Analysis of Open-Ended Statistics Questions with Many Facet Rasch Model

Neşe GÜLER*

Suggested Citation:

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research*, 55, 73-90. http://dx.doi.org/10.14689/ejer.2014.55.5

Abstract

Problem Statement: The most significant disadvantage of open-ended items that allow the valid measurement of upper level cognitive behaviours, such as synthesis and evaluation, is scoring. The difficulty associated with objectively scoring the answers to the items contributes to the reduction of the reliability of the scores. Moreover, other sources of error also affect reliability. When measurement involves more than one source of error, as in the case of scoring open-ended items, item response theory, which removes the restriction of the classical test theory, is preferred.

Purpose of Study: The purpose of the study is to assess the infit-outfit statistics and reliability coefficients of the scores for a statistics exam composed of open-ended items using the many facet Rasch model (MFRM) analysis for each source of variability (i.e., students, items, and raters) and to interpret the reliability of the scores.

Methods: In this study, MFRM was used to analyse the answers given to 10 open-ended items in a Statistics I course; the answers were provided by 55 third year graduate students of the Psychological Counselling and Guidance Department of the Faculty of Education in the fall semester of the 2010-2011 academic year. The scoring was performed by three raters who were experts in statistics and work as academic staff at the university. Thereby, this study contains the following three sources of variability (facets): students, items, and raters. Measurement reports, including infit and outfit statistics, separation indexes and reliability coefficients were calculated for each facet by FACET computer package programme.

Findings and Results: According to the MFRM analysis, the reliability coefficients for the student and item facets were .79 and .90, respectively; moreover, the separation indexes of the student and item facets were 1.95

* Dr. Sakarya University, Sakarya, Turkey, gnguler@gmail.com

and 2.95, respectively. Additionally, complete consistency was found between the raters in this study.

Conclusions and Recommendations: The MFRM makes important contributions to the analysis of measurement results, the development of measurement tools, the organization of appropriate measurement circumstances, and the provision of effective training for raters. Because it is believed to provide important information, the use of the MFRM might be recommended when analysing the results obtained from exams in which open-ended items are used and through which important decisions about the students' future are made.

Keywords: Open-ended questions, reliability, many facet Rasch model

Introduction

One measurement tool that is frequently used in education is the open-ended item. These items enable students to freely communicate their answers and allow educators to analyse the insufficiencies and mistakes that cannot be analysed using other typical measurement tools, such as multiple-choice items (Hong, 1984). Open-ended items also enable students to pursue the process of thinking and strategy formation. Thus, those items help educators to understand each student's level of knowledge and asses how they structure their knowledge. Compared to multiple choice items, open-ended items offer three distinct advantages. Firstly, they remove the ability to select correct answers by chance, thus lessening measurement errors associated with other methods, which ultimately ensures the ability to obtain more reliable results. Secondly, open-ended items also eliminate the students' ability to select the correct answer through a process of elimination. For instance, for the equation $2(X + 4) = 38 - X$, students can find the correct answer with a multiple choice item by simply replacing X with the values given in the options and computing the answers, thereby succeeding despite not using the approach that is being assessed. In this instance, the measurement is actually assessing a different approach than the intended one, thereby leading to a decrease in the construct validity of the test. This is not an issue with open-ended items. Thirdly, open-ended items do not permit students to use corrective feedback in order to find the correct answer after a failed attempt; note that this is an issue with multiple-choice items (Bridgeman, 1992). For example, when a student fails to find the correct answer among the options on a multiple choice item, they can return to the question and employ a new strategy in an attempt to find the answer.

Despite the advantages, open-ended items also present some disadvantages. For example, a large portion of the time allocated for the entire assessment must be dedicated to writing the answers rather than thinking about them. This means fewer items can be included for a given amount of time permitted for the assessment, which lowers the sensitivity and the content validity of the measurement tool (Özçelik, 1998). Moreover, one of the most significant disadvantages of open-ended items that allow the valid measurement of upper level cognitive behaviours, such as

synthesis and evaluation, is scoring. The difficulty associated with objectively scoring the answers to the items contributes to a reduction in the reliability of the scores. Therefore, a great many studies have analysed the scoring of items (Alharby, 2006; Geer, 1988; Güler & Gelbal, 2010a; Hong, 1984; Levia, Rios, & Martinez, 2006). When open-ended items are rated as correct-incorrect (0-1), intra-rater reliability (i.e., whether or not a rater gives the same scores to the same answers at different times) and inter-rater reliability (i.e., whether or not different raters give the same scores to a specific answer) must be tested. The method commonly used to assess these aspects in the classical test theory (CTT) is the calculation of the Pearson product-moment correlation coefficient between the scores. A correlation value close to 1 suggests that consistency between scorings (i.e., rating reliability) is high, whereas a value close to 0 means that scoring reliability is low. However, note that the correlation coefficient is influenced by the size of the sample and that it is independent of the score averages; in fact, these are the greatest restrictions to its use. In this case, the t test between the score averages should be used when two scoring situations are available, and ANOVA should be used alongside the correlation coefficient in cases with more than two scoring situations (Goodwin, 2001; Güler & Gelbal, 2010b).

In addition to the scoring of answers given to for open-ended items, other sources of error can affect reliability. For example, different reliability coefficients can be calculated for various sources of error, such as the internal consistency for each item on the entire test and test-retest reliability (i.e., the consistency between answers given by the same student for the same items at different times), both of which are available with the CTT. Note that reliability cannot be assessed using a method through which all the sources of error are simultaneously assessed. In cases of measurement that contain more than one source of error (e.g., the case of scoring open-ended items), the generalizability theory and item response theory are preferred because they remove the restriction of the CTT. This study examines the reliability of scores from open-ended statistical items by using the many facet Rasch model (MFRM), which is an extension of the Rasch model developed by Linacre (1989). The MFRM is part of the item response theory, and the sources of student, item, and rater variability are treated together.

Many Facet Rasch Model

The MFRM is conceptually similar to regression analysis; i.e., the dependent variable is the logistic transformation of the probability of the rates of scores that a student can achieve on an item, and the independent variables are the sources of variability (facets), such as a student's level of ability, an item's level of difficulty, and a rater's level of severity/leniency in scoring (Randall & Engelhard, 2009). Thus, a MFRM in which three facets are available can be stated as follows:

$$\text{Log}\,(P_{sirc} / P_{sirc\text{-}1}) = B_s - D_i - C_r - F_c$$

P_{sirc}: The probability of student "s" being rated on item "i" by rater "r" with category c.

$P_{sirc\text{-}1}$: The probability of student "s" being rated on item "i" by rater "r" with category c-1.

B_s: The ability of student "s.".

D_i: The difficulty of item "i."

C_r: The severity of rater "r."

F_c: The difficulty of the step up from category c-1 to category c.

MFRM enables parameter predictions that are independent of the sample in relation to each facet. More specifically, item response theory is based on the supposition that no interactions exist between the facets; in contrast, the generalizability theory supposes that interactions do exist between them (Smith & Kulikowich, 2004). The latter approach allows for the observation of the levels of different facets, such as students' levels of ability, raters' levels of severity/leniency in scoring, and the levels of difficulty for items, on a single linear scale (i.e., usually called the logit scale). On such a scale, each student's level of ability is included as predictions that are independent of the distributional properties of certain items and independent of the scores given by certain raters. Similarly, predictions can be made for the levels of difficulty for the test items and raters' levels of severity/leniency, both of which are independent of the distributional properties of the other facets (Smith & Kulikowich, 2004). Moreover, students who are not considered to be a source of variability (facet) in the generalizability theory and who are described as an object of measurement are also considered as a source of variability in the MFRM. Thus, it is possible to simultaneously calculate separate reliability coefficients for each facet (e.g., students, items and raters) (Alharby, 2006).

The probabilities of students' answers are called "log-odds" and are represented on a logit table with "log-odds" units or "logits" units. Increasing positive values on the logits table reference high abilities for students, a high level of difficulty for the items, and increased severity for the raters; accordingly, high negative values are related with lower levels of ability for students, lower levels of difficulty for the items, and leniency in scoring for the raters. The visibility of the levels of each facet on such a logit table allows the researcher to see the ordering of the levels of the facets and the size of the difference between each element of each facet (Güler & Gelbal, 2010a; Hetharman, 2004).

In the MFRM, the infit and outfit statistical values are used to evaluate the suitability of the data. Additionally, the reliability of the separation index is also examined for each facet. This coefficient is an indicator of the extent to which the elements in the source of each variability (e.g. individuals or items) are separated from each other and is the proportion of the real score variance to the observed score variance. It is calculated using the following equation:

$$R = SD^2 - MSE / SD^2$$

Here, SD^2 represents the observed variance in a facet, while MSE represents the squared average of the prediction error (i.e., the square of the standard error) (Engelhard, 1994; Randall & Engelhard, 2009). Andrich (1982) gave detailed explanation how the separation index is obtained as well as the relationship with the KR-20 coefficient in detail (as cited in Engalhard, 1994, pg.62). In accordance with the

determination of rater consistency, high values for the reliability of separation index for raters predicate that there are significant differences in rater consistency.

The FACET programme, i.e., a software package that computes the MFRM, can be used to produce a table showing the unexpected responses in addition to logit scale, infit and outfit statistics and reliability coefficients for each facet. In cases with facets with low reliability, unexpected response tables can provide important information to diagnose what the source of unacceptable reliability value. In fact, the table shows which rater has scored which student's answer in an unexpected way. For instance, in a case with three raters, if the first rater has given a low score to the tenth student for their answer to the second item, whereas the second and third raters have given higher scores for the same response, this table will contain information to reveal this situation. Thus, it helps to detect unexpected responses when they emerge and helps the researchers to determine what the sources of low reliability and to plan more reliable measurement conditions. The research conducted by Nakamura (2002) offers information on the education of raters and on the revision of items; it is a good reference for those who wish to obtain further knowledge on this particular aspect.

This study uses the MFRM analysis to compute the separation indexes and reliability coefficients of the scores for a statistics exam composed of open-ended items for each facet (i.e., students, items, and raters) and to interpret the reliability of the scores.

Method

Researc Design

This is a descriptive survey and qualitative research method was used. In the study the MFRM to analyse the answers given to 10 open-ended items.

Study Group

This study uses the MFRM to analyse the answers given to 10 open-ended items in a Statistics I course; the answers were provided by 55 third year graduate students of the Psychological Counselling and Guidance Department of the Faculty of Education in the fall semester of the 2010-2011 academic year. Twenty nine of the students were female and twenty six were male.

Research Instrument and Procedure

The scoring was performed by three raters who were experts in statistics and work as academic staff at the university. In order to prepare the answer key for the test, the raters answered the items separately and then compared their answers; consequently, they agreed on the common answers for the answer key. Moreover, in case answers provided by students required comments from the raters, all potential answers were also noted. Thus, an answer key was jointly formed, and the raters used this common answer key to independently perform scorings of the tests. One of the statistic items is shown in Table 1 as an example.

Table1

Examples of the Statistics Questions

Question: The number of correct answers from 8 students on a 20-item mathematics exam.

Students	Number of correct answers
1	6
2	13
3	16
4	19
5	9
6	1
7	6
8	10

According to table;

a) Find the mode, median, and mean of the data.

b) Draw the histogram of the data and interpret it (At most 3 sentences).

The standard deviation for the number of correct answers is 6. The same students also took physics, chemistry, and Turkish exams, each of which included 20 items, and the standard deviations for the number of correct answers in the physics exam, the chemistry exam, and the Turkish exam were 8, 3, and 5, respectively.

According to these standard deviations, interpret the variability of the students' correct answers in 3 sentences or less.

As such, this study contains the following three sources of variability: students, items, and raters. Measurement reports, including infit and outfit statistics and standardized residual values, were calculated for each facet. The FACET computer package programme developed by Linacre (2007) was employed in the analyses of the answers.

Results

The logit table for the scores given by the three raters for the answers from the 55 students on the 10 item test is presented in Figure 1.

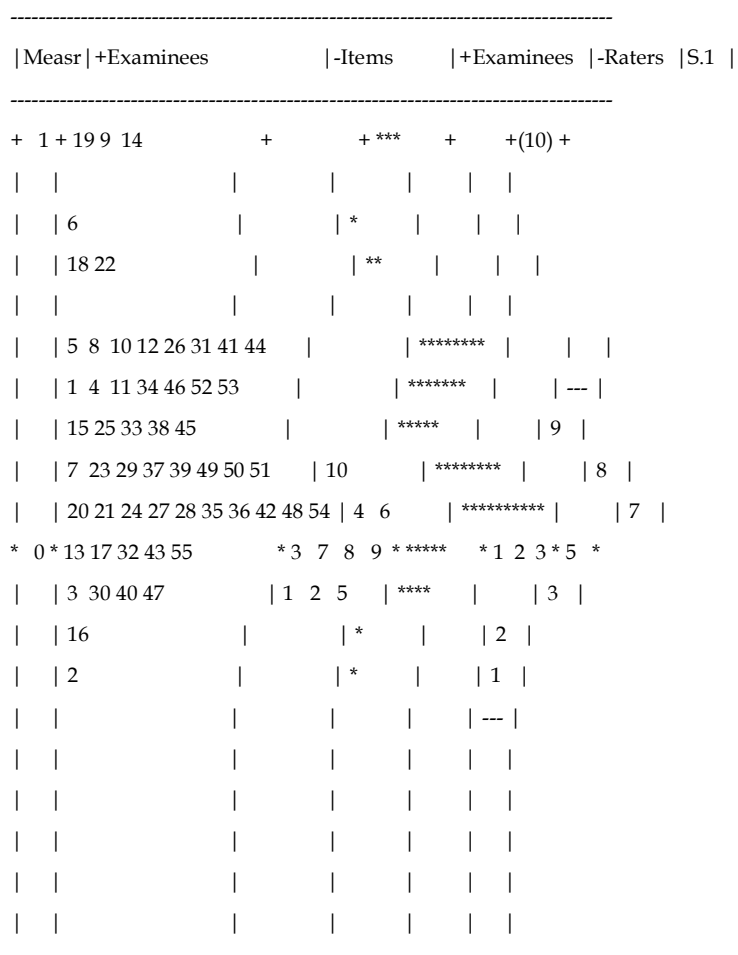

Figure 1: Logit Map for Three Facets

This figure presents the results for all the sources of variability on a single linear scale. As can be seen in the column where the students (i.e., "examinees") are shown, the students were ordered based on their scores from -1 to 1, i.e., from those with the fewest correct answers to those with the most correct answers. Thus, student 2 showed the least ability, whereas students 19, 9, and 14 demonstrated the highest ability level. In the column showing the items ranked according to difficulty, the item closest to -1 is the item with the lowest level of difficulty (i.e., the least difficult), and

as the item approaches 1, its level of difficulty increases (i.e., it becomes more difficult). Hence, the least difficult items are number 1 and 2, and the most difficult item is number 10. In the raters' column, the movement from -1 to 1 demonstrates a movement from the most lenient rater to the strictest rater. From Figure 1, the three raters included in the study performed at the same level of severity/leniency; i.e., they all scored at the level of 0 logits in terms of scoring. Table 2 shows the analysis report in relation to the students.

Table 2

Students' Measurement Report

Obsvd Score	*Obsvd Count*	*Obsvd Average*	*Fair-M Average*	*Measure*	*Model S.E.*	*Infit*		*Outfit*		*Nu*	*Student*
						MnSq	*Zstd*	*MnSq*	*Zstd*		
300	30			(2.00	1.60)	Maximum				19	19
298	30	9.9	9.94	.95	.41	0.3	0	0.3	0	9	9
298	30	9.9	9.94	.95	.41	0.3	0	0.2	0	14	14
297	30	9.9	9.91	.83	.31	0.5	0	0.7	0	6	6
295	30	9.8	9.85	.69	.22	1.2	0	0.8	0	18	18
295	30	9.8	9.85	.69	.22	1.1	0	0.4	0	22	22
290	30	9.7	9.70	.54	.14	0.8	0	0.3	0	26	26
290	30	9.7	9.70	.54	.14	0.7	0	0.5	0	44	44
289	30	9.6	9.67	.52	.13	0.7	0	0.4	0	10	10
289	30	9.6	9.67	.52	.13	0.8	0	0.4	0	10	10
289	30	9.6	9.67	.52	.13	0.6	0	0.5	0	41	41
288	30	9.6	9.65	.50	.13	0.7	0	0.4	0	8	8
288	30	9.6	9.65	.50	.13	0.7	0	0.9	0	31	31
286	30	9.5	9.59	.47	.12	0.6	0	0.4	0	5	5
281	30	9.4	9.45	.42	.10	0.7	0	0.9	0	1	1
280	30	9.3	9.42	.41	.10	1.1	0	0.8	0	11	11
280	30	9.3	9.42	.41	.10	1.0	0	0.5	0	46	46
279	30	9.3	9.39	.40	.09	0.7	0	0.7	0	52	52
278	30	9.3	9.36	.39	.09	1.2	0	0.5	0	34	34
276	30	9.2	9.30	.37	.09	0.6	0	0.4	-1	4	4
274	30	9.1	9.24	.36	.08	0.9	0	0.7	0	53	53
267	30	8.9	9.04	.32	.07	1.3	0	1.6	0	25	25
265	30	8.8	8.98	.31	.07	1.1	0	0.9	0	15	15
255	30	8.5	8.67	.26	.06	0.7	0	0.8	0	33	33
255	30	8.5	8.67	.26	.06	1.2	0	1.0	0	38	38
256	30	8.5	8.70	.26	.06	0.7	0	0.6	0	45	45
252	30	8.4	8.57	.25	.06	1.0	0	0.9	0	49	49
249	30	8.3	8.48	.24	.06	1.1	0	1.0	0	23	23
238	30	7.9	8.11	.20	.06	0.8	0	1.0	0	7	7
235	30	7.8	8.01	.19	.05	0.8	0	0.8	0	29	29
234	30	7.8	7.98	.19	.05	1.1	0	1.1	0	50	50
231	30	7.7	7.88	.18	.05	0.9	0	0.9	0	37	37
233	30	7.8	7.94	.18	.05	0.9	0	1.0	0	39	39
228	30	7.6	7.77	.17	.05	1.4	1	1.5	1	51	51

Table 2 Continue

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit		Outfit		Nu	Student
						MnSq	Zstd	MnSq	Zstd		
220	30	7.3	7.49	.15	.05	1.2	0	1.1	0	35	35
220	30	7.3	7.49	.15	.05	1.2	0	1.2	0	54	54
201	29	6.9	7.11	.12	.05	1.4	1	1.5	1	28	28
197	30	6.6	6.67	.09	.05	1.1	0	1.0	0	27	27
196	30	6.5	6.63	.09	.05	1.0	0	1.0	0	42	42
190	30	6.3	6.41	.08	.05	1.2	0	1.2	1	24	24
187	30	6.2	6.30	.07	.05	1.2	0	1.1	0	48	48
181	30	6.0	6.08	.06	.05	1.3	1	1.3	1	20	20
183	30	6.1	6.15	.06	.05	0.8	-1	0.8	-1	21	21
183	30	6.1	6.15	.06	.05	0.9	0	0.9	0	36	36
167	30	5.6	5.56	.03	.04	0.8	-1	0.8	-1	43	43
154	30	5.1	5.09	.01	.04	1.1	0	1.0	0	13	13
146	30	4.9	4.80	-.01	.04	1.0	0	1.0	0	17	17
139	30	4.6	4.54	-.02	.04	1.0	0	0.9	0	55	55
136	30	4.5	4.43	-.03	.04	0.9	0	0.9	0	32	32
109	30	3.6	3.48	-.08	.05	0.9	0	0.8	0	47	47
104	30	3.5	3.31	-.09	.05	1.2	0	1.1	0	3	3
106	30	3.5	3.37	-.09	05	1.0	0	1.2	0	30	30
101	30	3.4	3.20	-.10	05	1.1	0	1.4	1	40	40
79	30	2.6	2.46	-.15	05	1.2	1	1.7	1	16	16
31	30	1.0	0.94	-.33	.08	1.4	0	1.0	0	2	2
226.7	30.0	7.6	7.61	.26	.09	0.9	-0.0	0.9	-0.1	mean(cou.=55)	
67.2	0.1	2.2	2.30	.27	.08	0.3	0.7	0.3	0.8	S.D.	

RMSE (Model): .12 Adj. S. D.: .24 Seperation: 1.95 Reliability: .79

Fixed (all same) chi-square: 398.2 d.f.: 53 significance: .00

Random (normal) chi-square: 43.5 d.f.: 52 significance: .79

According to the last column in Table 2, student 19 is the most capable student (i.e., with the logit score of 2.00), and student 2 is the least capable student (i.e., with the logit score of -0.33). The infit and outfit statistics should be examined in order to check the consistency between the data and the model (Randall & Engelhard, 2009). The outfit statistic is the mean-square of the residuals between the observed data and the expected data and is quite sensitive to the unexpected extreme values (Engelhard, 1994). For instance, it is sufficiently sensitive to detect a student giving an incorrect answer to an easy question although he/she gave correct answers most of the other questions. On the other hand, the infit statistic is less sensitive to extreme values than the outfit statistic. The desired value for the infit statistic is 1. Values above 1 indicate that the data contains more variance than expected, whereas values below 1 indicate that the data contains less variance than expected (i.e., interdata dependence) (Hetherman, 2004). In the case of fit between the data and the model, the expected

value for both mean-squares is 1. The value limits mentioned in the literature for both the infit and the outfit statistics are rather similar. The acceptable values range between 0.6 and 1.5 according to Lunz, Wright, and Linacre (1990), whereas Turner (2003) reported the acceptable range as 0.5 to 1.5. Thus, students numbered 9, 14, 22, 26, 10, 12, 8, 5, 4, and 25 display infit and outfit statistics outside the acceptable values. Finally, the separation index in the last line was 1.95, and the reliability coefficient was 0.79. Note that the reliability coefficient is interpreted to be equivalent to Cronbach's alpha or to the generalizability coefficient (Nakamura, 2002). Thus, the internal consistency coefficient of the test is acceptable. The values for the items are shown in Table 3.

Table 3

Items' Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit		Outfit		Nu	Items
						MnSq	Zstd	MnSq	Zstd		
914	162	5.6	6.36	.18	.02	1.1	1	1.3	1	10	10
1108	162	6.8	7.96	.07	.02	1.0	0	1.2	0	4	4
1134	162	7.0	8.13	.06	.02	1.0	0	0.8	0	6	6
1195	162	7.4	8.50	.02	.03	1.0	0	0.8	0	9	9
1243	162	7.7	8.75	-.01	.03	1.0	0	0.7	-1	8	8
1271	162	7.8	8.88	-.03	.03	1.0	0	0.7	-1	3	3
1281	162	7.9	8.93	-.04	.03	1.0	0	0.7	0	7	7
1336	162	8.2	9.16	-.08	.03	1.4	2	1.4	1	5	5
1334	161	8.3	9.18	-.09	.03	0.9	0	0.6	-1	1	1
1352	162	8.3	9.22	-.09	.03	0.8	-1	0.5	-1	2	2
1216.8	161.9	7.5	8.51	.00	.03	1.0	0.1	0.9	-0.5	Mean	(count:10)
128.4	0.3	0.8	0.83	.08	.00	0.2	.1	0.3	1.0	S.D.	

RMSE (Model): .03 Adj. S. D.: .08 Seperation: 2.95 Reliability: .90

Fixed (all same) chi-square: 105.2 d.f.: 9 significance: .00

Random (normal) chi-square: 9.1 d.f.: 8 significance: .34

According to Table 3 and Figure 1, the most difficult item is number 10 (i.e., with the logit value of 0.18), and the easiest items are 1 and 2 (i.e., both with the logit value of -0.09). A close examination of the infit and outfit statistics of the items clearly reveals that those values are within acceptable limits for all of the items (i.e., the values are between 0.5 and 1.6). The separation index for the items was 2.95, whereas the reliability coefficient was calculated as 0.90. Upon examining the order of items according to the level of difficulty, we found that the first two items were the easiest

for the students, while the final item was the most difficult; moreover, the rest of the items were close to each other in terms of difficulty, i.e., at the medium level. When the items on a test vary in terms of difficulty level, ordering them from easiest to most difficult will help reduce students' test anxiety, thereby allowing the educator to obtain the students' real scores and raising the reliability of the scores. From this perspective, the items in the study were well organised according to the level of difficulty. The values for the raters are shown in Table 4.

Table 4

Raters' Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit		Outfit		N	Raters
						MnSq	Zstd	MnSq	Zstd		
3991	540	7.4	8.55	.01	.01	1.0	0	0.9	0	3	3
4074	539	7.6	8.70	.00	.01	0.8	-2	0.8	-1	1	1
4103	540	7.6	8.74	-.01	.01	1.2	2	0.9	0	2	2
4056.0	539.7	7.5	8.66	.00	.01	1.0	0.2	0.9	-0.8	Mean	(count:3)
47.5	0.5	0.1	0.08	.01	.00	0.1	2.2	0.0	0.2	s.d.	
RMSE (Model): .01			Adj. S. D.: .00		Seperation: .00			Reliability: .00			
Fixed (all same) chi-square: 1.5			d.f.: 2		significance: .48						

The consistency of the scores assigned by the raters is very important for the reliability of the scoring. Differences in scoring can often be observed between raters in an environment with more than one rater even when the raters have been provided with a well-designed, shared working programme to help them assign consistent scores. Since the differences in scoring are reflected in the students' scores, this creates a bias and threatens the reliability of the scores (Nakamura, 2000). From this perspective, the data in Table 4 indicates that the three raters in this study approached scoring with very similar severity/leniency. More specifically, rater 3 was the most severe (i.e., with a logit value of 0.019), whereas rater 2 was the most lenient (i.e., with a logit value of -0.01); additionally, the logit value for rater 1 was 0. In contrast to the separation indexes for the students and the items, the desirable value for the separation index for the raters is close to 0. If complete consistency exists among the raters' scoring, then the separation index will be 0 (Nakamura, 2000; Linacre, 1989). As is evident in Table 4, the separation index for the raters is 0; therefore, complete consistency was achieved with the raters in this study. In addition to these results, the statistics for the categories can be seen in Table 5.

Table 5

Category Statistics

Data				Quality Control			Step calib.		Expectation				
Sc.	*counts used*	*%*	*cm. %*	*Av. mean*	*Exp. mean*	*Out fit*	*mean*	*s.e.*	*mean catg.*	*at -.5*	*most prob from*	*Thurst. threshold at*	*Cat peak pr.%*
0	267	16	16	-.03	-.03	.8			(-.60)		low	Low	100
1	5	0	17	.03	-.01	.9	3.96	.08	-.31	-.44		-.08	1
2	18	1	18	-.01	.02	.4	-1.27	.08	-.19	-.24		-.08	3
3	25	2	19	.06	.06	.6	-.29	.08	-.11	-.15		-.06	4
4	6	0	20	.06	.09	.5	1.50	.08	-.05	-.08		-.04	1
5	162	10	30	.13	.12	.8	-3.19	.08	.00	-.03		-.04	17
6	3	0	30	.05	.16	.4	4.13	.07	.05	.03		.05	0
7	5	0	30	.24	.20	.5	-.33	.07	.11	.08		.05	0
8	37	2	33	.24	.25	.7	-1.78	.07	.19	.15		.05	3
9	41	3	35	.32	.30	1.1	.17	.06	.31	.24		.07	3
10	1050	65	100	.37	.37	1.1	-2.91	.06	(.58)	.44	.00	.08	100

The data in Table 5 shows that the frequencies of the values in the 0.5 to 10 category ranging in the 0 to 10 scoring category are very low. These mid values are not very often used in scoring. Thus, a similar study should be conducted for scoring with three categories containing the scores 0.5 to 10.

Discussion and Conclusions

The MFRM has been used in many measurement settings to simultaneously provide a considerable amount of useful information about many facets (e.g., students, items, occasions, or raters) in a single analysis (Atılgan, 2005; Baştürk, 2010; Nakamura, 2002; Nakamura, 2000; Semerci, 2011). This study used the MFRM to analyse the scores that three raters assigned to the answers given by 55 students on a statistics test containing 10 open-ended items. This enables us to gather detailed information on each facet in the assessment and to interpret the results as a whole. As demonstrated in the study, the information for each facet is accessible through separate analysis report tables, and the position of students, items, and raters in relation to each other on the logit table easily communicates information about the three facets. The extent to which separation indexes and reliability coefficients for each facet and the data as a whole yield reliable results can be examined in the analysis report tables. Furthermore, the infit and outfit statistics and the presence of any unsuitable elements within each facet can be identified. Consequently,

inappropriate items can be changed or removed from the prospective measurement tool, and when unsuitability among the raters is identified, required education can be provided to alleviate the situation (Nakamura, 2002). The data in this study indicates that the reliability was not adversely affected by any of the sources of variability. The inter-rater reliability in particular is very good. Since the infit and outfit statistics did not exceed the desired values, no unexpected responses were identified; however, too many scoring categories (i.e., 10) were used, and categories other than 0, 5, and 10 were not used very much. Therefore, if that statistics test was used again in the future, a reduction in the scoring categories available to the raters would be more appropriate.

Therefore, the MFRM makes important contributions to the analysis of test results by easily allowing the simultaneous assessment of many perspectives; moreover, it can be used in the development of measurement tools, in the organisation of the appropriate measurement circumstances, and in the provision of effective training for raters (Kim, Park, & Kang, 2012; Looney, 2012; Nakamura, 2000; Revesz, 2012). Because it is believed to provide important information, the use of the MFRM might be recommended when analysing the results obtained from exams in which open-ended items are used and through which important decisions concerning the students' futures are made. In addition to this, since one unique part of the MFRM is the detailed information in the rater measurement report, it can be used not only for educational settings but also for other assessment conditions with more than one rater (e.g., performance assessment in medicine, engineering, or art).

References

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

Atılgan, H. (2005). Analysis of special ability selection examination for music education department using many-facets Rasch measurement (İnönü University Case). *Eurasian Journal of Educational Measurement*, 20, 62-73.

Baştürk, R. (2010). Evaluating of research assignments with many facets Rasch measurement model. *Measurement and Evaluation in Education and Psychology*. 1(1), 51-57.

Bridgeman, B. (1992). A comparison of quantative questions in open-ended and multiple-choice formats. *Journal of Educaitonal Measurement*. 29, 3, 253-271.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*. 31, 2, 93-112.

Geer, J. G. (1988). What do open-ended questions measure? *Public Opinion Quarterly*, 52, 3, 365-371.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science,* 5 (1), 13-14.

Güler, N. & Gelbal, S. (2010 (a)). A study based on classical test theory and many facet Rasch measurement. *Eurasian Journal of Educational Research*, 38, 108-125.

Güler, N. & Gelbal, S. (2010 (b)). Studying reliability of open ended mathematics ıtems according to the classical test theory and generalizability theory. *Educational Sciences: Theory & Practice*, 10, 2, 1011-1019.

Hetherman, S. C. (2004). *An application of multi faceted Rasch measurement to monitor effectiveness of the written composition in English in the new york city department of education.* Unpublished phd. dissertation. Teacher College, Colombia University, Colombia.

Hong, L. K. (1984). List processing free responses: analysis of open-ended questions with word processor. *Qualitative Sociology*, 7, 2, 98-109.

Kim, Y., Park, I. ve Kang, M. (2012). Examining Rater Effects of the TGMD-2 on Children with Intellectual Disability. *Adapted Physical Activity Quarterly*. 29, 346-365.

Leiva, F. M., Montoro, F. J. & Martinez, T. L. (2006). Assessment of ınterjudge reliability in the open-ended questions coding process. *Quality & Quantity*, 40, 519-537.

Linacre, J. M. (2007). *A user's guide to FACETS. Rasch model computer programs*. Chicago, IL.

Linacre, J. M. (1989). *Many facet Rasch measurement*. Unpublished phd. dissertation. University of Chicago, Chicago.

Looney, M. A. (2012). Judging Anomalies at the 2010 Olympics in Men's Figure Skating. *Measurement in Physical Education and Exercise Science*. 16, 55-68.

Lunz, M. E., Wright, B. D. & Linacre, J. M. (1990). Measuring the ımpact of judge severity on examination scores. *Applied Measurement in Education*. 3, 4, 331-345.

Nakamura, N. (2002). Teacher assessment and peer assessment in practice. *Educational Studies*, 44, 143. 204-215.

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

Özçelik, D. A. (1998). *Ölçme ve değerlendirme* [Measuremet and evaluation] (2nd ed.). ÖSYM Yayınları, Ankara.

Randall, J. & Engelhard, G. Jr. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement*, 46, 1, 1-18.

Revesz, A. (2012). Working Memory and the Observed Effectiveness of Recasts on Different L2 Outcome Measures. *Language Learning*. 62, 1, 93-132.

Semerci, Ç. (2011). Mikro Öğretim Uygulamalarının Çok-Yüzeyli Rasch Ölçme Modeli ile Analizi [Analyzing microteaching applications with many-facet Rasch measurement model]. *Eğitim ve Bilim/ Education and Science*, 36 (161), 14-25.

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

Turner, J. (2003). *Examining on art portfolio assessment using a many facet Rasch measurement model*. Unpublished phd. dissertation. Boston College, Boston.

Açık Uçlu İstatistik Maddelerine Verilen Cevapların Çok Yüzeyli Rasch Modeli ile Analizi

Atıf:

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research*, 55, 73-90. http://dx.doi.org/10.14689/ejer.2014.55.5

Özet

Problem Durumu: Eğitimde kullanılan ölçme araç ve yöntemlerinden biri de açık-uçlu maddelerdir. Açık-uçlu maddeler, öğrencilerin cevaplarını kendi ifadeleriyle özgürce aktarabilmelerini sağlarken diğer bazı ölçme araçlarıyla analiz edilemeyen eksikleri/hataları analiz edebilmeyi de mümkün kılar. Açık-uçlu maddeler öğrencilerin düşünme ve strateji kurabilme sürecinin izlenmesini; öğrencinin bilgi düzeyinin ve bilgiyi nasıl yapılandırdığının daha geçerli şekilde anlaşılabilmesini sağlar. Açık-uçlu maddelerin, çoktan seçmeli maddelere göre başlıca üç avantajı bulunmaktadır: 1. Şans başarısını ortadan kaldırarak bu sebeple oluşacak ölçme hatasını azaltıp; daha güvenilir sonuçlara ulaşılmasını sağlar. 2. Çoktan seçmeli maddelerde öğrenci, doğru cevabı seçeneklerden giderek de bulabilmektedir. Ancak bu tür bir sağlama yapılarak doğru cevaplamak, açık-uçlu maddelerde mümkün değildir. Örneğin; 2(X+4)= 38-X eşitliğinde X değerinin bulunmasında; öğrenci seçeneklerde verilen değerleri denklemde yerine koyarak doğru cevabı bulabilir. Halbuki öğrenciden bilmesi istenilen çözüm yolu bu değildir. Bu durum, ölçmenin istenilen yapıdan farklı bir yapıyı ölçmesine sebep olacaktır ki bu da testin yapı geçerliğinin düşmesine yol açar. Açık-uçlu maddelerde yapı geçerliğini tehdit eden bu tür bir faktör bulunmamaktadır. 3. Çoktan seçmeli maddelerin doğasında yer alan istenmeyen düzeltici dönütün yapılmasına izin vermez. Öğrenci, doğru cevabı seçeneklerde bulamayınca soruya tekrar dönüp yeni bir stratejiyle cevabı bulma yoluna gitmektedir. Açık-uçlu maddelerde bu tür bir durum söz konusu değildir.

Özellikle sentez ve değerlendirme gibi üst düzey bilişsel davranış basamaklarının geçerli bir şekilde ölçülebilmesini sağlayan açık-uçlu maddelerin en önemli

dezavantajı ise puanlanmasıdır. Açık-uçlu maddelere verilen cevapların objektif puanlanmasındaki güçlük, elde edilen puanların güvenirliğini düşüren önemli sebeplerden biridir. Açık-uçlu maddelere güvenirliği etkileyen farklı hata kaynakları da bulunmaktadır. Sınavın bütününü oluşturan her bir maddenin iç-tutarlılığı, farklı zamanlarda aynı maddelere aynı öğrencilerin verdikleri cevaplar arasındaki tutarlığı ifade eden test-tekrar test güveniliği gibi klasik test kuramı (KTK)'nda yer alan herbir hata kaynağı için farklı güvenirlik katsayıları hesaplamak mümkündür. KTK'da tüm hata kaynaklarının ve bunlar arasındaki etkileşimin birlikte aynı anda ele alınabildiği bir yöntemle güvenirliğin hesaplanması mümkün olmamaktadır. Açık-uçlu maddelerde olduğu gibi hata kaynaklarının birden fazla olduğu ölçme durumlarında KTK'nın bu sınırlılığını ortadan kaldıran genellenebilirlik ve madde tepki kuramlarının (MTK) kullanılması tercih edilmektedir.

Bu çalışmada, açık-uçlu istatistik maddelerinden alınan puanların güvenirliği; öğrenci, madde ve puanlayıcı yüzeyinin birlikte ele alındığı MTK'da yer alan, Rasch modelinin bir uzantısı olan çok yüzeyli Rasch modeli (ÇYRM) kullanılarak incelenmiştir.

Araştırmanın Amacı:Bu çalışmada, açık-uçlu maddelerden oluşan istatistik sınavı puanlarının ÇYRM analiziyle herbir yüzey (öğrenciler, maddeler ve puanlayıcılar) için uyum indeksleri ve güvenirlik katsayılarının bulunması, sonuçlar doğrultusunda puanların güvenirliğinin yorumlanması amaçlanmıştır.

Araştırmanın Bulguları: Araştırmada yer alan 55 öğrencinin 10 maddeye verdiği cevapların üç puanlayıcı tarafından puanlanmasıyla elde edilen veriler logit cetvelle incelenmiştir. Bu cetvelde tüm yüzeylerin sonuçlarını ortak bir doğrusal ölçek üzerinde görmek mümkündür. Cetvelde, öğrenci sütunu incelendiğinde, -1'den 1'e maddeleri en az doğru cevaplayan öğrencilerden en çok doğru cevaplayanlara doğru bir sıralama yer almaktadır. Böylece, en az başarı gösterenin 2. (logit puanı -0.33); en yüksek başarı gösterenlerin 19. (logit puanı 2.00), 9. ve 14. öğrenciler olduğunu açıkça görmek mümkündür. Maddelerin yer aldığı sütunda da -1'e en yakın madde, güçlük düzeyi en düşük (en zor) iken; 1'e yaklaştıkça maddelerin güçlük düzeyleri artmakta (en kolay)'dır. Böylece, en zor 1. ve 2.; en kolay 10. maddenin olduğu görülmektedir. Puanlayıcı sütununda -1'den 1'e; en cömert puan verenden en katı puanlayıcıya doğru bir gidiş söz konusudur ve üç puanlayıcının da puanlamadaki katılık-cömertlik düzeylerinin aynı olduğu (0 logits düzeyinde) görülmektedir. Verilerin, modele uyumunu iç ve dış uyum istatistikleri göstermektedir. Dış-uyum, gözlenen ile beklenen veriler arasındaki artıkların kareler ortalamasıdır ve beklenmedik uç değerlere karşı oldukça duyarlıdır. İç-uyum ise dış-uyuma göre uç değerlere karşı daha az duyarlıdır. İç-uyum için istenilen değer 1 olup; daha büyük değerler verilerin beklenenden daha fazla değişim gösterdiğini, daha küçük değelerse beklenenden daha az değişim olduğunu (veriler arası bağımlılık) gösterir. Verilerin modele uyumlu olması durumunda her iki kareler ortalaması için de beklenen değerler 1'dir. Alan yazında uyumun olduğunu söyleyebilmek üzere; dış ve iç uyum için belirtilen sınır değerler çok büyük farklılıklar göstermemektedir. Kabul edilebilir değerler (0.6, 1.5) ya da (0.5, 1.5) aralığında yer almaktadır. Buna göre; 9, 14, 22, 26, 10, 12, 8, 5, 4, 25 numaralı öğrenciler kabul edilebilir sınırların dışında iç ya da dış

uyum değerleri göstermişlerdir. Son olarak, ayırma indeksinin 1.95 ve güvenirlik katsayısının .79 olduğu görülmüştür. Buradan, testin iç-tutarlılık katsayısının kabul edilebilir düzeyde olduğu söylenebilir.

Maddeler için elde edilen analiz sonuçları incelendiğinde (Şekil 1), en zor 10. (logit değeri 0.18), en kolay 1. ve 2. maddeler (logit değeri -.09) dir. Maddelerin iç ve dış uyumları incelendiğinde, tüm maddelere ilişkin bu değerlerin kabul edilebilir sınırlar (0.5, 1.6) içinde yer aldığı görülmektedir. Maddelere ait ayırma indeksi 2.95, güvenirlik katsayısı .90 olarak bulunmuştur. Bu bilgiler dışında, maddelerin güçlük düzeylerine göre sıralaması incelendiğinde, ilk iki maddenin öğrencilere en kolay, en son maddenin en zor geldiği ve diğer maddelerin güçlük düzeyleri açısından birbirine yakın ve orta düzeyde olduğu görülmektedir. Sınavlarda yer alan maddelerin güçlük düzeyleri farklılık gösterdiğinde, maddelerin kolaydan-zora doğru sıralanması öğrencilerin sınav kaygısının düşmesine yardımcı olacak; bu durum öğrencilerin gerçek puanlarını görebilmemize dolayısıyla puanların güvenirliğinin artmasına katkı sağlayacaktır. Bu açıdan incelendiğinde, maddelerin güçlük düzeylerine göre iyi organize edildiği söylenebilir.

Puanlayıcı ölçme raporu incelendiğinde, üç puanlayıcının puanlama açısından birbirine çok yakın cömertlik-katılık düzeyinde olduğu söylenebilir. Puanlamada en katı 3. (logit değeri 0.01), en cömert 2. (logit değeri -0.01) puanlayıcının ve 1. puanlayıcının 0 logit değerine sahip olduğu görülmektedir. Puanlayıcılara ilişkin ayırma indeksinin, öğrenciler ve maddeler için elde edilenlerin aksine 0'a yakın bir değer alması istenir. Puanlayıcıların puanlamaları arasında tam bir tutarlılık söz konusuysa, ayırma indeksi 0 olacaktır. Puanlayıcıların ayırma indeksi 0 olup, puanlayıcılar arasında tam bir tutarlılığın olduğu söylenebilir.

Ayrıca, çalışmada yer alan puan kategorilerine ilişkin yer alan raporda kategoriler için frekans ve yüzde değerleri elde edilmiştir. Buna göre; 0, 5 ve 10 puanları arasında kalan değerlerin frekansları oldukça azdır. Bu ara değerlerin puanlamada çok kullanılmadığı söylenebilir. Böylece, benzer bir çalışmanın sadece 0, 5 ve 10 puanlarından oluşan üç kategori üzerinden düzenlenmesinin uygun olacağı söylenebilir.

Sonuç ve Öneriler: ÇYRM analizi, ölçmedeki farklı yüzeylere ilişkin hem tek tek ve ayrıntılı bilgi edinmemizi sağlarken hem de bir bütün olarak elde edilen sonuçların yorumlanmasına izin vermektedir. Herbir yüzeye ilişkin bilgiler ayrı ayrı sunulan analiz tablolarıyla incelenebilirken, yüzeylerin ortak logit-cetveliyle tüm yüzeylerin birbirine göre durumu aynı anda kolayca görülebilmektedir. Analiz rapor tablolarında, herbir yüzey için ayırma indeksi ve güvenirlik katsayısı değerleriyle verilerin bir bütün olarak ne ölçüde güvenilir sonuçlar verdiği; iç ve dış uyum katsayılarıyla herbir yüzeyde yer alan elemanlar içinde uyumsuzluk gösterenlerin olup olmadığı teşhis edilebilmektedir. Böylelikle, maddeler arasında uyumsuzluk gösteren bir maddenin daha sonraki ölçmeler için düzeltilmesi/ölçme aracından çıkarılması, puanlayıcılar için gözlenen uyumsuzluk durumunda puanlayıcılara gerekli eğitim programlarının düzenlenmesi sağlanabilir. ÇYRM'nin ölçme sonuçlarının pek çok yönüyle aynı anda ve kolaylıkla incelenebilmesinde, ölçme

araçlarının geliştirilmesinde, uygun ölçme koşullarının düzenlenmesinde önemli bilgiler sunduğu söylenebilir. Özellikle öğrencilerin geleceğine ilişkin önemli kararların alındığı birden fazla puanlayıcının bulunduğu sınavlardan elde edilen sonuçların analizinde ÇYRM'nin kullanılması önerilebilir.

Anahtar kelimeler: Açık-uçlu maddeler (açık-uçlu sorular), Güvenirlik, Çok yüzeyli Rasch modeli