publications.

Orsmond, P. , Merry, S. & Reiling, K. (1997). A Study in self-assessment: Tutor and students perceptions of performance criteria, Assessment and Evaluation in Higher Education, 22 (4), 357–369.

Supovitz, J. A., MacGowin, A., and Slattery, J. (1997). Assessing agreement: An examination of interrater reliability of portfolio assessment in Rochester, New York. Educational Assessment, 4 (3), 237-259.

Sluijsman, D. & Moerkerke, G. (1999). Student involvement in performance assessment: A research Project. Europen Journal of Open and Distance Learning.
http://www.eurodl.org/materials/contrib/1999/assessment_issue/sluijsmans/

Sluijsmans, D. (2002). Establishing learning effects with integrated peer assessment tasks. Retrieved June 19, 2013,from
http://www.heacademy.ac.uk/assets/documents/resources
/database/id437_establishing_learning_effects.pdf

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility, Optimising New Modes of Assessment: In Search of Qualities and Standards Innovation and Change in Professional Education. 1, 55-87.

Vavrus, L. (1990). Put portfolios to the Test. Instructor. 100 (1), 48-53.

Wainer, H. and Steinberg, L. S. (1992) Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A Bidirectional validity study, Harvard Educational Review, 62, 323-336.

Wright, B.D. and Linacre, J.M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8 (30), 370.