Id: 888565

Participatory Educational Research (PER)
Vol. 8(4), pp. 334-345, December 2021
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.21.93.8.4

Investigating Musical Aptitude Examination with a Many-Facet Rasch Model

Neşe Öztürk Gübeş[*]
Faculty of Education, Burdur Mehmet Akif Ersoy University, Burdur, Turkey
ORCID: 0000-0003-0179-1986

Article history
Received:
01.01.2021

Received in revised form:
25.04.2021

Accepted:
28.04.2021

Key words:
Many-facet Rasch model;
Rater reliability;
Musical aptitude examination

The aim of this study is to show how a many-facet Rasch measurement model (MFRM) can be used for quality control whilst monitoring a musical aptitude examination. The data used in this study was gathered from a musical aptitude examination which was applied in 2019-2020 academic year for selecting teacher candidates to a music education department in one public university in Turkey. In this study, the total scores of musical singing and playing exams were used. The study group of this research is consisted of 164 candidates and five specialists who rated the musical performance of candidates. A three-facet Rasch model was used including student (n=164), rater (n=5), and task (n=2). Data was gathered with fully crossed design. MFRM analysis showed good fit the data. The reliability of separation index for students was very high and it indicated that the musical aptitude examination differentiate among students in terms of their musical performance. The reliability of the rater separation index was found as 0.00 and it suggested that raters rated students' musical performance with very similar levels of severity/leniency and they were interchangeable. The results of task measure showed that musical singing task is harder than musical playing task. The results of bias analyses showed that there is no bias based on rater by task and rater by student interactions. However, student by task interaction has some bias measures.

Introduction

Performance assessment is generally used express to a variety of assessments that rest on observation and judgement. In performance assessment, a rater or an assessor usually observes a performance and judges its quality. Ratings, although they constitute a rich source of data for decision makers, are unfortunately exposed to subjectivity (Myford & Wolfe, 2003). Therefore, consistency of the scores given by different raters has to be determined.

A performance assessment is very important in music education. Psychomotor behaviours that are singing and playing constitute musical performance (Atak Yayla, 2004). Classical Test Theory (CTT) measurement models are often used to assess consistency estimates of musical performance (Wesolowski, Wind & Engelhard, 2016). Although reliability estimates in CTT

[*] Correspondency: nozturk@mehmetakif.edu.tr

are powerful for determining the general quality of test scores in questions, it has some limitations: (1) only one measurement error can be determined in each CTT estimate and effects of multiple source of error cannot be estimated at one time (2) CTT behaves all errors to be random therefore systematic measurement error cannot be separated from random measurement error in CTT reliability estimates (3) CTT provides a sole standard error of measurement estimate for all individuals (Weir, 2005).

On the other hand, item response theory (IRT) models the probabilistic distribution of individuals' success at the item level and it focuses on the item-level information. IRT includes group of models and defines the correspondence between latent variables and their manifestation. For dichotomously scored test items there are three IRT models: one-parameter, two-parameter and three-parameter IRT Models (de Ayala, 2009; Fan, 1998). In the one-parameter model it is assumed that item difficulty is purely an item feature that influences examinee performance. The one-parameter logistic (1PL) model is also called Rasch model (Hambleton, Swaminathan & Rogers, 1991). Rasch model is mathematically equivalent to 1PL model. In the Rasch model the discrimination parameter is fixed at a value of 1.00 for all items whereas the constant of discrimination value in the 1PL model does not have to be equal 1.0 (de Ayala, 2009). The equation for the Rasch model is (Baker, 2001):

$$P(\theta) = \frac{1}{1+e^{-1(\theta-b)}} \qquad (1)$$

In the equation above, *b* is the difficulty parameter and *Θ* is the ability parameter. In fact, Rasch model has many advantages. One of the most important advantage it has measurement invariance. If the given data fit the Rasch model, examinee measures (such as item, task or rater measures) are invariant (or sample free) across different groups of examinees (Eckes, 2009).

Many-facet Rasch measurement models (MFRM) are extensions of the one-parameter Rasch model (Andrich, 1988; Rasch, 1980; Wright and Masters, 1982 as citied in Engelhard & Myford, 2003). Rasch model was generalized by Linacre (1989) to examine the quality of judgments about students' performance assessment which may cover multiple facets (i.e., rater severity, item difficulty and student ability). Through maintaining the same mathematical properties, the Rasch model has been extended to rating scale and partial credit scores. The rating scale model is used for analysing attitude scales and rated assessments. The equation for three-facet model (student ability, item difficulty and rater severity) can be written as (Eckes, 2009; Engelhard, 1994; Linacre, 2020):

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - a_j - \tau_k \qquad (2)$$

where:
p_{nijk}: probability of examinee *n* receiving a rating of *k* on criterion *i* from rater *j*,
p_{nijk-1}: probability of examinee *n* receiving a rating of *k-1* on criterion *i* from rater *j*,
θ_n: proficiency of examinee *n*,
β_i: difficulty of task *i*,
a_j: severity of rater *j*,
τ_k: difficulty of receiving a rating of k relative to a rating k-1.

MFRM is used for analysing rating data, summing up rating patterns regarding main effects of group-level for variables (or "facets") of ratings. Using MFRM approach researchers can look at individual –level effects of the elements in each facet. Various facets are analysed simultaneously but statistically independently and calibrated onto the logit scale (Engelhard & Myford, 2003). MFRM also corrects each observed score for the presence of systematic measurement error. For example, if we want to consider the impact of rater severity on examinee ability estimates, ability estimates are signified as a function the facet elements that produced them. A student who is rated by two lenient raters will have an ability estimate which is adjusted to reflect the raters' leniency. Another advantage of the MFRM is that it can show and correct the systematic error related with interactions between facets. For instance, we can introduce a rater-by-group interaction term in the model for determining the degree which a rater rates a particular group differently (Wolfe & Dobria, 2008).

In the evaluation of musical performances, raters are exposed to fast, real-time decision making processes due to immediate reactions to trait deductions (Thompson, Williamon & Valentine, 2007). Therefore, musical evaluation processes are affected by rater discernment and holistic paradigms. However, it is very important to evaluate music performances objectively and music performance assessments are arranged with the purpose to evaluate both processes and products of performance systematically and objectively (Wesolowski et al., 2016). In literature, CTT methods (such as Pearson's r, Cohen's kappa, Spearman correlation coefficient) were used to determine rater reliability of music performance in numerous studies (Birel & Albuz, 2014; Dalkıran, 2008; Ece & Kaplan, 2008; Engur, Çeliktaş, Demirbatır, 2015; Gün & Demirtaş, 2015; Kurtuldu, 2010; Öztürk & Güdek, 2016). Nevertheless, there are a few ones (Akın & Baştürk, 2012; Atılgan, 2005; Girgin, 2020; Köse, Acay Sözbir & Kalender, 2016) for which MFRM analysis were used to evaluate performance in music education. The aim of this study is to show how a many-facet Rasch measurement model (MFRM) can be used for quality control when monitoring a musical aptitude examination.

Method

Research Design

In this study, a three-facet Rasch model was used including student, rater, and task for quality control for the musical aptitude examination. Since the existing case was described as it is and without any effect this study is a descriptive research (Karasar, 2005).

Data Source

The data used in this study was gathered from a musical aptitude examination which was applied in 2019-2020 academic year for selecting teacher candidates to a music education department of one public university in Turkey. The aptitude examination was conducted in three fields: musical hearing-reading-writing, musical singing and musical playing.

Musical hearing-reading-writing field examination

The musical aptitude examination was conducted with three subfields: musical hearing, reading and writing fields. Students were asked to (a) write two different melodies with their measures, notes and period in musical writing (dictate) exam; (b) decode the melody which was determined by the music commission in musical reading exam; c) repeat two melodies which are played on the piano with their voice in melody repetition exam. The score scale for evaluating the musical hearing-reading-writing field is shown in Table 1.

Table 1. The score scale for evaluating the musical hearing-reading-writing field

Musical Writing		Musical Reading (Decode –solmization)	Melody Repetition		Total
Tonal dictation 4x6=24	Tonalite dictation 4x6=24	4x5=20	4x4=16	4x4=16	100

As seen in Table 1, the maximum score for musical writing, reading and melody repetition exams are as respectively 48, 20, 32 and totally 100 points.

Musical singing field examination

In the musical singing field examination, attributes and skills of candidates' voice usage are measured. The musical singing field is evaluated regarding three subfields: quality and capacity of voice, technique and musical. The score scale for evaluating the musical singing field is shown in Table 2.

Table 2. The score scale for evaluating the musical singing field

	Quality and capacity of the voice	Technique	Musical	Total
Score	40	30	30	100

As seen in Table 2, maximum scores for quality and capacity of voice, technique and musical parts are respectively 40, 30, 30 and totally 100. While evaluating the quality and capacity of the candidate's voice four criterions are considered: (1) rotundity, (2) compass, (3) timbre, and (4) health and cleanness of voice. The technique dimension of the musical singing field examination is evaluated based on six criterions, these are: (1) to vocalize the musical work with correct posture, (2) to vocalize the musical work correctly, (3) to vocalize the musical clearly within its tonality, (4) to vocalize the musical work metrically and rhythmically correct, (5) to vocalize the musical work with correct articulation, (6) to vocalize the musical work clearly and understandably. Lastly, the musical field is evaluated with regarding five criterions: (1) to sentence correctly, (2) to vocalize the musical work with its original – authentic speed, (3) to vocalize the musical work as a whole, (4) to interpret the musical work with a style which is appropriate to its dynamics, (5) to vocalize the musical work meaningfully and appropriately to its character.

Musical playing field examination

The candidates' playing skills and attributes are measured in musical playing exam. Candidates can play instruments such as a piano, guitar, mandolin, zither, violin, viola, violoncello, recorder and alike. The musical work which is chosen by candidates should be appropriate to the instrument's playing characteristics and its composer and voice tone should be definite. The score scale for evaluating the musical playing field is shown in Table 3.

Table 3. The score scale for evaluating the musical playing field.

	Technique	Musical	Total
Score	60	40	100

As seen in Table 3, students can get maximum 60 points from technique and 40 points from musical; in total they can get maximum 100 points. In the musical playing field, six criterions are considered while evaluating the technique dimension: (1) to play the instrument with the correct grip, (2) to apply right and left hand technique correctly, (3) to play the musical work correctly, (3) to play the musical work cleanly within its tonality, (4) to play the musical work metrically and rhythmically correct, (5) to play the musical work with correct articulation. Lastly, while evaluating the musical dimension these four criterions are considered: (1) to vocalize the musical work within its original-authentic speed, (2) to sentence the musical work correctly, (3) to vocalize the musical work as a whole, (4) to interpret the musical work suitable for its dynamics and character style.

In this study, the total scores of musical singing and playing fields of musical aptitude exam were used. The reason for excluding the total scores of musical hearing-reading-writing exam is that raters gave joint scores for each candidate's performance.

Study Group

The study group of this research consisted of 164 candidates who applied to musical aptitude examination and five specialists in music education. Originally there were six specialists in music education who rated 165 candidates musical performance. As two raters rated the performances together (not independently) only one rating of them was used in this study. One of students' scores was not obtained so the study was conducted with 164 students and 5 raters.

Data Analysis

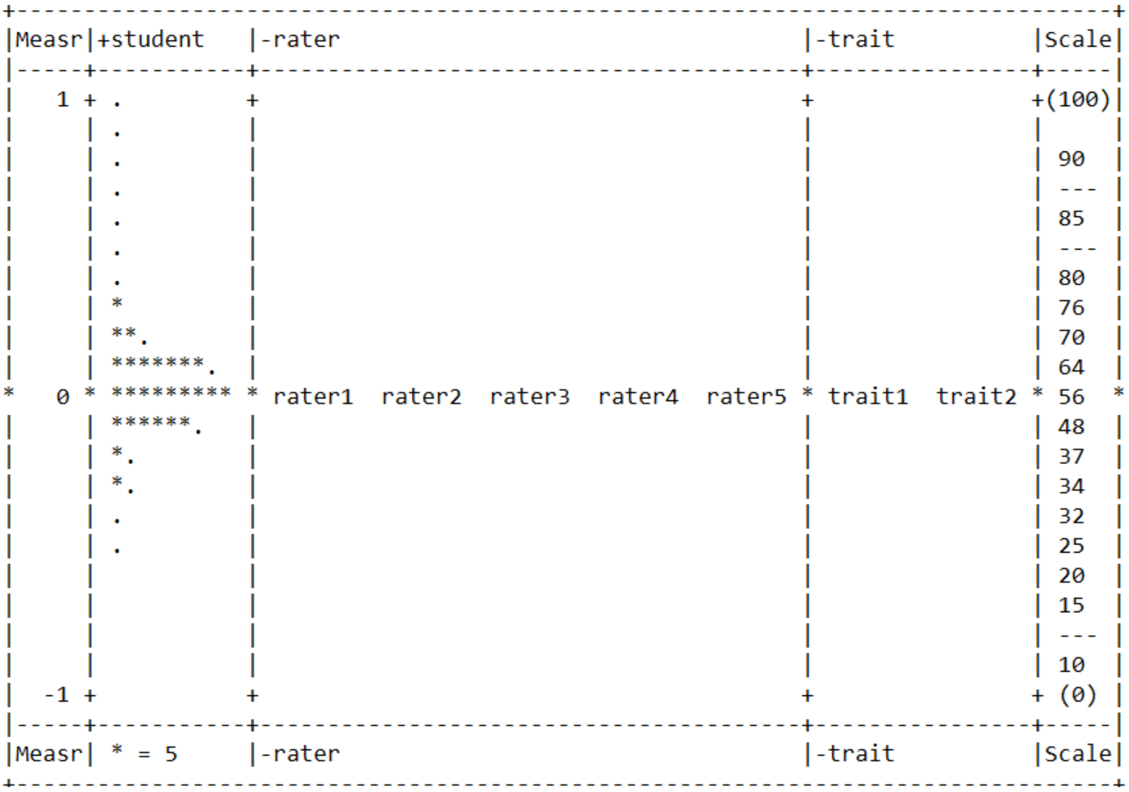
To estimate students' ability, task difficulty, rater severity and bias of the scores were analysed using Minifac, Version 3.83.3 (Linacre, 2020), student version of FACETS computer program. A three-facet Rasch model was used including student (n=164), rater (n=5), and task (n=2). Rating scale model (Andrich, 1978) was used in the analyses. Data was gathered with fully crossed design. Each of 164 students' musical performance was rated by five raters on two tasks (musical singing and playing).

Bias analyses were conducted for all two-way interactions of three facets including rater x task, rater x student and student x task interactions. The rater x task interaction was conducted for determining whether raters rated both of the tasks in the same severity or leniency. The rater x student interaction was conducted for determining whether raters rate some students more severely or leniently than others. The student x task interaction was conducted for determining whether the student responded consistently to the task in a way which is both different from other students and different from his/her own behaviour in relation to other tasks (Haiyang, 2010).

Results

The data-model fit was investigated by checking the standardized residuals. When the data fit the model, the 5% of the standardized values should not exceed ±2 interval and %1 of the standardized values should not exceed ±3 interval (Linacre, 2020). The results showed that 2.20% of the total standardized values [36 out of 1640 (164x2x5 data] are outside ±2 interval and there are not any standardized values outside ±3 interval. It can be said that model-data fit of this study is satisfactory for further analyses.

Figure 1 shows the variable map for the three facet crossed design. The variable map displays relative abilities of students, the relative severity of raters, the relative difficulties of traits and the scale steps.



```
+---------------------------------------------------------------------------------+
|Measr|+student   |-rater                                  |-trait          |Scale|
|-----+-----------+----------------------------------------+----------------+-----|
|   1 + .         +                                        +                +(100)|
|     | .         |                                        |                |     |
|     | .         |                                        |                | 90  |
|     | .         |                                        |                | --- |
|     | .         |                                        |                | 85  |
|     | .         |                                        |                | --- |
|     | .         |                                        |                | 80  |
|     | *         |                                        |                | 76  |
|     | **.       |                                        |                | 70  |
|     | *******.  |                                        |                | 64  |
*   0 * ********* * rater1  rater2  rater3  rater4  rater5 * trait1  trait2 * 56  *
|     | ******.   |                                        |                | 48  |
|     | *.        |                                        |                | 37  |
|     | *.        |                                        |                | 34  |
|     | .         |                                        |                | 32  |
|     | .         |                                        |                | 25  |
|     |           |                                        |                | 20  |
|     |           |                                        |                | 15  |
|     |           |                                        |                | --- |
|     |           |                                        |                | 10  |
|  -1 +           +                                        +                + (0) |
|-----+-----------+----------------------------------------+----------------+-----|
|Measr| * = 5     |-rater                                  |-trait          |Scale|
+---------------------------------------------------------------------------------+
```

Figure 1. Variable map for three facets

In the first column of the map, there is a logit scale and all measures of students, raters and traits are placed on this scale (Eckes, 2009). The second column (labeled "students") shows the ability estimates of students on the musical aptitude examination and each star represent five students and a dot symbolize one or two students. While higher scoring students are appearing at the top of the variable map, lower-scoring students are appearing at the bottom of the variable map. The third column (labeled "rater") of the variable map in Figure 1 shows the distribution of rater severity or leniency measures when rating the performance in students' musical aptitude examination. We can say that all of five raters performed at the same level of severity/leniency. As seen from the variable map, they all scored at the level of 0 logits. The fourth column (labeled "traits") compares the two traits (musical singing and playing) in terms of their relative difficulties. As can be seen, both of traits are at a similar level of difficulty and their difficulties are at the level of 0 logits.

More detailed measurement results on each of three facet's measurement report are presented in Table 4 through Table 6.

Table 4. Summary of students' measurement report

				Infit		Outfit	
M (SE)	SD (SE)*	Min	Max	M	SD	M	SD
0.05 (0.06)	0.23 (0.03)	-0.45	0.98	0.88	0.94	0.88	0.94

RMSE (Model): 0.06 Adj. S.D.: 0.22 Separation: 3.43 Strata: 4.91 Reliability: 0.92
Fixed (all same) chi-square: 123.9 d.f.: 162 significance: 0.99

*SD refers to the spread of scores between students. SE refers to the spread of estimates for a student.

In Table 4, the summary results of the students' facet are reported. It indicates that students' ability ranged between -0.45 logit and 0.98 logit, with a mean of 0.88 and standard deviation of 0.94. The mean SE is 0.06 and SE shows the precision of the estimates of students' ability. The relatively low SE is result of the dataset included more than one score for each test taker. The chi-square test is significant at p< 0.001 and it means that students varied related to ability being measured. The students' separation ratio of 3.43 in Table 4 indicates that the spread of students' performance is 3.43 times larger than the precision of those measures (Myford & Wolfe, 2004). In other words, the variance among students is importantly bigger than the error of estimates and the test distinguishes 164 students into three distinct levels in terms of the ability being measured (Barkaoui, 2013). The reliability of separation index for students is 0.92 and this index is similar to coefficient alpha (Myford & Wolfe, 2004), which indicates that the assessment distinguishes between students in terms of being measured (Bond & Fox, 2001). This index implies that how well the elements within student facet are separated in order to define reliably the facet (Engelhard & Myford, 2003). It also implies that raters could reliably distinguish among the students. The students are well differentiated related to their performance level.

The other information in Table 4 is a summary of fit statistics for student facet. The mean fit (0.88) is closed to the expected value of 1.00 (SD=0.94). The mean of outfit statistic is also 0.88 with standard deviation of 0.94. In Rasch analyses, infit and outfit statistics are reported as mean squares in the form of chi-square statistics divided by their degrees of freedom. The mean square fit statistics show the compatibly of the data with the model (Bond & Fox, 2001). The desired value for the infit statistic is 1.00, if the observed data fit the model. Linacre (2004) indicated that the range for fit statistics between 0.5 and 1.5 is adequate for measurement. In this study, Linacre's (2004) ranges for fit statistics were taken in consideration, based on mean of infit and outfit statistics, and it can be said that model is compatible with the data.

Table 5. Rater measurement report

	Measure	Model SE	Infit	Outfit	Corr. PrBis
Rater 1	0.00	0.01	1.01	0.90	0.70
Rater 2	0.00	0.01	0.93	0.83	0.71
Rater 3	0.01	0.01	0.97	0.85	0.71
Rater 4	-0.01	0.01	1.05	0.91	0.69
Rater 5	0.00	0.01	1.02	0.89	0.70
Mean (n=5)	0.00	0.01	1.00	0.88	0.70
SD	0.01	0.00	0.05	0.03	0.01

RMSE: 0.01 Adj (True) S.D.:0.00 Separation= 0.00 Strata=0.33 Reliability (not inter rater)=0.00
Fixed (all same) chi-square: 1.9 d.f.: 4 significance: 0.74
Inter-rater agreement opportunities: 3280 Exact agreements: 1569 = 47.8% Expected: 393.9 = 12%

Note. SE= Standard error. Infit and outfit statistics are mean-square statistics.

Table 5 reports results for the rater facet. It shows that difference in severity between Rater 1, Rater 2 and Rater 5 is 0.00 logit and the maximum difference is 0.02 and it is between Rater 3 and Rater 4. The chi-square value of 1.9 with degrees of freedom 4 is not significant (p>0.05) and the non-significant chi-square test indicates that raters were equal in severity when evaluating students (Myford & Wolfe, 2004). The rater separation ratio is 0.00, it measures the spread of the rater severity measures relative to the precision of those measures and it means that there is not any difference between rater severities. The reliability of the rater separation index is 0.00 and it suggests that raters were rating at very similar levels of severity and they were interchangeable. This reflects undesirable variation between raters in levels of severity and close to zero value is desirable for it (Myford & Wolfe, 2004).

Table 5 reports observed and expected percentages of exact rater agreement. The percentage of observed exact agreement between raters is 47.8 and it is much higher than the expected rate (12%). Linacre (2020) indicates that when the observed agreement rate is approximately equal, it can be said that "raters may be behaving like independent raters". (s. 211); if the observed agreement rate is higher than the expected rate, "raters may be behaving like rating machines" (s.211). He also emphasized that this is a typical behaviour of rater psychology; they have a mental pressure to agree with the expectations of others. In this study, we can say that pressure has increased observed agreement from 12.0% to 47.8%.

Table 5 also shows fit statistics for the rater facet. When we examine the fit means-square indices for Raters 1 through 5, we see that they range from 1.05 to 0.83 and there is not any misfitting rater, they get values within 0.5 and 1.5 (Linacre, 2004). Also, raters have fit indices that are very close to the expected value of 1.00. Rater fit indices indicate that the ratings are compatible with the MFRM model. The average interrater correlation value is equal to 0.70 and we can say that the ratings of raters exhibit a medium level of agreement.

Table 6. Trait measurement report

	Measure	Model SE	Infit	Outfit
Trait 1 (Musical singing)	0.01	0.01	1.02	0.90
Trait 2 (Musical playing)	-0.01	0.01	0.98	0.86
M (n=2)	0.00	0.01	1.00	0.88
SD	0.01	0.00	0.03	0.03
RMSE: 0.01 Adj (True) S.D.: 0.01 Separation: 2.23 Strata: 3.31 Reliability 0.83				
Fixed (all same) chi-squared: 6.0 d.f.: 1 Sig. : 0.01				

Note. SE= Standard error. Infit and outfit statistics are mean-square statistics.

As seen in Table 6, musical singing which labelled as "Trait 1" (-0.01 logit) is harder than musical playing which labelled as "Trait 2" (0.01 logit). The chi-square value of 6.0 with 1 degree of freedom is statistically significant (p<0.05) which indicates that traits are significantly different from each other in terms of their difficulty. The trait separation ratio of 2.23 indicates that the spread of the trait difficulty measures is 2.23 times larger than the precision of their measures. The reliability of separation index value of 0.83 is high and suggests that raters could reliably distinguish among traits (Myford & Wolfe, 2004). Table 6 also shows fit statistics for each trait. The infit and outfit mean-square indices range from 0.86 to 1.02 and they are close to the desired value of 1.00. We can say that there is not any misfit task, as they have fit indices range between 0.5 and 1.5 (Linacre, 2004).

Results for Bias Analyses

For bias analyses; rater x task, rater x student and student x task interactions were investigated. The t-statistic is the control parameter for investigating bias. The t-statistics is obtained by dividing the bias measure by its standard error. The hypothesis "There is no bias apart from measurement error" is checked by t-statistics. If the number of observations exceeds 30, a t-statistic is normally distributed, i.e., a z-statistic. Statistically significant bias shows that the difference between the element measure for this interaction and the overall element measure is greater than the difference we would expect to see by chance. If t-statistic equal or outside to ±2, it is reported as bias (Linacre, 2020, p. 185). The summary statistics for rater x task, rater x student and student x task interactions are given in Table 7.

Table 7. Summary statistics for the bias analyses

	Type of Interaction		
Statistics	Rater x task	Rater x student	Student x task
N combinations	10	820	328
% large t-values*	0	0	31.10
Minimum t	-1.02	-1.70	-4.41
Maximum t	1.02	1.41	3.78
M	0.00	0.00	-0.03
SD	0.67	0.32	1.75

Note. *Percentage of absolute t-values ≥2.

The first bias analysis was conducted for rater x task interaction for determining whether raters rated both tasks in the same severity or leniency. As seen in Table 7, in this data set, 10 (5 rater x 2 task) possible two-way interactions were examined. The t –statistics change within ±1.02 and they are in within ±2.00. It can be said that there is no bias based on rater by task interaction. The second bias analysis was conducted for the rater x student interaction. In this data set, there are possible 820 (164 students x 5 raters) interactions. As seen in Table 7, t-statistics gets a value between -1.70 and 1.41 and they are in within ±2.00. The results showed that there is not any significant bias based on rater x student interaction. The third bias analysis was conducted for student x task interaction. The possible 328 (164 student x 2 task) interactions were examined. As seen in Table 4, the results showed that the t-statistics change within -4.41 and 3.78. Among 328 possible interactions, 102 interactions got t-statistics outside the ±2.00 range, in other words 31.10% (N=102) showed significant bias. Table 8 presents four examples of significantly biased interactions.

Table 8. Examples for significantly biased student by task interactions

Students ID	Task	Student Ability (Logit)	Task Difficulty (Logit)	Expected Score	Observed Score	Discrepancy	Error	t-score
151	Task1	-0.09	0.01	266.45	340	14.71	0.06	3.78
94	Task1	0.04	0.01	331.46	400	13.71	0.15	2.73
32	Task2	-0.09	-0.01	179.98	135	-9.00	0.13	-3.05
9	Task2	-0.17	-0.01	136.83	100	-7.37	0.15	-3.14

As shown in Table 8, with t-statistics 3.78 “Student 151” and with t-statistics 2.73 “Student 94” performed better than expected in Task1 (musical singing). On the contrary, with t-statistics -3.05 “Student 32” and with t-statistics -3.14 “Student 9” performed worse than expected in Task 2 (musical playing).

Discussion and Conclusion

An MFRM analysis is a very valuable tool for examining the effects of the different facets and their interactions on scores (Barkaoui, 2013). In the current study, the MFRM analysis was used to check the quality control of a musical aptitude examination. A three-facet Rasch analysis was used including students (n=164), rater (n=5) and task (n=2). A fully-crossed data was used in MFRM analysis and each of student performance on two tasks (musical singing and playing) was rated by five independent raters.

MFRM analysis showed good fit the data according to Linacre’s (2020) benchmarks. There were only 2.20% of the total standardized values outside the ±2 interval and there were not any standardized values out of range ±3. The chi-square statistics for students’ measure showed that students were not equal in terms of the musical aptitude and the reliability of separation index for students' was very high and it also indicated that the musical aptitude examination differentiates among students in terms of their musical performance. It also implies that raters could reliably distinguish among the students. The students are well differentiated in terms of their levels of performance. Similar results were found in Atılgan's (2005) study. He investigated the Musical Aptitude Examination for Music Education Department by using MFRM analysis. Based on high reliability index for students he concluded that the relevant musical aptitude examination differentiate among students' musical performance.

The results showed that the percentage of observed exact rater agreement was higher than the expected rate. According to Linacre's (2020) view we can say that "raters may be behaving like rating machines" (p. 211) and "this is a typical behaviour of rater psychology" (p.211). The musical aptitude examination is one of the most important performance exams for selecting students to musical education departments of universities. Therefore, the raters who are also lecturers in these music departments may have a mental pressure to agree with the expectation of other raters. The reliability of the rater separation index was found as 0.00 and this value is the most desirable value for raters (Engelhard & Myford, 2003). It suggested that raters rated students' musical performance at very similar levels of severity/leniency and they were interchangeable. The past researches in music area (Akın & Baştürk, 2012; Köse et al., 2016) displayed that there was not a significant difference between raters severity, on the other hand in some research (Atılgan, 2005; Girgin, 2020) it was found that there was a significant difference between raters severity.

The results of task measure showed that musical singing task is harder than musical playing task. Musical singing and musical playing both require a certain skill but singing requires much more courage. It is also related to the type of the related person's voice. If the person has a good voice, he/she can sing a song very well but if he/she has a poor voice it will not be as desired (Dineen, 2015).

The results of bias analyses showed that there is no bias based on rater by task and rater by student interactions. However, student by task interaction has some bias measures. The rater by task interaction indicates the degree of the particular rater by task combination deviating from the expectations of model produced (Myford & Wolfe, 2004). Based on non-significant t-values for the rater by task interaction we can say that raters do not show any misfit from expected ratings, they assigned ratings that were fairly consistent with the expected ratings for Task 1 (musical singing) and Task 2 (musical playing). The rater by student interaction investigates whether each rater appeared to show differential severity/leniency while rating students (Eckes, 2009; Myford & Wolfe, 2004). Based on non-significant t-scores of rater by student interaction we can say that raters do not show a differential severity/leniency effect while rating students. However, the results for student by task interaction shows 31.10% (n=102) significant bias measures. Based on this result, it can be said that the difficulty level of a particular task was not the same or they were significantly difficult for 31.10% students. Atılgan (2005) investigated rater by student, task by student and rater by task interaction. He found that rater by student interaction has %10.74, student by task interaction has 4.63% and rater by task interaction has 57.69% bias measures.

Music performances have many factors such as student ability level, difficulty of the task, the severity of the raters and the structure of the rating scale that can contribute to the variability of observed scores (Wesolowski et al., 2016). MFRM analysis is very powerful tool that enables researchers to calibrate all measurement facets or factors simultaneously on an equal-interval logit scale and researchers are able to evaluate the severity of the rater on the same scale as the ability of student and the difficulty of task to be rated (Myford & Wolfe, 2004). In a music performance evaluation area, the MFRM analyses can be used effectively for examining students' musical proficiency, raters' severity and tasks' difficulty.

References

Akın, Ö. & Baştürk, R. (2012). The evaluation of the basic skills in violin training by many facet Rasch model. *Pamukkale University Journal of Education*, *31*(1), 175-187.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 357-374.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage.

Atak Yayla, A. (2004, April). *Müziksel performansın ölçülmesi [Measuring musical performance]*. Paper presented at the 1924-2004 Musiki Muallim Mektebinden Günümüze Müzik Öğretmeni Yetiştirme Sempozyumu [The Symposium of Training Music Teachers from Music Teaching School to the Present Day], Isparta.

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

Baker, F. B. (2001). *The basics of item response theory* (2nd edition). College Park: ERIC Clearinghouse on Assessment and Evaluation.

Barkaoui, K. (2013). Multi-facet Rasch analysis for test evaluation. In Kunnan, A. J. (Ed.), *The companion to language assessment* (pp. 60-4-1079). US: John Wiley & Sons.

Birel, A. S. & Albuz, A. (2014). Evaluation and test of graded scoring key (rubric) prepared for performance assessment in teaching violoncello. *Atatürk University Journal of Social Sciences Institute, 18*(3), 207-281.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates Publishers.

Dalkıran, E. (2008). Measurement of performance in violin education. *Journal of Yüzüncü Yıl University Education Faculty, 5*(2), 116-136.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford.

Dineen, M. (2015). Re: Which requires more skill in music, singing or playing an instrument [Web log comment]. Retrieved from https://thesession.org/discussions/35070.

Ece, A. S. & Kaplan, S. (2008). Müzik özel yetenek seçme sınavının puanlayıcılar arası güvenirlik çalışması [The study of inter rater reliability of music special ability examination]. *Milli Eğitim [National Education],* 177, 36-49.

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.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*, 93-112.

Engelhard, G. & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model* (College Board Research Report No. 2003-1 ETS RR-03-01).

Engur, A., Çeliktaş, H., & Demirbatir, R. E. (2015). A study on testing reliability of 2013 musical aptitude test scores conducted by music department in Uludağ University. *Procedia Social and Behavioral Sciences, 197*, 821-825.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Girgin, D. (2020). An investigation of the songs created by student-teachers in music via an interdisciplinary approach based on the RASCH measurement model and MAXQDA analysis program. *International Online Journal of Education and Teaching (IOJET), 7*(4). 1661-1687.

Gün, E. & Demirtaş, H. O. (2015). International Journal of Social Science, 40, 157-165.

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 majör graduates. *Chinese Journal of Applied Linguistics, 33*(2), 87-102.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications, Inc.

Karasar, N. (2005). *Bilimsel araştırma yöntemi[Scientific research method],* Ankara: Nobel Publication Distribution.

Köse, İ. A., Acay Sözbir, S. & Kalender, C. (2016). Examination of the violin playing skills by means of Rasch model. *Journal of Abant İzzet Baysal University Faculty of Education, 16* (Silk Road Special Issue), 2339-2349.

Kurtuldu, M. K. (2010). Validity and reliability of evaluation scale in piano education directed for the grades. *Electronic Journal of Social Sciences, 9*(31), 224-232.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95–110.

Linacre, J. M. (2020). A user's guide to FACETS Rasch-Model computer programs. Available online www.winsteps.com.

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

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

Öztürk, D. & Güdek, B. (2016). An assay concerning improving the graded scoring method (rubric) for rating the violoncello performance. *Journal of Academic Music Research, 2*(3), 1-20.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristic of performance evaluation. *Music Perception: An Interdisciplinary Journal, 2*5(1), 13-29.

Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2016). *Rater analyses in music performance assessment: Application of the many facet Rasch model*. Paper presented at the 5th International Symposium on assessment in music education, Williamsburg, VA.

Weir, C. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Wolfe, E. W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 71–85). Los Angeles: Sage.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.