International Journal of Assessment Tools in Education
2019, Vol. 6, No. 2, 193–201
https://dx.doi.org/10.21449/ijate.533517
Published at http://www.ijate.net http://dergipark.org.tr/ijate *Research Article*

An Empirical Study for the Statistical Adjustment of Rater Bias

Mustafa İlhan *

[1] Dicle University, Ziya Gokalp Education Faculty, Department of Mathematics and Science Education, Diyarbakir, Turkey

ARTICLE HISTORY

Received: 28 February 2019
Revised: 24 April 2019
Accepted: 30 April 2019

KEYWORDS

Bias adjustment,
Rater bias,
Many facet Rasch model

Abstract: This study investigated the effectiveness of statistical adjustments applied to rater bias in many-facet Rasch analysis. Some changes were first made in the dataset that did not include *rater* × *examinee* bias to cause to have *rater* × *examinee* bias. Later, bias adjustment was applied to rater bias included in the data file, and the effectiveness of the statistical adjustment was further examined. The outcomes pertaining to the datasets with and without bias, and to which the bias adjustment was applied, were compared. It was concluded that diversities created by *rater* × *examinee* bias in examinees' ability estimation, item difficulty indices and measures of rater severity and leniency were, to a large extent, eliminated by bias adjustment. This result indicates that the bias adjustment using many-facet Rasch analysis is a viable way to control rater bias.

1. INTRODUCTION

The tests used in education and psychology are categorized as objective tests and subjective tests by the type of scoring (McNamara, Erlandson, & McNamara, 2013). Objective tests consist of the items based on selecting a correct answer from the options provided, such as multiple-choice, true-false, and matching questions (Haladyana, 1997). Scores on objective test do not vary according to the rater, which means that objective tests have higher rater reliability. These tests can be rated easily and quickly, and so they are budget-friendly (Bennett, Ward, Rock, & LaHart, 1990). Subjective tests, on the other hand, use the items that require students to construct their responses, such as open-ended questions. Subjective tests scores tend to vary according to the rater (Bennett, 1991). For this reason, in subjective tests raters are one of the variability sources that affect students' test scores (Eckes, 2005). Rater-based factors are undesired and systematic rater behaviors that lead to the inclusion of variance irrelevant to the construct being measured to students' test scores, and are known as rater effect (Eckes, 2005; Hoyt, 2000). Rater effect includes rater severity and leniency, halo effect, central tendency

CONTACT: Mustafa İLHAN ✉ mustafailhan21@gmail.com Dicle University, Ziya Gokalp Education Faculty, Department of Mathematics and Science Education, Diyarbakir, Turkey

ISSN-e: 2148-7456 /

effect and range restriction (Saal, Downey, & Lahey, 1980). Bias is also a form of rater effect (Myford & Wolfe, 2004).

1.1. Rater Bias (Differential Rater Severity/Leniency)

Bias is raters' unexpectedly severe or lenient scoring regarding an aspect of the assessment process (Knoch, Read & von Randow, 2007). Rater bias can be related to examinees (*rater x examinee*), items (*rater x item*) or both (*rater x examinee x item*). "*Rater x examinee"* bias refers to raters' tendency to give higher or lower scores based on students' prior performances or demographics such as gender, age, and cultural factors (Aubin, St-Onge, & Renaud; 2018; Kumar, 2005). "*Rater x item"* bias refers to whether raters grade all the items on a test with the same severity or leniency (Haiyang, 2010). "*Rater x examinee x item"* bias refers to raters' assignment of lower or higher scores than expected to some students for their performance on some items.

In order to avoid rater bias, rater training (Knoch, Read, & von Randow, 2007; Fahim & Bijani, 2011) and blind scoring using rubrics have been suggested (Hogan & Murphy, 2007). Studies, however, have shown that rater bias can persist despite these precautions. For example, Kondo Brown (2002) investigated whether teachers who received rater training are biased towards some candidates or certain criteria while evaluating university students' Japanese second language writing ability. The performance of 234 university students was graded by three raters using an analytic rubric. The study results showed significant interactions between raters and students, and rater and rating criteria that indicated bias. In a different study by İlhan (2015), 104 students' responses to eight open-ended mathematics questions were graded by seven raters. Despite the training provided to the raters and using a rubric for the scoring, rater bias was not entirely eliminated. Knoch, Read and von Randow (2007) compared the effectiveness of face-to-face and online rater training. They also concluded that rater training cannot completely prevent bias in scoring.

1.2. Statistical Adjustment of Rater Bias

The fact that the rater bias persists in spite of using rubrics, blind scoring and rater training brings up the question of whether bias can be statistically adjusted. Indeed, there are studies in the literature on the statistical correction of the rater effects. Raymond and Houston (1990) conducted a study with the purpose of determining and correcting for rater effects in performance assessment. In the research four different procedures; ordinary least squares, weighted least squares, the Rasch model (with a two facets design that includes only raters and examinees and that provides results similar to the Wright and Masters (1982) rating scale model) and data imputation via the E-M algorithm were considered on a simulated data set. The results of the research showed that each of the methods yields more accurate estimates of true levels of performance than the classical approach of summing observed ratings. In the Houston, Raymond and Svec's (1991) study the methods of ordinary least squares, weighted least squares and imputation of the missing data were examined for correcting rater severity and leniency. In the study, simulation data was used and root-mean-squared-error (RMSE) was employed in order to assess the accuracy of the methods in estimating true scores. The research results indicated that the three correction methods used consistently outperformed the procedure of averaging the observed ratings. In another study by Raymond and Viswesvaran (1993), it was aimed to elucidate a simple and flexible method to statistically control for specific types of rating error. In accordance with this purpose, three different models namely ordinary least squares; weighted least squares; and ordinary least squares, subsequent to applying a logistic transformation to observed ratings were performed to data obtained from an oral examination where each of 115 examinees graded by four raters. The study results revealed that the models used for correction of ratings increases reliability. In addition to the methods used in the

researches listed, the literature also includes an approach based on the many-facet Rasch model (MFRM) proposed by Linacre (2018) to adjust rater bias statistically; however, there are no empirical studies of its effectiveness.

1.3. Aim of the Study

This study aimed to test the effectiveness of MFRM statistical adjustment of rater bias empirically. It investigates the effects of statistical bias adjustment on estimating the abilities of examinees, on the difficulty indices of items and on measures of rater severity/leniency.

2. METHOD

2.1. Model of the Study

This study focused on testing a model for the process of bias adjustment, and was therefore designed as basic research. Basic research, rather than seeking answers to real-life problems, addresses issues that offer theoretical contributions to science, build theories and generate new knowledge (Connaway & Powell, 2010). Basic research also formalizes theories and tests hypotheses involving abstract concepts (Bailey, 1994).

2.2. Participants

The participants included 95 eighth-grade students, of whom, 49 (51.58%) were female, and 46 (48.42%) were male. Three mathematics teachers graded their responses to open-ended questions.

2.3. Data Collection Tools

In the study two data collection tools was used. The first was the Mathematics Achievement Test developed by İlhan (2016). This test contains six open-ended questions. According to results reported by İlhan (2016), the test had a one-dimensional structure. It explained 31.18% of variance ratio, and the factor loads of its items were found to range between .51 and .64.

This study's second data collection tool was a rubric used to grade the students' responses to the open-ended questions. This rubric was also developed by İlhan (2016). The rubric has a holistic structure and four categories: *inadequate* (0), *needs to be developed* (1), *dood* (2) and *very good* (3). Ilhan (2016) indicated that these categories were intended to reflect the adequacy of responses on five levels: understanding of the problem, method of solving the problem, the processes carried out to solve the problem, the accuracy of the results obtained and how the solution was obtained.

2.4. Data Collection, Psychometric Characteristics, and Analysis

Data were collected in the spring term of 2018. Administering the achievement test to the 95 eighth-grade students was the first stage of data collection. Their responses were graded by three mathematics teachers. The rubric used in the study had been introduced to the raters beforehand. The raters were also told that they should rate all answers to the one question before moving to the next question, and that they should not include variables outside the construct measured, such as appealing handwriting and spatial organization of the responses. After the rating, the data were analyzed using MFRM. FACETS software was used for the analysis.

Statistical indicators of whether the Rasch analysis assumptions were met were investigated firstly. Rasch analysis has three assumptions: unidimensionality, local independence, and model-data fit (DeMars, 2010). However, there is no need to test each assumption one by one since they are all related. That is to say, model-data fit indicates that the unidimensionality assumption has been met (Lee, Peterson, & Dixon, 2010), which indicates that there is no problem with local independence (Nandakumar & Ackerman, 2004). Therefore, the fundamental assumption that needs to be tested is whether there is model-data fit (Güler, İlhan,

Güneyli, & Demir, 2017). This assumption is tested by examining standardized residuals. The number of standardized residuals outside the ±2 range should not exceed 5% of the total number of data, and those outside ±3 should not exceed 1%, according to Linacre (2018). In this study, the total number of data was 1,710 since it involved 95 students, six items and three raters (95×6×3). The number of standardized residuals outside the range of ±2 was found to be 76 (4.44%) and the number of standardized residuals outside the range of ±3 was found to be 16 (0.94%). This indicated adequate model-data fit, and that the assumptions of Rasch analysis had been met.

After determining that the assumptions were met, the psychometric characteristics of the study data were investigated. The results for reliability and model-data fit in MFRM are shown in Table 1. The infit and outfit indices in all three of the examinee, item and rater facets were within the range of .5 and 1.5, the recommended criteria for their interpretation (Wright & Linacre, 1994). These fit indices indicate model-data fit and the validity of the measurements.

Table 1. Results for reliability and model-data fit in MFRM.

Facet	Infit	Outfit	Separation Index	Reliability	df	Chi square
Examinee	.99	1.01	2.19	.83	94	443.00**
Item	.99	1.01	13.20	.99	5	857.20**
Rater	.99	1.01	5.51	.97	2	62.40**

** $p<.001$

Table 1 shows that the chi-square value for the rater facet was significant, and that the reliability coefficient and separation index were high. This indicated a significant difference between the raters' severity and leniency. Despite this difference, the values reported for the facets of item and examinee indicated that the measures were reliable because the chi-square values for the facets of examinee and item were significant, the reliability coefficients exceeded .80, and the separation indices were higher than 2 (Linacre, 2012). Thus, the students' performances on the different test items can be rated independently, and examinees with different mathematical performances were distinguished with high reliability.

Following the psychometric investigation of the study data, the datasets were prepared for bias adjustment. The comparison of the analysis outcomes obtained from a dataset not involving rater bias and the analysis outcomes reached in case of the inclusion of bias in this dataset and the adjustment of the bias included was thought to be the most convenient way to set forth the effectiveness of the statistical adjustment applied. For this reason; while preparing the dataset for the bias adjustment, the original rater biases were excluded from the dataset to create a dataset with no apparent rater bias–the unbiased dataset. The results of analysis indicated significant relationships between rater 1 and examinee 84 (bias size=1.57, t=2.66) and rater 2 and examinee 23 (bias size=1.75, t=2.66). These two raters graded two examinees mentioned more leniently than expected. Therefore, the data for examinees 84 and 23 were excluded from the dataset, creating a dataset where three raters graded 93 students' responses to six open-ended mathematics questions and no rater bias. This dataset's measurements of examinees' ability levels, item difficulty indices, rater severity/leniency were used as the criteria for the effectiveness of bias adjustment.

In the second stage of the testing the effectiveness of bias adjustment, some changes were made in the dataset so that it would contain "*rater* × *examinee*" bias. The grading of rater 1 for examinees 1 to 10 and rater 2 for examinees 11 to 20 were increased by one in some parts of the test and by two in others, creating a dataset where bias was encountered in 20 of the 279 [(93 examinees) × (3 raters)] possible interactions between raters and examinees.

In the final stage, the bias adjustment formula was applied to the biases included in the dataset. A fourth facet, bias adjustment, was incorporated in the analysis, along with the three facets of rater, examinee and item. In this facet, *rater* × *examinee* biases in the dataset were listed, and bias adjustment was applied to the grading of rater 1 for examinees 1 to 10 and rater 2 for examinees 11 to 20. No other bias adjustments were done. The *rater* × *examinee* interactions to which the bias adjustment was applied were encoded as 1, and the *rater* × *examinee* interactions where bias adjustment was not necessary were encoded as 2. Thus, syntax containing the four facets of rater, examinee, item and bias adjustment were prepared for many-facet Rasch analysis. At this point, the comparison of the three datasets proceeded.

In this study, the consistency between the ability estimations in the unbiased, biased and adjusted dataset was examined using Pearson's product-moment correlation and the paired samples *t*-test. Correlation analysis and the *t*-test were done using IBM SPSS 20 software. The effect of bias adjustment on item difficulty indices and rater severity/leniency could not be statistically determined since the number of items was limited to six, and the number of raters to three. It was possible only to investigate how close item difficulties and measures regarding raters were to the values in the unbiased dataset.

3. RESULT and DISCUSSION

This section includes the study's results. The ability estimations in the unbiased, biased and adjusted datasets are shown in Table 2. As Table 2 shows, there were significant differences between ability estimations in the unbiased and biased datasets. These differences were valid for almost all participants, but were more explicit for the first 20 students who served as a source for the *rater* × *examinee* bias. Table 2 showed that the ability estimations after the application of statistical adjustment to the *rater* × *examinee* bias were significantly closer to the ability scores in the unbiased dataset. This means that the effect of rater bias on the examinee's ability estimations can be controlled by bias adjustment. However, in order to reach a more powerful judgement, the relationship between the ability estimations in the unbiased, biased and bias adjusted datasets needed to be tested statistically. In order to determine statistically how much bias adjustment brings the ability estimations closer to the ability estimations in the unbiased dataset, correlation analysis and the paired samples *t*-test were done. Their outcomes are shown in Table 3.

Table 3 shows that there was a positive, powerful and significant relationship between ability estimations in the unbiased dataset and biased datasets [r=.896, p<.001]. However, it should not be overlooked that there was a significant difference between the mean ability scores in these two datasets [$t_{(92)}$=5.03, p<.001]. Better to say, *rater* × *examinee* bias did not have a great impact on the ordering of the examinees' ability levels, but significantly affected their ability estimations. A comparison of the ability estimations in the bias adjusted and unbiased datasets found a perfect positive relationship [r=.996, p<.001]. No significant difference was found between the ability estimations in the two datasets [$t_{(92)}$=1.11, p>.05]. This indicated that the effects created by the *rater* × *examinee* bias on the ability estimations can be, to a large extent, eliminated by bias adjustment. The effect of bias adjustment on item difficulty indices and rater severity and leniency measurements are shown in Table 4.

Table 2. Ability estimations in the unbiased, biased and adjusted bias datasets.

Examinee Number	No bias	Bias	Adjusted bias	Examinee Number	No bias	Bias	Adjusted bias	Examinee Number	No bias	Bias	Adjusted bias
E1	-0.08	0.58	0.00	E32	-1.04	-0.97	-1.06	E63	0.28	0.25	0.28
E2	1.46	1.73	1.44	E33	-0.44	-0.42	-0.46	E64	-1.26	-1.17	-1.29
E3	-0.08	0.58	0.00	E34	-0.17	-0.17	-0.18	E65	-0.63	-0.60	-0.65
E4	-0.54	0.25	-0.48	E35	-0.08	-0.08	-0.08	E66	-0.83	-0.78	-0.85
E5	-3.15	-1.07	-2.78	E36	-0.73	-0.69	-0.75	E67	-0.35	-0.33	-0.36
E6	-1.50	-0.42	-1.58	E37	-1.04	-0.97	-1.06	E68	-1.50	-1.39	-1.53
E7	-1.38	-0.25	-1.28	E38	-1.26	-1.17	-1.29	E69	0.46	0.41	0.46
E8	-0.83	0.08	-0.73	E39	-0.54	-0.51	-0.55	E70	-0.08	-0.08	-0.08
E9	-0.93	-0.17	-1.13	E40	-2.21	-2.06	-2.26	E71	-0.35	-0.33	-0.36
E10	0.01	0.58	0.00	E41	0.37	0.33	0.37	E72	-0.63	-0.60	-0.65
E11	0.10	0.58	0.09	E42	-1.76	-1.63	-1.79	E73	-0.26	-0.25	-0.27
E12	0.10	0.50	-0.03	E43	0.01	0.00	0.01	E74	-1.04	-0.97	-1.06
E13	-0.63	0.08	-0.65	E44	-0.35	-0.33	-0.36	E75	0.73	0.67	0.74
E14	0.28	0.67	0.21	E45	-0.63	-0.60	-0.65	E76	-0.54	-0.51	-0.55
E15	-1.76	-0.33	-1.35	E46	1.23	1.14	1.26	E77	-0.83	-0.78	-0.85
E16	0.10	0.58	0.09	E47	-1.90	-1.77	-1.94	E78	-0.93	-0.87	-0.96
E17	-0.83	-0.08	-0.91	E48	-1.26	-1.17	-1.29	E79	-0.08	-0.08	-0.08
E18	-0.08	0.41	-0.15	E49	-1.15	-1.07	-1.17	E80	-0.83	-0.78	-0.85
E19	0.28	0.67	0.21	E50	-0.73	-0.69	-0.75	E81	-0.08	-0.08	-0.08
E20	-0.17	0.33	-0.27	E51	-0.54	-0.51	-0.55	E82	-0.26	-0.25	-0.27
E21	-1.15	-1.07	-1.17	E52	-0.54	-0.51	-0.55	E83	-0.17	-0.17	-0.18
E22	-0.44	-0.42	-0.46	E53	-0.35	-0.33	-0.36	E84	-0.54	-0.51	-0.55
E23	0.46	0.41	0.46	E54	-0.35	-0.33	-0.36	E85	-2.21	-2.06	-2.26
E24	-2.21	-2.06	-2.26	E55	-0.83	-0.78	-0.85	E86	0.73	0.67	0.74
E25	-2.60	-2.43	-2.65	E56	-1.62	-1.51	-1.66	E87	0.01	0.00	0.01
E26	0.64	0.58	0.65	E57	-0.63	-0.60	-0.65	E88	-0.08	-0.08	-0.08
E27	0.19	0.16	0.19	E58	-0.93	-0.87	-0.96	E89	0.64	0.58	0.65
E28	-0.17	-0.17	-0.18	E59	-0.73	-0.69	-0.75	E90	0.55	0.50	0.55
E29	0.10	0.08	0.10	E60	-0.54	-0.51	-0.55	E91	-1.04	-0.97	-1.06
E30	-0.26	-0.25	-0.27	E61	-0.93	-0.87	-0.96	E92	-1.50	-1.39	-1.53
E31	-0.93	-0.87	-0.96	E62	-0.35	-0.33	-0.36	E93	-0.44	-0.42	-0.46

Table 3. Correlation analysis and paired samples *t*-test results for the comparison of the ability estimations in the unbiased, biased and adjusted datasets.

Comparison	Dataset	Mean (Logit)	Standard Deviation	r	df	t
No bias – Bias	No bias	-.55	.80	.896**	92	5.03**
	Bias	-.36	.75			
No bias – Adjusted bias	No bias	-.55	.80	.996**	92	1.11*
	Adjusted bias	-.56	.79			

* $p>.05$, ** $p<.001$

Table 4. Item difficulty indices and rater severity and leniency measurements in the unbiased, biased and adjusted datasets.

	Item Difficulty Indices		
	No bias	Bias	Adjusted bias
I1	.86	.74	.85
I2	-1.40	-1.21	-1.38
I3	.53	.48	.53
I4	.14	.13	.13
I5	-1.29	-1.16	-1.33
I6	1.16	1.02	1.20

	Rater Severity/ Leniency Measures		
	No bias	Bias	Adjusted bias
R1	.00	-.10	-.01
R2	-.31	-.33	-.30
R3	.30	.43	.31

Table 4 shows that the item difficulties and rater measurements in the unbiased and biased datasets were quite different. On the other hand, the item difficulties and rater measurements in the adjusted dataset were extremely close to those of the unbiased dataset. In other words, the differences caused by *rater* × *examinee* bias in the item difficulty indices and rater severity and leniency measurements were largely eliminated by bias adjustment, although not entirely.

4. CONCLUSION

This study investigated the effectiveness of MFRM statistical adjustment of rater biases. Its dataset, which did not include any *rater* × *examinee* bias, was altered to involve r*ater* × *examinee* bias. Then, bias adjustment was applied to the rater biases included in the dataset, and the effectiveness of the statistical adjustment was tested. Ability estimations, item difficulties, and rater measurements in the bias adjusted dataset were compared to those in the unbiased dataset. The correlation analysis results failed to indicate complete consistency, despite a strong relationship between ability estimations in the dataset that did not include *rater* × *examinee* bias and the biased dataset. On the other hand, it was determined that there was excellent consistency between the ability estimations calculated after bias adjustment and ability estimations in the unbiased dataset.

A significant difference was also found between the ability estimations in the dataset that did not include *rater* × *examinee* bias and the ability estimations in the biased dataset. No significant differences were found between ability estimations in the bias adjusted dataset and those in the unbiased dataset. All these results reveal that the effects of rater biases on examinees' ability estimations can be eliminated by bias adjustment. This was also the case for item difficulty indices and rater severity and leniency measurements. A comparison of the three datasets determined that differences caused by *rater* × *examinee* bias on item difficulties and rater measurements were almost entirely eliminated.

5. IMPLICATIONS for PRACTICE

This study's results indicate that MFRM bias adjustment can serve as a way to minimize the effects of rater bias. However, it should be underlined that this does not mean that statistical bias adjustment can replace other methods of reducing rater bias such as rater training, blind scoring or using rubrics. The most accurate interpretation based on research results is that statistical adjustment should be performed for observed biases when rater bias occurs despite precautions such as using rubrics or training raters. More clearly, just as statistical controls can be used to support physical controls, but not replace them in scientific researches, bias adjustment should be considered a way to support rater training, blind scoring or the use rubrics, not as an alternative to them.

ORCID

Mustafa İlhan https://orcid.org/0000-0003-1804-002X

6. REFERENCES

Aubin, A. S., St-Onge, C., & Renaud, J. S. (2018). Detecting rater bias using a person-fit statistic: A Monte Carlo simulation study. *Perspectives on Medical Education*, *7*(2), 83-92. http://dx.doi.org/10.1007/s40037-017-0391-8

Bailey, K. (1994). *Methods of social research*. New York: The Free.

Bennett, R. E. (1991). On the meanings of constructed response. *ETS Research Report Series*, *2*, 1-46. http://dx.doi.org/10.1002/j.2333-8504.1991.tb01429.x

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). Toward a framework for constructed response items. *ETS Research Report Series*, *1*, 1 - 29. http://dx.doi.org/10.1002/j.2333-8504.1990.tb01348.x

Connaway, L. S., & Powell, R. R. (2010). *Basic research methods for librarians*. Santa Barbara, CA: Libraries Unlimited.

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221. http://dx.doi.org/10.1207/s15434311laq0203_2

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, *1*(1), 1-16. Retrieved from http://www.ijlt.ir/portal/files/401-2011-01-01.pdf

Güler, N., İlhan, M., Güneyli, A., & Demir, S. (2017). An evaluation of the psychometric properties of three different forms of Daly and Miller's writing apprehension test through Rasch analysis. *Educational Sciences: Theory & Practice*, *17*(3), 721-744. http://dx.doi.org/10.12738/estp.2017.3.0051

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, *33*(2), 87 - 102. Retrieved from http://www.celea.org.cn/teic/90/10060807.pdf

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

Hogan, T. P., & Murphy, G. (2007) Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, *20*(4), 427-441. http://dx.doi.org/10.1080/08957340701580736

Houston, W. M., Raymond, M.R., & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement*, *15*(4), 409-421. http://dx.doi.org/10.1177/014662169101500411

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*(1), 64–86. http://dx.doi.org/10.1037/1082-989X.5.1.64

İlhan, M. (2015). *The identification of rater effects on open-ended math questions rated through standard rubrics and rubrics based on the SOLO taxonomy in reference to the many facet Rasch model*. Doctoral dissertation, Gaziantep University, Gaziantep, Turkey. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

İlhan, M. (2016). *Comparison of the ability estimations of classical test theory and the many facet Rasch model in measurements with open-ended questions*. *Hacettepe University Journal of Education*, *31*(2), 346–368. http://dx.doi.org/10.16986/HUJE.2016015182

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26-43. http://dx.doi.org/10.1016/j.asw.2007.04.001

Kondo Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3 - 31. https://doi.org/10.1191/0265532202lt218oa

Kumar, DSP D. (2005). Performance appraisal: The importance of rater training. *Journal of the Kuala Lumpur Royal Malaysia Police College*, *4*, 1 - 15. Retrieved from http://rmpckl.rmp.gov.my/Journal/BI/performanceappraisal.pdf

Lee, M., Peterson, J. J., & Dixon, A. (2010). Rasch calibration of physical activity self-efficacy and social support scale for persons with intellectual disabilities. *Research in Developmental Disabilities*, *31*(4), 903-913. http://dxdoi.org/10.1016/j.ridd.2010.02.010

Linacre, J. M. (2012). *Many-facet Rasch measurement: Facets tutorial*. Retrieved from http://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2018). *A user's guide to FACETS Rasch-model computer programs*. Retrieved from https://www.winsteps.com/manuals.htm

McNamara, J. F., Erlandson, D. A., & McNamara, M. (2013). *Measurement and evaluation: Strategies for school improvement.* New York, NY: Routledge.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applıed Measurement*, *5*(2), 189-227. Retrieved from http://jimelwood.net/students/grips/tables_figures/myford_wolfe_2004.pdf

Nandakumar, R., & Ackerman, T. A. (2004). Test modeling. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 93-105). Thousand Oaks, CA: Sage.

Raymond, M. R., & Houston, W. M. (1990). *Detecting and correcting for rater effects inperformance assessment* (ACT Research Rep. No. 90-14). Iowa City, American College Testing. Retrieved from http://www.act.org/content/dam/act/unsecured/documents/ACT_RR90-14.pdf

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30*(3), 253-268. http://dx.doi.org/10.1111/j.1745-3984.1993.tb00426.x

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413-428. http://dx.doi.org/10.1037/0033-2909.88.2.413

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370-371. Retrieved from https://www.rasch.org/rmt/rmt83b.htm