International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 998–1012
https://doi.org/10.21449/ijate.1058300
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Using Rasch analysis to examine raters' expertise Turkish teacher candidates' competency levels in writing different types of test items

Ayfer Sayin[1,*], **Mehmet Sata**[2]

[1]Gazi University, Faculty of Education, Department of Educational Sciences, 06500, Ankara, Türkiye
[2]Ağrı İbrahim Çeçen University, Faculty of Education, Department of Educational Sciences, Kars, Türkiye

ARTICLE HISTORY

Received: Jan. 15, 2022
Revised: Oct. 12, 2022
Accepted: Nov. 29, 2022

Keywords:
Test item,
Raters' expertise,
Many Facet Rasch,
Validity,
Reliability.

Abstract: The aim of the present study was to examine Turkish teacher candidates' competency levels in writing different types of test items by utilizing Rasch analysis. In addition, the effect of the expertise of the raters scoring the items written by the teacher candidates was examined within the scope of the study. 84 Turkish teacher candidates participated in the present study, which was conducted using the relational survey model, one of the quantitative research methods. Three experts participated in the rating process: an expert in Turkish education, an expert in measurement and evaluation, and an expert in both Turkish education and measurement and evaluation. The teacher candidates wrote true-false, short response, multiple choice and open-ended types of items in accordance with the Test Item Development Form, and the raters scored each item type by designating a score between 1 and 5 based on the item evaluation scoring rubric prepared for each item type. The study revealed that Turkish teacher candidates had the highest level of competency in writing true-false items, while they had the lowest competency in writing multiple-choice items. Moreover, it was revealed that raters' expertise had an effect on teacher candidates' competencies in writing different types of items. Finally, it was found that the rater who was an expert in both Turkish education and measurement and evaluation had the highest level of scoring reliability, while the rater who solely had expertise in measurement and evaluation had the relatively lowest level of scoring reliability.

1. INTRODUCTION

Language is the most effective means by which human beings convey their feelings and opinions. Language education is a developmental process which starts at birth – even before birth – and continues a lifetime. Thus, Turkish education programs that also constitute the basis of other disciplines are based on four fundamental skills, namely reading, writing, listening and speaking. The Ministry of National Education (MoNE) reports that "The Turkish Education Program is regarded as the development of language skills and competencies and a prerequisite to learning, personal and social development and acquisition of vocational skills" (2019). This statement indicates that language skills essentially form the basis of other disciplines. It is

*CONTACT: Ayfer SAYIN ayfersayin@yahoo.com Gazi University, Faculty of Education, Department of Educational Sciences, Assessment and Evaluation in Education, 06500 Beşevler, Ankara, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

known that teacher quality has an important role in students' reaching the learning outcomes in education programs. It is important to utilize valid and reliable tools not only to identify the extent to which students reach the learning outcomes in the program and to make decisions about students, but also to provide students with effective feedback. Thus, in the present study, the aim was to examine Turkish teacher candidates' competencies in writing different types of items to measure reading comprehension skills. With respect to the reading comprehension skill in Turkish education programs, the aim is for students to read fluently and to accurately comprehend the texts they encounter in their daily life by using the right methods, to critically interpret and evaluate what they read, and to adopt the habit of reading (MoNE, 2019). Reading comprehension skills are observed to have an important place in the Turkish language test section of exams administered within the school transitional system in Turkey. Furthermore, the importance of developing students' reading comprehension skills is also highlighted in such international test administrations as PIRLS and PISA. As in all skills and competencies, it is essential not only to equip students with reading comprehension skills but also to measure these skills in a valid and reliable way. In parallel to the changes in the expertise expected of an individual in the 21st century, the changes in teaching and learning environments should be reflected in the measurement tools as well. In other words, in an education system where the development of students' higher order skills is aimed at, measurement tools are also expected to have the quality of measuring higher order skills (Sayın & Kahraman, 2020).

During pre-service trainings, teachers receive training in writing items in accordance with item writing principles and writing items that can measure not only lower-level skills but also higher order skills. Test development includes the processes of individuals' use of knowledge, abilities, talents, areas of interest, attitudes and other characteristic expertise to develop items and transform them into a test format within the framework of a plan. It also includes the procedures of identifying the appropriate test administration conditions, how the scoring of the test performance is to be done and how the scores are to be announced to the test takers (Crocker & Algina, 2008). Even though details regarding test development, which includes numerous steps and a long process, vary in different sources (Linn & Gronlund, 2000; Walsh & Betz, 1995), test development is comprised of the following steps: identifying the purpose of the test, defining the constructs to be measured via the test, writing the items, revising the items based on expert opinion, preparing the pilot form, conducting a pilot study, scoring, item analysis, selection of items, and finalizing the test (Baykul, 2000). However, such institutions as the Higher Education Council (HEC) and MoNE in Turkey, which administer high scaled tests, are unable to conduct their pilot studies during the test development process owing to issues of confidentiality. In-class tests are also developed generally without a pilot study, based solely on expert opinion, because of the small number of participants and other reasons. In other words, the test development process is completed at the stage when items are evaluated based on expert opinion. Thus, expertise of the experts to evaluate the test items formed during test development comes forward. It is imperative that items measuring the target learning outcome be developed in accordance with measurement and evaluation principles. Even if it has a correct response, an item that is not well-structured may not serve its purpose. For this reason, it was ensured that the raters participating in the present study to evaluate the test items had diverse expertise.

Since the study aimed to determine the effect of rater qualifications in evaluating the different item-type writing skills of pre-service teachers, the multi-faceted Rasch model was used. It gives individual and group-level statistics on a single comparable scale (logit scale) (Linacre, 1993). In addition, the multi-faceted Rasch model contributes to the reliability and validity of the measurements in determining the expected effects of the variability within the scope of the research (e.g., the mutual interactions between the rater and the item type). When a multi-faceted Rasch bias analysis is performed, the researcher looks for evidence in the rater's scoring

pattern (Myford & Wolfe, 2003). The effects of rater biases, beliefs, or personal characteristics on scoring behavior can be studied using the multi-faceted Rasch measurement model approach. Similarly, the effects of the rater's past experiences on the scoring behavior can be examined. The multi-surface Rasch approach was preferred in this context in the related research.

When the rater effect is mentioned, it was examined whether the raters were experienced (Barkaoui, 2010; Davis, 2016; Erman Aslanoğlu & Şata, 2021; Kim, 2020) or the scoring rigidity within themselves (Anthony, Styck, Volpe, & Robert, 2022; Jones & Bergin, 2019; Kamış & Doğan, 2017; Primi, Silvia, Jauk, & Benedek, 2019). In this research, the effect of the field expertise of the raters was examined, which is quite significant in terms of both the examination and the result. Since it is essential that the people who will work in the test development process give information about their expertise; similarly, it is expected to contribute to the field by giving feedback on item types and seeing which item types the pre-service teachers are better.

Just as the in-class learning outcomes to be measured and their levels vary, the item types to be included in a test also vary because true-false and short response items that are appropriate for measuring all kinds of learning outcomes at lower levels may not be conducive to measuring higher order level skills (Özçelik, 2010b). Hence, including different types of items in a test to form evidence for content validity is also important. Gorin (2007) and Sireci (2007) state that for any condition of assessment, there generally needs to be more than one test and item type.

1.1. Research Questions

1. Do raters' expertise influence the process of evaluating teacher candidates' competency levels when developing test items?
2. Do Turkish teacher candidates' competencies differ when writing different test items?
3. What kind of interaction exists between raters' expertise and teacher candidates' competency levels in writing different test items?

2. METHOD

2.1. Research Model

In the present study, the relational survey design, one of the quantitative research methods, was employed. The aim in a relational survey model is to examine the existence and degree of a relationship between two or more variables without any intervention (Büyüköztürk et al., 2018; Karasar, 2018).

2.2. Study Group

The study group of the present study was comprised of 84 Turkish teacher candidates whose %71 (n=60) is female, and 29% (n=24) is male. They are at the 6th term of thecurriculum , and the teacher candidates started to write the items ten weeks after attending their measurement and evaluation course. The test items developed by the teacher candidates were scored by three raters with different expertise. One of the raters was an expert in measurement and evaluation (Rater 3), one was an expert in Turkish language education (Rater 2), and the final rater was an expert in both Turkish education and measurement and evaluation (Rater 1).

2.3. Data Collection Tools

The data collection process was performed in two stages. First, the Turkish teacher candidates were required to develop a test consisting of different types of items. Subsequently, the items produced were evaluated.

2.3.1. *Item writing*

After the 12 hours of face-to-face education that teacher candidates received during the test development unit in the measurement and evaluation course, they formed specification tables based on the learning outcomes regarding reading comprehension skill in the Turkish education program. As the curriculum is spiral in nature, there are similarities between the prescribed learning outcomes for different grade levels. After the preparation of the specifications table, the teacher candidates were asked to write the learning outcomes planned to be measured by means of true-false, short response, multiple choice and open-ended items. After matching the learning outcomes with the appropriate item type, the teacher candidates passed onto the stage of selecting texts. By its very nature, the reading comprehension skill is shaped based on the type of text used. Such expertise as length of text, style of expression and statements have a direct impact on the type and level of the item to be developed (Sayın & Takıl, 2017). The items based on the related learning outcomes that were written based on the selected or written texts in accordance with the points to be considered in text selection were written on the item writing form. The form consisted of five sections: the related learning outcome(s), text, instruction, items, and answer key. In addition, at the beginning of the form was included a section on the item writing principles to be considered for each item type. The teacher candidates wrote a total of 14 items: 5 true-false, 5 short response, 5 multiple choice and 1 open-ended. As the teacher candidates initially organized their texts, and then wrote items based on these texts, the probability of copying their items from elsewhere was minimized. Moreover, the items written by the teacher candidates were checked for originality via a software before the rating stage began.

2.3.2. *The Scoring of the items*

The test consisting of different item types and developed by the teacher candidates within the scope of this study was scored with the use of a holistic rubric developed for each test item by the researchers. Taking into consideration the qualities that test items need to possess, the researchers based the rubric on a five-point measurement scale. Each item type was scored within its own category. During the scoring stage, three experts were asked to assign a score for each item. With the aim of identifying the impact of raters' expertise on scoring, the raters' areas of expertise showed variation. The first rater (Rater 1) was an expert in both Turkish education and measurement and evaluation. The second rater (Rater 2) was an expert in Turkish education but did not have direct expertise in measurement and evaluation. The third rater (Rater 3) was an expert in measurement and evaluation but did not have direct expertise in Turkish education. Using the holistic rubric, the raters independently rated all the item types written by all the teacher candidates.

After the holistic rubric was prepared and used, data was collected for the validity and reliability of the measurements (Appendix 1). Factor analysis was utilized for the validity of the measurements, and the McDonald (1999) ω coefficient was employed for reliability purposes. Since the factor loading of each criterion is different (since the congeneric measurement is in item), the omega coefficient, which makes a more consistent estimation, was used (Osburn, 2020). Prior to an exploratory factor analysis (EFA) for validity, the underlying assumptions of this analysis need to be tested. Hence, the statistical analyses to test the assumptions revealed that the requiured minimum sample size (minimum five people per variable) was met, there was no outliers or loss of data in the data set, there was a linear relationship among the criteria of the measurement tool, and all the variables showed a normal distribution. Ater all the assumptions were found to be met, whether or not the data set could be factorized was examined, and it was revealed that it could be (for the related data set the Kaiser-Meyer-Olkin value was found to be .654, and the Bartlett's sphericity test was found to be statistically significant (χ^2(fd) = 37.411 (6), p = .000)). According to the EFA results, it was found that the

measurement tool represented a single factor structure (The variance explained was 46.67%, and the factor loadings of the criteria were 0.803, 0.653, 0.595, 0.665, respectively). After evidence for the validity of the measurements was obtained, the McDonald ω coefficient was used to assess the reliability of the measurement tool. As a result of the analysis run via the Mplus (version 8) package program, the McDonald ω coefficient was found to be .733. Based on these findings, it can be claimed that the measurements obtained from the holistic rubric used to assess the teacher candidates' competencies in writing different types of test items were valid and reliable.

2.4. Data Analysis

In the present study, which aimed to evaluate teacher candidates' competency levels in writing different types of test items, the many facet Rasch analysis (Linacre, 2012) was used as it was appropriate for the nature of the study. Since more than one variable source can be analyzed simultaneously in many facet Rasch analysis, it can be used in many different designs. In the present study, there are three dimensions (source of variability): raters, teacher candidates, and item type. All the variability sources in the study were taken into consideration, and a full factorial design, in which all the raters, all the teacher candidates, and all the item types were evaluated, was utilized. During data analysis, the guidelines defined by Myford & Wolfe (2003, 2004) were taken into consideration. In accordance with these guidelines, the statistics of the group, followed by those of the individuals, were presented. As many facet Rasch analysis is a member of the item response theory, it rests on certain assumptions that need to be met (Farrokhi, Esfandiari & Schaefer, 2012; Farrokhi, Esfandiari & Vaez Dalili, 2011). These assumptions are unidimensionality, local independence and model-data fitting. In terms of the first assumption – unidimensionality – as stated in the measurement tool section, it was identified that the holistic scoring rubric was based on a single factor; that is, it met (the) unidimensionality assumption. Since the unidimensionality of a measurement tool indicates local independence, it was accepted that the assumption of local independence was also met. Finally, the standardized residual values were examined for the model-data fitting. To meet the assumption of model-data fitting, the number of standardized residual values that do not fall within the ±2 interval must not be more than 5% of the total observation numbers. Also, it is reported that the standardized residual values that do not fall within the ±3 interval should not be more than 1% of the total number of data (Linacre, 2017). When the standardized residual values were examined, it was found that there were 51 (5.06%) values within the ±2 interval and 11 items (1.09%) within the ±3 interval, thus concluding that the model-data fitting was at an acceptable level (total number of observations 3x4x84 = 1 008).

3. FINDINGS

In the present study, which aimed to evaluate Turkish teacher candidates' competency levels in writing different types of items, initially the impact of raters' expertise on the evaluations was examined. Within this scope, the measurement reports for the rater dimension were obtained and presented in Table 1.

As can be observed in Table 1, the discrimination ratio for the group level statistics, discrimination index and discrimination index reliability values were low (<0.70). The reliability of the discrimination index is interpreted as Cronbach's alpha coefficient, and values below .70 indicate that the reliability of individuals in discrimination according to their performance is low (Marais & Andrich, 2008).

Table 1. *Measurement report for the rater dimension*

Rater	Logit	Standard error	Infit		Outfit		*t*-value	Rasch-Kappa
			MnSq	ZStd	MnSq	ZStd		
Rater 1	+0.12	0.08	0.95	-0.40	0.78	-2.10	1.50	0.44
Rater 2	+0.02	0.08	0.95	-0.40	1.09	0.80	0.25	0.34
Rater 3	-0.15	0.07	1.08	0.70	1.18	1.70	-2.14	0.31
Mean	0.00	0.08	1.00		1.01			
SD	0.14	0.00	0.07		0.21			

Model, Sample: RMSE = .08 Standard deviation = .08
Discrimination ratio=1.43 Discrimination index = 2.25
Discrimination index of reliability= 0.67
Model, Fixed (all same) chi square=6.20 *df* =2 *p*= .04
Model, Random (normal) chi square =1.50 *df* = 1 *p*= .22
Observed inter-rater agreement: 67.00%
Expected inter-rater agreement: 48.10%
Kappa inter-rater reliability statistics: 0.37

$t_{critical}$ (0.05, 2) = 4.30; $\chi^2_{critical}$ (0.05, 2) = 5.99

Thus, this indicates that the scores of the raters who evaluated the teacher candidates' competency levels in writing different item types showed slight variations. The p-value for the fixed effects chi-square value regarding the statistical variation was found to be 0.04. A chi-square value that is higher than the critical chi-square value indicates that the measurements show a statistically significant difference. In other words, it indicates that raters' expertise had an impact on the evaluations. When the t-value for each rater was examined, and since the critical t-value was observed to be small, it was revealed that the evaluations made by the raters in the study showed similarity in levels of strict versus lenient scoring.

Even though there was no statistically significant difference between the raters' lenient or strict scoring levels, the examination of each rater's Rasch-Kappa values showed that the first rater had a higher level of reliability when compared to that of the other two raters. Accordingly, it was deduced that raters' expertise had an effect on teacher candidates' competency levels in writing different types of test items. An examination of raters' expertise revealed that the rater who had expertise in both Turkish education and measurement and evaluation had the highest level of reliability in scoring. Then followed the rater with expertise in solely Turkish education. The lowest reliability in scoring among the three raters belonged to the rater who had expertise solely in measurement and evaluation.

In the process of writing different items of Turkish teacher candidates, the measurement report on the item type related to a statistical difference according to item type was examined. This measurement report by item type is presented in Table 2. As can be observed in Table 2, the discrimination ratio for item types, the discrimination index and the discrimination reliability values are very high (>0.70). Moreover, the chi square value was found to be statistically significant. Accordingly, a variation was revealed between the competency levels of the teacher candidates in writing different types of test items. In order to identify the source of this variation at the group level, the variables at the individual level were examined. Initially, the logit values were calculated for each item type; the highest and lowest logit values were found to be 0.89 and -0.82, respectively. A positive logit value indicates a high level of item writing competency, while a negative logit value indicates a low competency level. Accordingly, the Turkish teacher

candidates' competency levels in writing true-false type of items were found to be high, while their competency levels in writing multiple choice items was found to be low.

Table 2. *Measurement report for the dimension of item type*

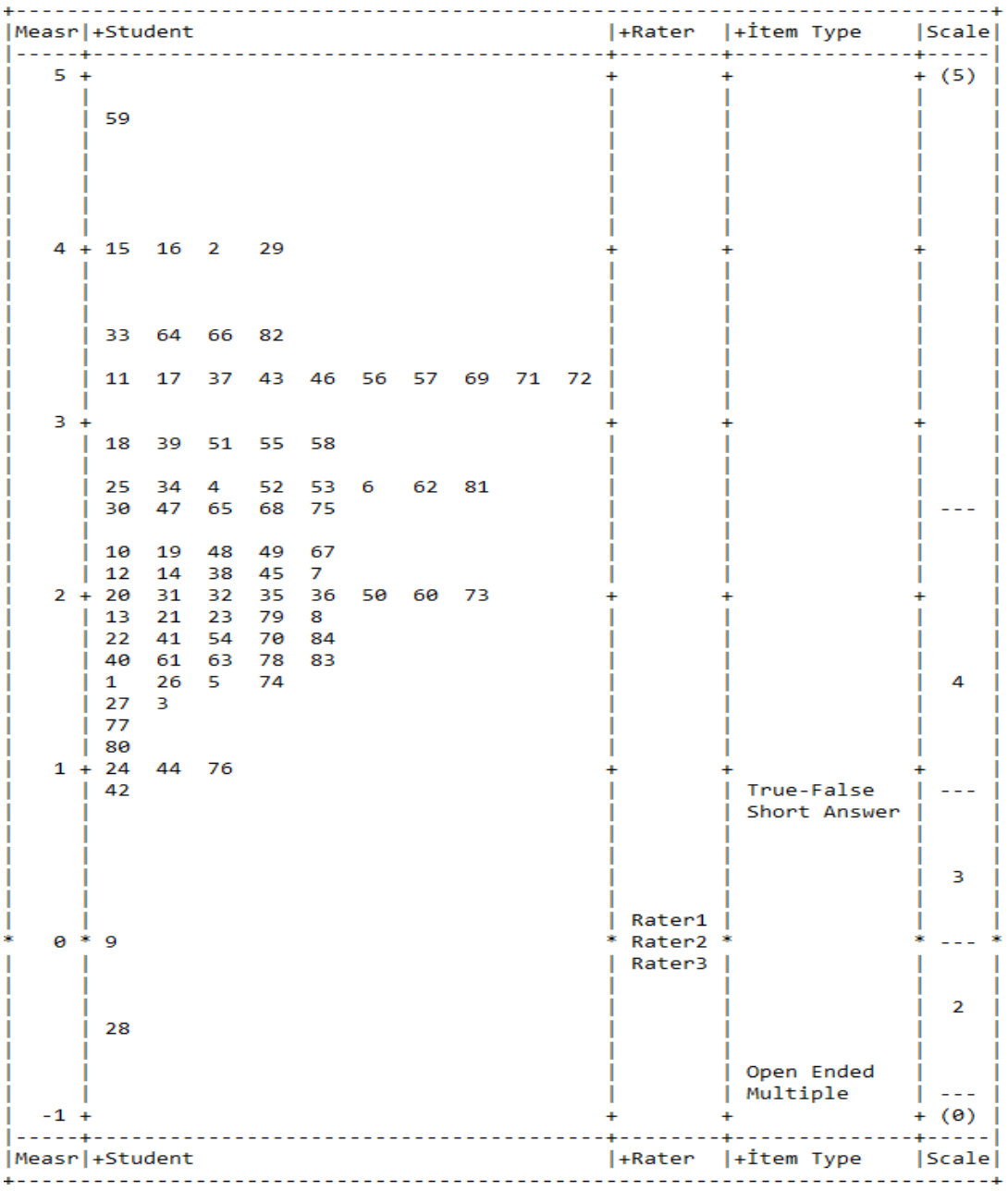
Item Type	Logit	Standard error	Infit		Outfit	
			MnSq	ZStd	MnSq	ZStd
True False	+0.89	0.12	0.87	-1.00	0.80	-1.40
Short response	+0.72	0.11	1.63	4.40	1.55	3.50
Open-ended	-0.79	0.08	0.71	-3.30	0.74	-2.90
Multiple choice	-0.82	0.07	1.06	0.60	0.97	-0.20
Mean	0.00	0.10	1.07		1.01	
Standard deviation	0.93	0.02	0.40		0.37	

Model, Sample: RMSE = .10 Standard deviation= .80
Discrimination ratio =9.44 Discrimination index =12.92
Discrimination index of reliability= .99
Model, Fixed (all same) chi square=269.10 *df*=3 *p*= .00
Model, Random (normal) chi square=3.00 *df*=2 *p*= .22

The standardized forms of the residual values were examined in order to determine in which item type the most unexpected scores were given during the raters' evaluation of different item types. The analyses revealed that there were 51 outlier values: 11 of these (21.57%) belonged to the first rater, while 19 (37.25%) and 21 (41.18%) of them belonged to the second rater and the third rater, respectively. An examination of which item type outliers were more existent revealed that there were 6 (11.76%) outliers in the multiple choice items, 7 (13.73%) outliers in the open-ended items, 8 (15.69%) outliers in the true-false items and 30 (58.82%) outliers in the short response items. Accordingly, it can be claimed that raters showed the lowest agreement in their scorings of short response items where the highest ratio of outliers were observed. That is, short response items were the most affected by raters' expertise. Appendix 2 depicts the distribution of the outliers (standardized residual values) by item type. The common map obtained by converting each of the variable sources (each dimension) addressed within the scope of the study into logit values is displayed in Figure 1.

Figure 1 shows that teacher candidates, raters, and competency levels in relation to item types were converted to the same logit measure. This common measure allows for a comparability among all variability sources. It is depicted that the most successful teacher candidate was candidate number 59, while the least successful candidate was candidate number 28. Similarly, it can be observed that while rater 1 was the most lenient scorer, rater 3 was the strictest scorer. In addition, it can be observed that the competence level for preparing true-false items was found to be high, while the competence level for preparing multiple choice items was low.

Figure 1. *Logit map of the variables in the study*



4. DISCUSSION, CONCLUSION and SUGGESTIONS

The present study aimed to utilize the Rasch analysis to examine the competency levels of Turkish teacher candidates in writing test items. In addition, the effect of the expertise of the raters who scored the items developed by teacher candidates was examined within the scope of the study. There are studies in which the tasks of teachers and prospective teachers are evaluated with multi-faceted Rash analysis (Erguvan & Aksu Dünya, 2021; Goodwin, 2016; Li, 2022). Because Rash analysis, the multi-faceted Rasch model, contributes to the reliability and validity of the measurements in determining the expected effects of the variability within the scope of the research (e.g., the mutual interactions between the rater and the item type). When a multi-faceted Rasch bias analysis is performed, the researcher looks for evidence in the rater's scoring pattern (Myford & Wolfe, 2003).

The conclusions derived from the Rasch analysis run on the data obtained from the test items developed by 84 Turkish teacher candidates and the data obtained from the 3 raters are as follows:

One conclusion that was arrived at was that raters' expertise had an impact on teacher candidates' competency levels in writing different types of items. When the raters' expertise were examined, it was observed that the most reliable scoring belonged to the rater who was an expert in both Turkish education and measurement and evaluation. Then followed the rater who was an expert in Turkish education, who was also observed to score in a reliable way (though with a lower reliability score). The least reliable rater was found to be the rater with expertise solely in measurement and evaluation. Most of the studies in the literature are those where the effect of a higher number of raters is investigated (Atılgan & Tezbaşaran, 2005; Bıkmaz Bilgen & Doğan, 2017; Kamış & Doğan, 2017). In addition, in a study by Erman Aslanoğlu & Şata (2021), it was reported that raters with similar expertise were effective in scoring items, and in a study by Kara & Kelecioğlu (2015), it was revealed that raters' expertise were effective in scoring reliability such as determining the cut-off values. In the literature, it is seen that rater qualities are examined more in the process of evaluating language skills (Song et al., 2014). In the study by Leckie and Baird (2011), it was determined that inexperienced raters were more rigid than experienced raters in assessing students' language skills. Similarly, Meadows & Billington (2010) stated that experienced that raters make more consistent assessments than others. In the study conducted by Wiseman (2012), on the other hand, students had two types of compositions, narrative and persuasion, scored by eight raters. It was determined that the scorers' scores changed according to different composition types. This result indicates that rater qualifications effectively score and support the study's results. Institutions such as the Higher Education Council and the Ministry of National Education develop and administer numerous tests, primarily tests that serve as references for the school transitional system. Owing to issues of confidentiality, institutions are unable to administer pilot studies of the test they develop and, hence, solely base their test development process on expert opinions. The present study revealed that test items should be developed by raters that have expertise both in the related subject domain and in the area of measurement and evaluation. Alternatively, the findings of the study indicate that an expert on the subject domain and an expert on measurement evaluation should work together. As opposed to studies reporting that raters should have similar expertise, the present study revealed that raters with different areas of expertise score with higher reliability. The findings of the present study indicate that even though the rater who was an expert solely in the subject domain performed a higher level of reliable scoring than the rater who was an expert solely in measurement and evaluation, it is concluded that together they will produce results with a higher level of reliability. Hence, it is recommended that they do the scorings together. A person who completes measurement and evaluation graduate programs has expertise in this field. Although people who graduated from different undergraduate programs participate in graduate education because there is no undergraduate program, generally, those who graduated from the field of digital education do postgraduate education. The reason for this is the limited number of graduate programs in universities and the high placement scores of the applicants. For this reason, finding an assessment and evaluation specialist in all disciplines is difficult. The results of this research show how important the cooperation between the subject matter expert and the measurement and evaluation expert is, and it is necessary to work together in the test development and scoring process.

After the education which the Turkish teacher candidates received in relation to measurement and evaluation and the test development process, they developed a test consisting of different types of items. Subsequent to the analyses, it was revealed that the teacher candidates had the highest level of competence in true-false items and then followed short response, and open-ended items. The teacher candidates' lowest competence among the different types of items

was observed to be in writing multiple choice items. This finding is consistent with the literature in that writing multiple choice items is difficult. Among the different types of items, the True-False item type can be described as an item type where there is a single statement which needs to be identified as true or false. Open-ended items are more difficult than short response items because they are written to measure higher level skills. Although often still, multiple-choice tests form the backbone of most standardized and classroom tests for various reasons. The advantages of multiple-choice assessments over most free-response assessments include lower costs for scoring, higher reliability, broader sampling of content, and the ability to obtain a wide range of scores (Gierl, Bulut, & Zhang, 2017; Fuhrman, 2018). In this study, pre-service teachers formed multiple-choice items at understanding, application, and analysis levels. Similarly, open-ended items were prepared to measure high-level skills. In other words, it is seen that pre-service teachers have the most difficulty in formulating items to measure high-level skills. This result is consistent with the literature. Asim, Ekuri, & Eni (2013) also determined in their study that pre-service teachers struggled to write multiple-choice items to measure high-level skills. Haladayna, Downing, & Rodriguez (2002) drew attention to the difficulty of writing multiple-choice items for teachers and pre-service teachers in their study where they determined the principles of test development. Özçelik (2010) asserts that multiple choice items can only be written after a certain period of preparation and experience. According to Özçelik (2010a), one must first start by writing short response items and by doing so learn how to write multiple choice items. Preparing a test consisting of multiple choice items would require quite a long period of time because writing the items requires not only expertise in the subject domain but also certain knowledge and skills in measurement and evaluation (Tan, 2012). The findings obtained in the present study are consistent with those reported in the related literature. However, further studies are needed on teacher candidates' practice in writing particularly open-ended and multiple choice test items. Teachers state that they are not competition at the item writing. For this reason, pre-service teachers need to gain theoretical knowledge about measurement and evaluation processes and practice. The findings obtained in the present study are consistent with those reported in the related literature. However, further studies are needed on teacher candidates' practice in writing, particularly open-ended and multiple-choice test items. However, reducing the measurement and evaluation course to 2 hours per week in 2020 makes this situation difficult. For this reason, increasing the course hours or taking a separate course before the service for test development is recommended.

When the raters' expertise and the interaction between different types of items were examined, it was found that raters' expertise were mostly influential on scoring of short response items. In other words, variations among the raters' scores were mostly observed in the short response items. Short response items are those where students provide a number, word or a sentence as a response (Özçelik, 2010b), and since there are no options in the item and the student needs to provide his/her own response, subjectivity can be involved in scoring these items (Tekin, 2004). When the scoring criteria of short response items were examined, it could be observed that short response items had such expertise as having a single correct answer, being understood in the same way by different people, being clear and comprehensible, and matching the measured target learning outcome. While the rater with expertise in solely measurement and evaluation assigned a high score to a single response to an item developed, by for instance student no. 52, the rater with expertise in solely Turkish education assigned a low score. As previously mentioned, these findings indicate the importance of collobarative work in scoring by an expert on the subject domain and an expert on measurement and evaluation during the development of test items.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ağrı İbrahim Çeçen University, 01/12/2021, E-95531838-050.99-25942.

Authorship Contribution Statement

Ayfer Sayin: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Mehmet Sata**: Methodology, Supervision, and Validation. Authors may edit this part based on their case.

Orcid

Ayfer Sayin https://orcid.org/0000-0003-1357-5674
Mehmet Sata https://orcid.org/0000-0003-2683-4997

REFERENCES

Anthony, C.J., Styck, K.M., Volpe, R.J., & Robert, C.R. (2022). Using many-facet rasch measurement and generalizability theory to explore rater effects for direct behavior rating–multi-item scales. *School Psychology. Advance online publication*. https://doi.org/10.1037/spq0000518

Asim, A.E., Ekuri, E.E., & Eni, E.I. (2013). A Diagnostic Study of Pre-Service Teachers' Competency in Multiple-Choice Item Development. *Research in Education, 89*(1), 13–22. https://doi.org/10.7227/RIE.89.1.2

Atılgan, H., & Tezbaşaran, A. (2005). Genellenebilirlik kuramı alternatif karar çalışmaları ile senaryolar ve gerçek durumlar için elde edilen g ve phi katsayılarının tutarlılığının incelenmesi. *Eğitim Araştırmaları, 18(1)*, 28-40.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31–57.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme*. ÖSYM Yayınları.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Puanlayıcılar Arası Güvenirlik Belirleme Tekniklerinin Karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8*(1), 63-78. https://doi.org/10.21031/epod.294847

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2018). *Eğitimde bilimsel araştırma yöntemleri*. Pegem Akademi. https://doi.org/10.14527/9789944919289

Crocker, L.M. & Algina, L. (2008). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135.

Erguvan, I.D. & Aksu Dünya, B. (2021). Gathering evidence on e-rubrics: Perspectives and many facet Rasch analysis of rating behavior. *International Journal of Assessment Tools in Education , 8*(2) , 454-474 . https://doi.org/10.21449/ijate.818151

Erman Aslanoğlu, A., & Şata, M. (2021). Examining the differential rater functioning in the process of assessing writing skills of middle school 7th grade students. *Participatory Educational Research (PER), 8*(4), 239-252. https://doi.org/10.17275/per.21.88.8.4

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal, 34*(1), 79-101. https://doi.org/10.37546/JALTJJ34.1-3

Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal, 15*(11), 76-83. https://doi.org/10.4304/tpls.1.11.1531-1540

Fuhrman, M. (1996) Developing Good Multiple-Choice Tests and Test Items, *Journal of Geoscience Education, 44*(4), 379-384. https://doi.org/10.5408/1089-9995-44.4.379

Gierl, M.J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research, 87*(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30(1)*, 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Gorin, J.S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8), 456-462. https://doi.org/10.3102/0013189X07311607

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment, *Applied Measurement in Education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Jones, E., & Bergin, C. (2019) Evaluating Teacher Effectiveness Using Classroom Observations: A Rasch Analysis of the Rater Effects of Principals, *Educational Assessment, 24*(2), 91-118. https://doi.org/10.1080/10627197.2018.1564272

Kamış, Ö. & Doğan, C.D. (2017). How consistent are decision studies in G theory?. *Gazi University Journal of Gazi Educational Faculty, 37*(2), 591-610.

Kara, Y., & Kelecioğlu, H. (2015). Puanlayıcı Niteliklerinin Kesme Puanlarının Belirlenmesine Etkisinin Genellenebilirlik Kuramı'yla İncelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 58-71. https://doi.org/10.21031/epod.47997

Karasar, N. (2018). *Bilimsel araştırma yöntemi* (33th ed.). Ankara: Nobel Yayıncılık.

Kim, H. (2020). Kim, H. Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis. *Lang Test Asia 10*(16), 1-23, https://doi.org/10.1186/s40468-020-00115-0

Leckie, G., & Baird, J.A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Read Writ*. https://doi.org/10.1007/s11145-022-10279-1

Linacre, J.M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, 7(1), 283-284.

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

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

Linn, R.L., & Grolund, N.E. (2000). *Measurement and assessment in teaching* (8th ed.). Merrill/Prentice Hall.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas, 9*(3), 200-215.

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

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. AQA Education.

Milli Eğitim Bakanlığı (2019). *Türkçe Dersi Öğretim Programı (İlkokul ve Ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. Sınıflar)*. MEB Yayınları.

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

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

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological methods, 5*(3), 343-355.

Özçelik, D.A. (2010a). *Ölçme ve değerlendirme.* Pegem Akademi.
Özçelik, D.A. (2010b). *Test geliştirme kılavuzu*. Pegem Akademi.
Primi, R., Silvia, P.J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 176–186. https://doi.org/10.1037/aca0000230
Sayın, A., & Kahraman, N. (2020). A measurement tool for repeated measurement of assessment of university students' writing skill: development and evaluation. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 113-130. https://doi.org/10.21031/epod.639148
Sayın, A., & Takıl, N.B. (2017). Opinions of the Turkish teacher candidates for change in the reading skills of the students in the 15 year old group. *International Journal of Language Academy, 5*(2), 266-284. http://dx.doi.org/10.18033/ijla.3561
Sireci, S.G. (2007). On validity theory and test validition. *Educational Researcher, 36*(8), 477-481. https://doi.org/10.3102/0013189X07311609
Song, T., Wolfe, E.W., Hahn, L., Less-Petersen, M., Sanders, R., & Vickers, D. (2014). *Relationship between rater background and rater performance*. Pearson.
Tan, Ş. (2012). *Öğretimde ölçme ve değerlendirme KPSS el kitabı*. Ankara: Pegem Akademi.
Tekin, H. (2004). *Eğitimde ölçme ve değerlendirme*. Yargı Yayınevi.
Walsh, W.B., & Betz, N.E. (1995). *Tests and assessment*. Prentice-Hall, Inc.
Wiseman, C.S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150-173. https://doi.org/10.1016/j.asw.2011.12.001

APPENDIX

Appendix 1. Rubric

Item Type	Criteria	Score
True/False	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (not containing only absolutely, etc. expressions, having only one correct answer, not giving clues, not being used one-to-one in the text, etc.)	5
Multiple-choice	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (having only one correct line, the structure of the options, appropriateness of the item root, etc.)	5
Short-answered	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (having only one correct answer, not giving clues, not being one-to-one in the text, limited response, etc.)	5
Open-ended	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (suitability for measuring high-level mental skills, the correctness of the answer key, etc.)	5

Appendix 2. *The distribution of standardized residual values by item type*