International Journal of Assessment Tools in Education
2021, Vol. 8, No. 3, 649–666
https://doi.org/10.21449/ijate.730977
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Examination of Common Exams Held by Measurement and Assessment Centers: Many Facet Rasch Analysis

Gulden Kaya Uyanik[1], **Tugba Demirtas Tolaman**[2,*], **Duygu Gur Erdogan**[2]

[1]Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Turkey.
[2]Sakarya University, Faculty of Education, Department of Turkish and Social Sciences Education, Sakarya, Turkey.
[3]Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Turkey.

ARTICLE HISTORY

Received: May 02, 2020
Revised: May 01, 2021
Accepted: July 08, 2021

Keywords:
Turkish Course,
Common Exam,
Many Facet Rasch Analysis,
Multiple-Choice Item,
Rater.

Abstract: This paper aims to examine and assess the questions included in the "Turkish Common Exam" for sixth graders held in the first semester of 2018 which is one of the common exams carried out by The Measurement and Evaluation Centers, in terms of question structure, quality and taxonomic value. To this end, the test questions were examined by three specialists with expertise in different fields in terms of structure, content, and taxonomic values. The test questions were then rated by raters with expertise in different fields according to the criteria set by the researchers. Hence, the study employed the descriptive survey model. The data obtained from the assessment of the questions were analyzed using the Many Facet Rasch Model (MFRM). According to the findings, of the 20 questions included in the exam, 5 (five) are in the category of "*Remembering*", 12 (twelve) in the category of "*Understanding*", 2 (two) in the category of "*Analyzing*" and 1 (one) in the category of "*Evaluating*." Accordingly, the number of questions that measure higher-order thinking skills was lower than the number of lower-level questions. In addition, the study contained three facets: raters, tasks (items), and criteria. There were no differences among the raters (a Turkish Education Specialist, a Program Development Specialist, and a Testing and Assessment Specialist) in terms of severity and leniency: all the raters were in agreement. Finally, in this study, the questions met the criteria measuring the structural features, while they failed to meet the criteria measuring the quality and clarity.

1. INTRODUCTION

The new century requires raising individuals who are not passive receivers of information,(i.e., who do not only obtain information but also question it, translate information into different forms according to changing conditions, use information effectively, and develop higher-order thinking skills, such as creative thinking, critical thinking, and comparison). Hill (2016) defines higher-order thinking skills as the ability to transcend the information provided, adopt a critical stance, make evaluations, develop meta-cognitive awareness, and use problem-solving skills. For this reason, education systems aim to raise individuals equipped with the skills needed in the 21st century, starting from the preschool period. Raising individuals who can meet the

*CONTACT: Tuğba DEMİRTAŞ TOLAMAN ✉ tdemirtas@sakarya.edu.tr Sakarya University, Faculty of Education, Department of Turkish and Social Sciences Education, Sakarya, Turkey

e-ISSN: 2148-7456 /

expectations of the era, keep up with current developments, have a sense of self-confidence, research, question, and realize themselves is what is expected from modern education systems (Anil, 2009). Language skills are an important tool in the acquisition of high-level mental skills. According to Gunes (2007), language is the most important tool for learning as well as developing mental skills. In Turkey, language skills consisting of reading, writing, speaking, and listening skills are acquired and developed by students in Turkish classes. Turkish lesson is not considered a course aiming to give information, but a process aimed at helping students to acquire and develop language skills (Kurudayioglu & Cetin, 2015; Karaduz, 2010; Gunes, 2011). The four basic language skills targeted by the Turkish course are also a basis for other courses. In other words, students' success in Turkish classes in understanding, interpreting, criticizing, and evaluating what they read and making inferences is a prerequisite for success in other courses as well as for overall academic performance. As a matter of fact, according to Cer (2018), one of the most important responsibilities in developing higher-order thinking skills in children falls upon the shoulders of mother tongue programs.

Constructivism-based curricula that have been implemented in Turkey since 2006 target not only basic language skills but also higher-order thinking skills. For example, Turkish Course Curriculum designed by the Republic of Turkey Ministry of National Education (2006) includes higher-order thinking skills such as critical thinking, creative thinking, problem-solving, and decision making. Also, in the revised Turkish Course Curriculum (2019), the structure and hierarchy of the learning objectives have been arranged in a way that contributes to the development of students' high-level cognitive skills as well as basic language skills. The curriculum aims to help students develop skills such as researching, exploring, interpreting, and constructing knowledge as well as accessing information from printed materials and multimedia sources and organizing, questioning, using, and producing information. In addition, it is aimed to help students understand, evaluate, and question what they read from a critical perspective.

All these skills are planned to be conveyed to students through the learning objectives specified in the curriculum. On the other hand, testing and assessment, which reveals whether the stated learning objectives are achieved by students, is carried out by teachers through classroom activities. In addition, the Ministry of National Education or Student Selection and Placement Center conducts testing and assessment on a national and local scale in order to place students in a higher education institution (Kardes-Birinci, 2014; Cepni, Ozsevgenc & Gokdere, 2003). Furthermore, to evaluate the Turkish education system according to international criteria, Turkey has participated in the Programme for International Student Assessment (PISA), a worldwide study by the Organisation for Economic Co-operation and Development (OECD).

Recent years have witnessed developments in cognitive, psychometric, and technological tools, concepts and theories in the assessment of education (Mislevy, 2006). One of the developments in Turkey is the "Turkish Language Test for Four Skills." This test is held to measure students' four basic language skills within the framework of the Turkey's 2023 Education Vision. The pilot implementation of the test, which was held by the Ministry of National Education to measure students' four language skills in an electronic environment and with a standard measurement tool, was carried out with the participation of 7th-grade students in 15 provinces. This test is important in that it was the first nation-wide practice to measure students' basic language skills in the mother tongue in line with international standards (Republic of Turkey Ministry of National Education, 2020).

Regardless of the level and content of education, measuring student learning throughout and at the end of the education process is a necessity (Buyukozturk, 2016). The main tools for educational measurement are tests and exams. They do not only measure students' knowledge and skills related to a particular area (Buyukozturk, 2016) but also are indicators of whether learning objectives specified in a curriculum are achieved. Downing (2006) underlines twelve

steps for effective test development: overall plan, content definition, test specifications, item development, test design and assembly, test production, test administration, scoring examination responses, establishing passing scores, reporting examination results, item banking, and test technical report. These twelve steps provide a structured, systematic process for developing effective exams/tests of all kinds.

The content definition is one of the most important steps of test development. When defining the content, a table of specifications is used. On the other hand, when developing a table of specifications, taxonomies are used. The taxonomies developed to be used in the educational field (Bloom 1956, Haladayna, 1997; Marzano & Kendall, 2007, etc.) are used not only to guide the development of curricula but also for the development of effective test questions suitable for learning outcomes and objectives. These learning taxonomies also provide standardization in education both at the national and international levels.

Revised Bloom's Taxonomy employed in this study is a revision of Bloom's Taxonomy developed by Bloom et al. in 1956. It was published in 2001 by a group of testing and assessment specialists, cognitive psychologists, curriculum theorists, and instructional researchers chaired by Lorin W. Anderson, who was once a student of Bloom (Anderson et al., 2001). Bloom's original taxonomy identified six levels within the cognitive domain, from the simple recall or recognition of facts to increasingly more complex and abstract mental levels. These six levels are (1) *knowledge*, (2) *comprehension*, (3) *application*, (4) *analysis*, (5) *synthesis* (6) *evaluation* (Anderson, 2005). On the other hand, Revised Bloom's Taxonomy contains two dimensions: knowledge dimension (factual knowledge, conceptual knowledge, procedural knowledge, and meta-cognitive knowledge) and cognitive process dimension (applying, analyzing, evaluating). Also, in the revised taxonomy, knowledge was renamed as "remembering," comprehension as "understanding," and synthesis as "creating" (Anderson, 2005). Thus, the original taxonomy was revised and provided with a structure more appropriate for the new century.

Exams/tests are administered through asking questions. The question at the center of learning is generally defined as a statement expressed to extract information from the learner (Hill, 2016). Asking and answering questions means engaging in a mental process. According to Dillion (2006), "one turns to logic, philosophy, and linguistics for analyses of the nature of questions, their relation to answers, and their function in discourse, that is, for a theory of questions." This indicates that questions are a complex but effective tool consisting of many skills.

Asking and answering questions is one of the activities/methods frequently used in communication. Since Socrates (469 BC - 399 BC), Socratic method and questions (Noddings, 2018) have been at the center of learning and teaching activities. Teachers ask questions for different purposes in their educational activities. According to Yildirim (2012), for example, questions are the most important tools to monitor student learning. The competence of teachers and students at this level has an important place in improving students' comprehension of what they read. Gunes (2012) lists the objectives of questions in Turkish teaching as motivating students, increasing their comprehension levels, helping them develop language and mental skills, and effectively conducting and evaluating the learning and teaching process. Andre (1979) argues that questions may be used in at least four different situations to guide student learning: questions can be used in classroom recitation or discussion; they can be inserted in text or other instructional media; they can be used on examinations; finally, students can ask questions of themselves while studying.

The nature of questions has a crucial impact on the progress/development of thought in the classroom. The questions asked by teachers not only define the framework of the lesson but also indicate teachers' expectations from students (Wilen, 1991). In addition, the level of

questions also affects the quality of thinking skills. According to Andre (1979), level-of-question refers to the nature of cognitive processing required to answer a question. A question may ask a learner to repeat or recognize some information exactly as it was presented in instruction. Such a question is typically referred as a knowledge, factual, or verbatim question. Factual questions are believed to involve less complex cognitive processing than questions requiring more than direct memory. Ates et al. (2016) stated that teachers tend to ask lower-level questions, that students' thinking levels are affected by the questions asked by teachers, and students do not usually ask questions at higher levels than those posed by their teachers. The authors also stated that teachers often use questions to measure and assess students' comprehension levels, rather than to improve their comprehension skills or enable them to develop higher-order thinking skills.

Also, Dillon (2006) argues that questions alone are insufficient to foster students' independent thinking and may limit their thinking abilities. To eliminate these limitations, the following methods are recommended: avoiding direct questions, confirming what is said, keeping silent (waiting). Shaunessy (2000) recommends that for students to develop creative, critical, and higher-order thinking skills, teachers should use divergent questions to provoke more questions and new inquiries rather than convergent questions that have one correct answer.

Questions are considered as one of the basic tools of thinking and are often employed in Turkish classes for different purposes. A thorough review of the relevant literature has yielded a number of studies examining the questions asked by teachers in Turkish tests, the questions included in teacher's books and workbooks, and the questions asked by teachers and students during Turkish classes. Kavruk and Cecen (2013), Cintas Yildiz (2015), Gufta and Zorbaz (2008) examined the test questions prepared by Turkish teachers, Bircan (2012), Yesilyurt (2012), and Aktas (2017) examined the test questions prepared by prospective teachers studying in the Turkish language teaching department, and Gocer (2016) examined the questions prepared by Turkish teachers enrolled in postgraduate education. Also, Cayhan and Akin (2015) examined the nation-wide TEOG (transition from primary to secondary education) test, and Demiral and Mensan (2017) compared the test questions developed by teachers with TEOG test questions and PISA test questions.

Besides, many studies have examined the questions in Turkish workbooks and teacher's books as well as reading comprehension questions included in student's books. Gocer (2008), Cecen and Kurnaz (2015) examined the questions in the measurement and evaluation sections at the end of each theme, Ozdemir et al. (2007) and Bozkurt et al. (2015) examined the questions in workbooks, Eroglu and Kuzu (2014) examined the grammar questions in workbooks, Onalan and Zengin (2015), Sarar-Kuzu (2013), and Celikturk-Sezgin and Gedikoglu-Ozilhan (2019) examined reading comprehension questions, and Durukan (2009) examined the questions in teacher's books. All these studies examined the test questions developed by Turkish teachers, the questions in Turkish textbooks, the questions in workbooks, reading comprehension questions, and the questions in teacher's books according to Bloom's Taxonomy or the Revised Bloom's Taxonomy and revealed that the examined questions do not address high-level mental skills, which is an important common finding.

1.1. Many Facet Rasch Model

The main problem of the study was answered by using the Many Facet Rasch Model (MFRM). The Rasch Model, which is a two-facet model based on Item Response Theory (IRT), is used in measurement situations where the Rasch Model is affected by different variability sources (raters, different measurement situations, etc.) other than individual and item facets. MFRM is a measurement model that can overcome the limitations of Classical Test Theory (CTT) (Anshel et al., 2009; Kim et al., 2012; Govindasamy et al., 2019; Uto, 2020). In the MFRM, predictions for each facet (individual, item, rater, situation, etc.) can be made independent of

other variability sources (Engelhard, 1994). For example; item parameters can be estimated independently of the severity/generosity levels of raters or other sources of variability that may affect the measurement results. In CTT, the ability levels of individuals in a test are estimated by the sum of their scores from test items. It is assumed that the difficulty levels of each item in the test (or the likelihood of participating in an item or not) are equal and / or their contribution to the total score is the same. However, if each item has a different contribution in the measured property, accepting the contribution of each item as equal in the total score causes biased results and the statistics based on this acceptance contain errors (Brinthaupt & Kang, 2012). Based on the results obtained from raw scores in CTT, individuals can only be ranked according to their ability levels and these scores obtained in the ranking scale cannot be collected. However, the mathematical model on which MFRM is based overcomes this limitation and by taking the natural logarithm of the raw data (log-odds), the measurement results are converted to the interval scale (logit) level. In addition to these, compliance statistics (INFIT and OUTFIT) can be determined for each variability source with a single analysis in MFRM. In addition, parameter estimates for each facet can be interpreted together on a common ruler (logit scale) (Linacre, 1989). Relative places of facets can be examined in this ruler with a common metric. Thus, for example; By observing the distribution of items, it can be determined at what level the item was absent / missing and at what level there were many items throughout the skill level (Brinthaupt & Kang, 2012). In addition, MFRM also provides descriptive information about other facets (eg raters) in the study. For example, in a measurement case involving more than one rater, one rater scoring more generous than the others; This "unexpected scoring situation" can be determined where all other raters give a high score and this rater gives a lower score (Linacre, 1989). When examined in the light of all this information, it was thought that MFRM was a suitable method for this study where 3 different raters evaluated based on 20 different criteria.

1.2. Purpose of the Study

In 2017, the “Monitoring, Research and Development Project of Measurement and Evaluation Practices” was launched by the Ministry of Education in our country. In the annual report prepared, the main objectives of this project are stated as follows;

- *Improving the measurement and evaluation capacities of the provinces,*
- *Revealing the acquisition levels of the students and teachers in a way to give feedback,*
- *Ensuring that teachers perform more qualified exams by using the Question Bank software to be created at the end of the project,*
- *Improving the capacity of conducting joint exams across the province.*

Within the scope of the project, measurement and evaluation centers have been established in 81 provinces and common exams have been started in most of these centers and these exams are still ongoing. Besides, these exams expand their content in terms of grade level and lesson each year. In this study, conducted in this regard, it was aimed to examine and assess the questions included in the province-wide “Turkish Common Exam” for sixth graders held in the first semester of 2018 by the Sakarya measurement and evaluation center. To this end, the test questions were examined by three specialists with expertise in different fields in terms of structure, content, and taxonomic values, and the obtained results were evaluated. The study is considered to be important in terms of revealing the structural features of province-wide common examinations and providing suggestions for implementation in line with the opinions of the specialists. On the other hand, it is important in terms of bringing a critical perspective to the exams and revealing the points that should be considered in the exams held in all measurement and evaluation centers throughout the country through the Sakarya Sample.

2. METHOD

This study examined the questions included in the province-wide "Turkish Common Exam" for sixth graders held in the first semester of 2018 by the Sakarya Provincial Directorate of National Education- Testing and Assessment Services Unit. The test questions were then rated by raters with expertise in different fields according to the criteria set by the researchers. Hence, the study employed the descriptive survey model (Karasar, 1998).

2.1. Study Group

The study group consists of a team of three raters (a Turkish Education Specialist, a Program Development Specialist, and a Measurement and Evaluation Specialist). Researchers came together to deal with the questions structurally. Then, they determined the structural criteria that should be included in a question in the context of curriculum development dimensions, assessment and evaluation dimensions and Turkish education program objectives in the light of the relevant literature. As a result of the decision of the researchers and the relevant literature review, 20 criteria have emerged. Each question addressed in the context of these 20 criteria was evaluated separately by the researchers. The researchers independently examined the 20 test questions included in the common exam using the 20-criteria assessment form developed by the researchers.

2.2. Data Collection Tools

In the research, firstly, literature review was conducted on the stages of test development and a list of criteria obtained from the related sources (Downing, 2006; Garden & Orpwood, 1996; Lane et al., 2015; Ozcelik, 2009; Webb, 2007) was developed. The list was examined by the specialists, who added additional criteria suitable for the purpose of the study, and a form for assessing test questions was finally created. The form was then examined by the testing and assessment specialist in terms of its scope and by a grammar specialist in terms of grammar. The form was edited and finalized according to the opinions of the specialists.

As a result of examinations and editions in terms of scope and grammar, a form consisting of 20 items was obtained. The first part of the form consists of 2 items (to find out whether the test questions are positively or negatively worded and to what step in Revised Bloom's Taxonomy the test questions correspond), aiming to describe the descriptive features of the test question. The second part contains 18 3-point Likert type items. Rating in the second part is as follows: 1= no, 2=partially, and 3=yes.The data collection tool used in the study is included in the appendix.

2.3. Data Analysis

The data obtained from the assessment of the questions were analyzed using the Many-Facet Rasch Model (MFRM). MFRM has a conceptual framework similar to regression analysis (Eckes, 2011), and with this analysis method, groups, raters, and items are categorized in a reliable manner (Basturk, 2009). In this model, when estimating an individual's ability or levels of items, other variables that may affect the results are taken into account; thus, more objective results are obtained (Stenner, 1990).

In the present study, three raters examined and rated the 20 multiple choise questions included in the common exam using an assessment form developed by the researchers. With MFRM analysis, the appropriateness of the questions, the consistency of the raters in rating, and the reliability of the examination criteria were tried to be determined. The study contained three facets: raters, tasks (items), and criteria. Mathematical formula for MFRM which is used for the study is:

$$ln[P_{nijk} / P_{nijk-1}] = E_j (B_n - D_i - C_j - F_k) \quad (1)$$

In Equation (1);

P_{nijk}; probability of all items being awarded,

E_j : a slope for the item characteristic curve associated with rater *j*.

B_n : the items trait level,

D_i : difficulty level of the item,

C_j : raters attitude:

F_k : difficulty of observing k'th category (Myford & Wolfe, 2004)

MFRM analyses were performed using the FACETS computer program developed by Linacre (2007).

3. RESULT / FINDINGS

Test questions were first examined in terms of expressing the item root as positive or negative form. Table 1 presents the findings.

Table 1. *Whether the test questions are positively or negatively worded.*

Question No	Positive / Negative	Question No	Positive / Negative	Question No	Positive / Negative	Question No	Positive/ Negative
1	Negative	6	Positive	11	Positive	16	Positive
2	Positive	7	Positive	12	Positive	17	Positive
3	Negative	8	Positive	13	Positive	18	Positive
4	Positive	9	Positive	14	Positive	19	Positive
5	Negative	10	Negative	15	Positive	20	Positive

As can be inferred from Table 1, of the 20 questions included in the common exam, 4 are negatively while 16 are positively worded questions. Among the many suggestions given to question authors to write multiple choice questions, the most common one is to avoid negatively worded questions (Chiavaroli, 2017). In terms of frequency of citation, one review of educational textbooks noted that 31 of the 35 authors specifically advise against negatively-worded multiple choise questions (Haladyna & Downing, 1989a, 1989b) When we look at the studies in the literature, the main reason for avoiding negative questions is the lowering of the validity of the test (Haladyna & Downing, 1989b; Case & Swanson, 2002). Researchers point to an increased risk of emerging associated technical defects, such as heterogeneous options or low cognitive levels that are seen to be encouraged by negatively worded questions (Chiavaroli, 2017). For this reason, it is desirable to write multiple-choice items as positive questions. In this context, negatively worded questions were examined. It was realized that these questions could also have been written as positively worded. For example, the first question (Which of the following statements cannot be inferred from the graph?) seeks to measure students' ability to read graphs. This question, which covers the learning objective of "interprets the information presented in graphs, tables, and charts", could, in fact, be asked as a positively worded question.

The test questions were also examined in terms of cognitive steps. For this, the raters examined the questions and decided to what step in Revised Bloom's Taxonomy the questions correspond. Table 2 presents the findings.

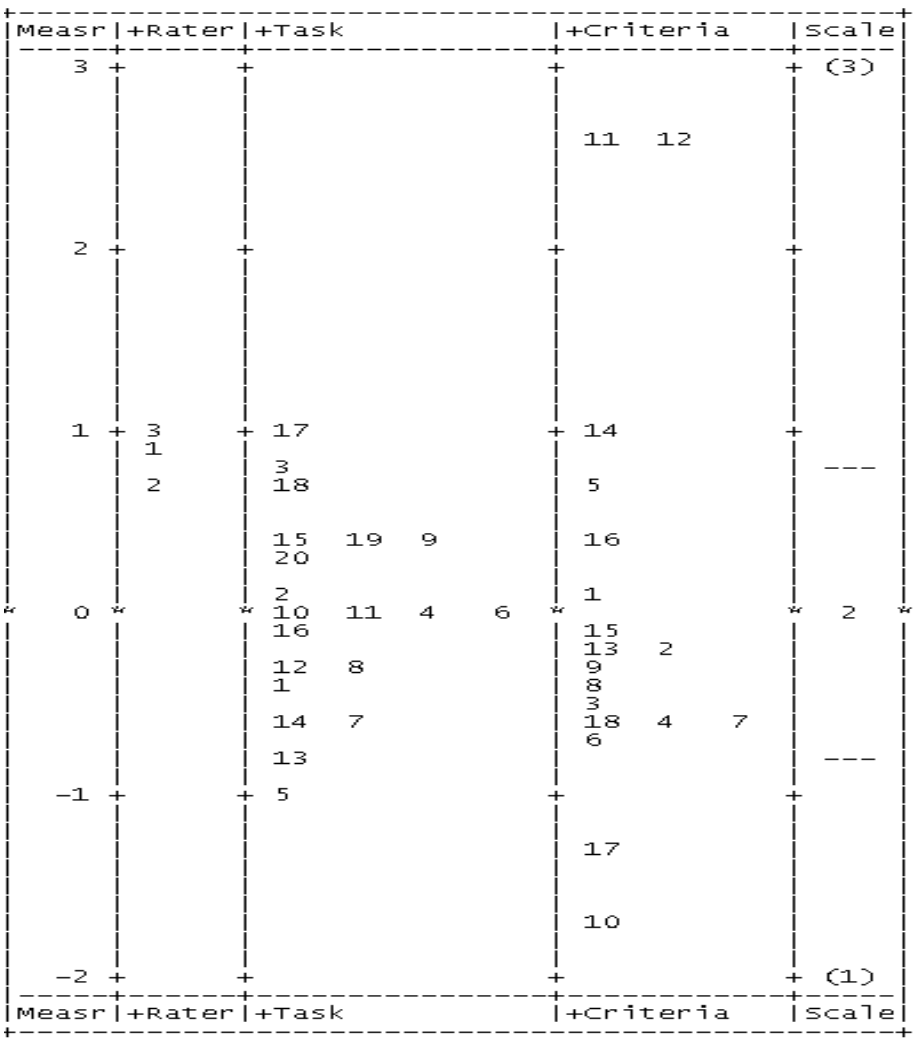
Table 2. *Distribution of the Test Questions by the Revised Bloom's Taxonomy.*

	Remembering	Understanding	Applying	Analyzing	Evaluating	Creating
Items	4-8-18-19-20	1-2-5-6-7-9-11-12-13-15-16-17	-	3-10	14	-

As can be inferred from Table 2, of the 20 questions included in the common exam, 5 correspond to "*Remembering*" step, 12 correspond to "*Understanding*" step, 2 correspond to "*Analyzing*" step, and 1 corresponds to "*Evaluating*" step. Hence, we can conclude that 85% of the questions correspond to "remembering" and "understanding" steps and that there is an insufficient number of questions for higher-order thinking skills.

The ratings of the raters based on the 18 criteria included in the second part of the assessment form were analyzed by MFRM, and the resulting variable map provided by FACETS (citation) software is given in Figure 1.

Figure 1. *Variable Map.*



The logit table in Figure 1 consists of five columns. The first column (Mease)contains the logit, the unit of measurement of the Rasch model. The rater, task (item), and criteria facets are interpreted at this unit level. The second column (Rater) contains the ratings of the raters. When interpreting the rater column, the rater with the highest ratings is considered the severest rater, while the rater with the lowest ratings is considered the most lenient rater. Accordingly, the severest rater was the Turkish Education Specialist (at 1.00 logit) and the most lenient raterwas the Program Development Specialist (at .70 logit). The third column lists the items according to the extent to which they meet the specified criteria. Accordingly, item 17 was the item that most met the criteria (at 1.05 logit), and item 5 was the item that least met the criteria (at -0.97 logit). The fourth column lists the criteria according to the extent to which they are met. Accordingly, the 11th criterion (The question does not contain clues for other questions) and

the 12th criterion (The answer to the question is not given in other questions) were the criteria most met by the test questions (at 2.61 logit), whereas the 10th criterion (The question is for higher-order thinking skills) is the criterion least met by the test questions (at -1.72 logit).

With the logit table, the three facets in the study are interpreted over a linear metric. In addition, detailed measurement reports were obtained for each facet by MFRM analysis. Table 3 presents the measurement results for the rater facet.

Table 3. *Measurement Results for the Rater Facet.*

Raters	Measure	Standard Error	Infit	Outfit
R3	1.00	0.08	1.08	1.09
R2	.94	0.08	.95	.87
R1	.70	0.08	1.00	1.04
Mean	.88	0.08	1.01	1.00
Standard Deviation	0.16	0.00	.07	.12
Reliability= .75	Separation index = 1.27	Chi-square = 8.0	*sd* = 2	*p* = .02

According to the measurement results given in Table 3, Rater 3 was the severest rater while Rater 1 was the most lenient rater, but there is no big difference between their logit values. The fit statistics show the degree of fit between the model and the data, and a value of 1.00 is considered a perfect fit (Hetherman, 2004). According to Wright and Linacre (1994), the acceptable range of infit and outfit values is between "0.5 and 1.5". The fact that the values are in this range indicates that the raters' ratings fit in with the model, in other words, none of the raters disrupted the model-data fit. Since the infit values are between 1.08 and .95 and outfit values between 1.09 and .87, it can be said that the model-data fit was achieved and that no rater disrupted the fit. Also, the raters' reliability index was calculated as .75. The reliability value refers to the difference in the raters' ratings and, like the Cronbach alpha reliability value, it ranges between 0 and 1 and is interpreted in a similar way. In addition, the separation index, which refers to the level of difference between the raters' ratings, is 1.27. Low separation indices indicate that the raters' ratings are consistent and that there is no big difference between the ratings. Thus, considering these two values, we can say that the severity and leniency of the raters were similar. The chi-square (chi-square = 8.0 p <.05) of this difference shows that the difference between the raters is statistically significant. When the logit table in Figure 1 is examined, it is seen that the least spread is among the raters.

Table 4 presents the measurement results the task facet. As can be inferred from Table 4, item 17 was the item that most met the criteria, in other words, it was found to be the most satisfactory item by the raters, whereas item 5 was the item that least met the criteria. The infit values of the items ranged between .63 and 1.34, and the outfit values ranged between .52 and 1.34. This indicates that the infit and outfit values of all the items are in an acceptable range. The average statistical value of the infit and outfit values was found to be 1.00. Accordingly, the average of the fit statistics in the item measurement being equal to 1 indicates that the model-data fit is perfect. The reliability and separation indices of the items were found to be .85 and 2.30, respectively. These values show that the items could be adequately separated in terms of the criteria.

Table 4. *Measurement Results for the Task Facet.*

Items	Measure	Standard Error	Infit	Outfit
I17	1.05	.28	1.10	1.07
I3	.77	.25	.97	.83
I18	.71	.24	.85	1.30
I19	.45	.22	.79	.65
I9	.40	.22	.88	.67
I15	.35	.22	.84	1.10
I20	.30	.21	1.29	1.04
I2	.09	.20	.92	.76
I4	.01	.20	.94	.78
I10	.01	.20	1.25	1.06
I6	-.03	.20	1.34	1.22
I11	-.03	.20	.96	.82
I16	-.11	.19	.63	.52
I12	-.26	.19	1.02	.93
I8	-.29	.19	1.03	.91
I1	-.36	.19	.94	.81
I7	-.65	.19	1.11	1.14
I14	-.65	.19	.95	.80
I13	-.79	.19	1.05	.93
I5	-.97	.19	1.20	.66
Mean	0.00	.21	1.00	1.00
Stn. Dev.	.54	.02	018	.44
Reliability= .85	Separation index = 2.30	Chi-square = 115.9	*sd* = 19	*p* = .00

Table 5 presents the measurement results for the criterion facet. As can be inferred from Table 5, the 11th criterion (The question does not contain clues for other questions) and the 12th criterion (The answer to the question is not given in other questions) were the criteria most met by the test questions, whereas the 10th criterion (The question is for higher-order thinking skills) was the criterion least met by the test questions. The infit values of the criteria ranged between .57 and 1.42, and the outfit values ranged between .60 and 1.35. Considering that the optimal range for fit statistics is between 0.5 and 1.5, all criteria contributed to a perfect model-data fit. The reliability and separation indices of the criteria were found to be .93 and 3.64, respectively. Accordingly, the criteria functioned reliably to separate the items according to the extent to which they met the criteria. In addition, the significant chi-square value (chi-square = 189.9, p <0.05) shows that there is a statistically significant difference between the difficulty levels of the criteria.

Table 5. *Measurement Results for the Criterion Facet.*

Criteria	Measure	Standard Error	Infit	Outfit
C11	2.61	.69	1.42	.87
C12	2.61	.69	1.42	.87
C14	1.02	.29	1.25	1.35
C5	.66	.25	.94	1.07
C16	.35	.21	.91	.81
C1	.14	.20	.93	1.03
C15	-.15	.18	1.25	1.19
C2	-.21	.18	1.03	1.11
C13	-.21	.18	1.18	1.07
C9	-.34	.17	.86	.84
C8	-.43	.17	1.24	1.19
C3	-.52	.17	.98	1.03
C4	-.57	.17	1.24	1.34
C7	-.57	.17	.75	.73
C18	-.63	.17	1.14	1.10
C6	-.74	.17	.96	.92
C17	-1.30	.17	.84	.88
C10	-1.72	.19	.57	.60
Mean	0.00	.25	1.09	1.00
Stn. Dev.	1.11	.16	.32	.20
Reliability= .93	Separation Index = 3.64	Chi-square = 189,9	*sd* = 17	*p* = .00

Table 6. *Measurement Results for the Rating Scale Facet.*

Criterion Ratings	Frequency	%	Cumulative %	Average Measurement	Expected Measurement	Outfit
1	226	21	21	-.04	-.11	1.3
2	236	22	43	.30	.44	.5
3	618	57	100	1.44	1.41	.9

Table 6 presents the measurement results for the scale facet (1=no, 2=partially, 3=yes). As can be inferred from Table 6, of all the ratings, 21.2% are 1 (no), 22% are 2 (partially), and 57% are 3 (yes). Accordingly, as the ratings increased (from 1 to 3), their usage rates also increased. The frequency of the ratings at a value of at least 10 indicates that the ratings functioned adequately and have a balanced distribution (Engelhard, 1994). Accordingly, considering the obtained frequency, we can say that the rating data are at the desired level. The outfit values of the criterion rating range between .5 and 1.3, which indicates that the rating fits the model.

4. DISCUSSION and CONCLUSION

In this study, 20 multiple-choice questions in the 6th grade Turkish course common exam conducted throughout the province by the Directorate of National Education Directorate of Assessment and Examination Services were examined by three different field experts through a form consisting of 20 criteria.

Our findings show that of the 20 questions included in the exam, 4 are negatively while 16 are positively worded questions. Using negative questions in a test affects the test reliability;

therefore, it is necessary to avoid negative questions. Negative expressions such as "not," "except" decrease the comprehensibility of the question, increasing the probability of the student making mistakes due to lack of attention. In addition, it takes more time for the student to answer such items (McMillan, 2013). Therefore, in common exams seeking to measure students' language skills, to achieve accurate measurement results and to exclude other variables like "attention" from the test results, it is recommended to avoid negative questions.

It was also investigated that, the questions examined in the study correspond to which step in the "Revised Bloom Taxonomy". Of the 20 questions, 5 corresponded to "*Remembering*" step, 12 corresponded to "*Understandin*g" step, 2 corresponded to "*Analyzing*" step, and 1 corresponded to "*Evaluating*" step. Accordingly, the number of questions that measure higher-order thinking skills was lower than the number of lower-level questions. Studies in the relevant literature have also reported similar findings. In the study conducted by Kavruk and Cecen (2013), the questions in the 6th, 7th, and 8th-grade tests developed by 38 Turkish teachers were examined according to Bloom's Taxonomy. As a result of the assessment, it was observed that most of the questions were at the level of knowledge, comprehension, and application that measure lower-level skills. In a similar study conducted by Cintas Yildiz (2015), the questions in the 5th, 6th, and 7th-grade tests were analyzed according to the Revised Bloom's Taxonomy, and most of the questions were found to be at conceptual knowledge step of the knowledge dimension and at the understanding step of the cognitive process dimension. In addition, studies conducted with taxonomies other than Bloom's Taxonomy reported similar results. Kocaarslan and Yamac (2018) examined reading comprehension questions in tests developed by Turkish teachers according to the reading comprehension taxonomy developed by Day and Park (2005) and stated that the questions were mainly for literal comprehension. The authors also found that only a few questions triggered learners' reorganization, prediction, and personal response skills while none aimed to assess inference and evaluation skills. Similarly, Ates (2011) found that teachers most frequently employ the strategy of asking questions and that they do not ask many questions to trigger students' higher-order thinking processes. This shows that teachers' questioning skills remain unchanged and continue in a traditional way, even as time progresses.

Lower-level questions that require memorization and conveying existing knowledge instead of generating new knowledge may be beneficial for the learning of disadvantaged children but does not contribute much to the development of normal and gifted children. In contrast, higher-level questions that require students to use higher-order thinking skills contribute to normal and gifted students in terms of cognitive development (Gall, 1984 as cited in Topcu, 2017). According to Akyol et al. (2013), the success (or failure) of Turkish students at questions requiring higher-order thinking skills (such as critical thinking) in international tests such as PIRLS, PISA, and TIMSS may be related to the level of questions they encounter in the teaching process and written materials. Higher-order thinking requires students to go beyond simple recall of facts and manipulate information and ideas. When teachers ask higher-level questions, they may initially see that students have difficulty in answering the questions or that they give answers consisting of only a few words. Therefore, the teacher should model for his/her students how to give a higher-level answer. Though it may take some time to train students to give higher-level answers, it will definitely produce positive outcomes (Peterson & Taylor, 2012). In fact, the Turkish Course Curriculum (Republic of Turkey Ministry of National Education, 2019) underlines the importance of developing tests and exams that contain various types of items that trigger students' higher-order thinking processes such as making inferences, critical thinking, analysis, visual reading, reasoning, and spatial skills.

The 18 criteria in the assessment form were analyzed by MFRM. The study contained three facets: raters, items, and criteria. There were no differences among the raters (a Turkish Education Specialist, a Program Development Specialist, and a Testing and Assessment

Specialist) in terms of severity and leniency: all the raters were in agreement. Commissions to develop common exams to be conducted through the central examination system should include specialists in different fields: a specialist in the lesson content, a program development specialist, and a testing and assessment specialist. The common exam assessed in this study was developed by a commission of Turkish teachers working in the Sakarya Provincial Directorate of National Education- Testing and Assessment Services Unit. The commission does not include a program development specialist or a testing and assessment specialist. However, the commissions to develop such province-wide common exams that will affect many students should include specialists with expertise in different fields. In addition, in-service training on testing and assessment approaches and tools and developing new types of questions should be given to the teachers in such commissions. As a matter of fact, Maden (2011) stated that Turkish teachers found complex the testing and assessment tools and methods in the 2006 Turkish Course Curriculum, and Erdogan (2017) stated that teachers do not make enough effort to improve their questioning skills. As a result, rather than creating their own questions or tasks to use in tests, teachers use readily available questions included in printed or online resources. In fact, we realized that some of the test questions included in the common exam were taken from other resources.

Considering all the criteria in the assessment form used in the study, of the 20 items, only 8 are in the range of 0 and 1 logit, 2 are at 0 logit, and 10 at a negative logit. This indicates that the questions failed to meet the criteria sufficiently. Also, it was observed that the criteria measuring the structural features of the questions were met while the criteria measuring the quality and comprehensibility of the questions were not met. This shows that though the exam development commission paid attention to the structural features of the test questions, they failed to attach sufficient importance to the quality of test questions. In other words, they took care to include multiple-choice items that seek to measure the learning objectives specified in the curriculum but failed to meet the criteria set for the quality of test questions.

Furthermore, the exam failed to measure the four basic language skills in the mother tongue: though the exam contained questions measuring students' grammatical knowledge and reading comprehension skills, there were no questions to assess students' listening, speaking, or writing skills. Turkish classes are aimed at helping students develop all four basic language skills. For this reason, and in order to develop tests measuring students' four basic language skills, the "Turkish Language Test for Four Skills" developed by Republic of Turkey Ministry of National Education (2020) should be examined thoroughly by teachers.

Overall, the study concludes that the questions included in the common exam was appropriate for the learning objectives specified in the Turkish Course Curriculum (Republic of Turkey Ministry of National Education, 2019) but failed to address higher-order thinking skills. Therefore, we recommend that exam development commissions to develop province-wide common exams that will affect many students should include specialists with expertise in different fields.

In this study, an exam for the Turkish course prepared and administered by the Sakarya Provincial Directorate of National Education- Measurement and Evaluation Center- is examined. In the future researches, it is recommended to examine the exams held in different provinces for both Turkish and different courses, compare the results obtained and thus determine the situation across the country.

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.

Authorship Contribution Statement

Gulden Kaya Uyanik: Investigation, Resources, Supervision, Methodology, Development of Data Collection Tool, Analysis, Writing the original draft. **Tugba Demirtas Tolaman**: Investigation, Resources, Development of Data Collection Tool, Writing the original draft. **Duygu Gur Erdogan**: Investigation, Resources, Development of Data Collection Tool, Writing the original draft.

ORCID

Gulden Kaya Uyanik https://orcid.org/ 0000-0002-8100-6994
Tugba Demirtas Tolaman https://orcid.org/ 0002-6632-9752
Duygu Gur Erdogan https://orcid.org/ 0000-0002-2802-0201

5. REFERENCES

Aktaş, E. (2017). Öğretmen adaylarının farklı metin türlerine yönelik soru sorma becerilerinin Yenilenmiş Bloom Taksonomisine göre değerlendirilmesi [Evaluation of the questioning skills of teachers candidates towards the different text types accordıng to The Renewed Bloom Taxonomy]. *Electronic Turkish Studies*, *12*(25), 99-118. http://dx.doi.org/10.7827/TurkishStudies.12274

Akyol, H., Yıldırım, K., Seyit, A., & Çetinkaya, Ç. (2013). Anlamaya yönelik nasıl sorular soruyoruz? [What kinds of questions do we ask for making meaning?]. *Mersin University Journal of Education*, *9*(1), 41-56.

Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education, *Studies in Education Evaluation*, *31*, 102-113. https://doi.org/10.1016/j.stueduc.2005.05.004

Anderson, L.W., Krathwohl, D.R., Airaisan, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). A taxonomy for learning, teaching and assessing: a revision of Bloom's Taxonomy of educational objectives. Addison Wesley Longman, Inc.

Andre, T. (1979). Does Answering higher-level questions while reading facilitate productive learning? *Review of Educational Research*, *49*(2), 280-318.

Anıl, D. (2009). Uluslararası öğrenci başarılarını değerlendirme programı (PISA)'nda Türkiye'deki öğrencilerin Fen Bilimleri başarılarını etkileyen faktörler [Factors effecting Science achievement of science students in Programme for International Students' Achievement (PISA) in Turkey]. *Education and Science*, *34*(152), 87-100.

Anshel, M.H., Weatherby, N.L., Kang M. & Watson, T. (2009). Rasch calibration of a unidimensional perfectionism ınventory for sport. *Psychology of Sport and Exercise*, *10*(2009), 210-216. https://doi.org/10.1016/j.psychsport.2008.07.006

Ateş, S. (2011). *Evaluatıon of Fifth-Grade Turkish Course Learnıng and Teaching Process in Terms of Comprehensıon* Instruction [Unpublished doctoral dissertation, Gazi University]. Gazi University Libraries.

Ateş, S., Güray, E., Döğmeci, Y., & Gürsoy, F. F. (2016). Öğretmen ve öğrenci sorularının gerektirdikleri zihinsel süreçler açısından karşılaştırılması [Comparison of questions of teachers and students in terms of level]. *Research in Reading and Writing Instruction*, *4*(1), 1-13.

Baştürk, R. (2009). Applying The Many – Facet Rasch Model to evaluate powerpoint presentation performance ın higher education. *Assesment And Evaluation in Higher Education*, *33*(4), 431-444. https://doi.org/10.1080/02602930701562775

Bircan, E. (2012). Türkçe öğretmeni adaylarının hazırladığı soruların yeniden yapılandırılan Bloom Taksonomisine göre değerlendirilmesi [Evaluation of the questions prepared by Turkish language teacher candidates according to The Revised Bloom's Taxonomy]. *Kastamonu University Journal of Education*, *20*(3), 965-982.

Bloom, B. (1956). Taxonomy of educational objectives. David Mckay. https://www.uky.edu/~rsand1/china2018/texts/Bloom%20et%20al%20-Taxonomy%20of%20Educational%20Objectives.pdf

Bozkurt, B.Ü., Uzun, G.L., & Lee, Y. (2015). Korece ve Türkçe ders kitaplarındaki metin sonu sorularının karşılaştırılması: PISA 2009 sonuçlarına dönük bir tartışma [A comparıson of reading comprehension questions in Korean and Turkish textbooks: A discussion on PISA 2009 results]. *International Journal of Language Academy*, *3*(4), 295-313. http://dx.doi.org/10.18033/ijla.327

Brinthaupt, T.M., & Kang, M. (2012). Many-faceted rasch calibraton: an example using the self-talk scale. *Assessment*, *21*(2), 241-249. https://doi.org/10.1177/1073191112446653

Büyüköztürk, Ş. (2016). Sınavlar üzerine düşünceler [Thoughts on exams]. *Kalem International Journal of Education and Human Sciences*, *6*(2), 345-356.

Case, S.M. & Swanson, D.B. (2002). *Constructing written test questions for the basic and clinical sciences.* 3rd Ed (rev.) National Board of Medical Examiners.

Cayhan, C., & Akın, E.(2015). TEOG sınavı Türkçe dersi sorularının Türkçe Dersi Öğretim Programındaki kazanımlar açısından değerlendirilmesi [The evaluation of Turkish lesson questions TEOG examinatıon in terms of Turkish lesson education program objectives]. *Siirt University Journal of Social Sciences Institute*, 5, 106-114.

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research, and Evaluation*, *22*(1), 3. https://doi.org/10.7275/ca7y-mm27

Çeçen, M. A., & Kurnaz, H. (2015). Ortaokul Türkçe dersi öğrenci çalışma kitaplarındaki tema değerlendirme soruları üzerine bir araştırma [Student workbook of secondary school Turkish course: A research on theme evaluation questions]. *Journal of Karadeniz Social Sciences*, *7*(2).

Çeliktürk Sezgin, Z., & Gedikoğlu Özilhan, Y. G. (2019). 1.-8. sınıf Türkçe ders kitaplarındaki metne dayalı anlama sorularının incelenmesi [Examining text-based comprehension questions in Turkish textbooks of the 1st- the 8st graders]. *Journal of Mother Tongue Education*, *7*(2), 353-367. https://doi.org/10.16916/aded.530191

Çepni, S., Özsevgenç, T. & Gökdere, M. (2003). Bilişsel gelişim ve formal operasyon dönem özelliklerine göre ÖSS fizik ve lise fizik sorularının incelenmesi [Examination of SSE physics and high school physics questions according to cognitive development and formal operation period features]. *Journal of National Education*, *157*, 30-39.

Çer, E. (2018). A comparison of mother-tongue curricula of successful countries in PISA and Turkey by higher-order thinking processes. *Eurasian Journal of Educational Research*, *73*, 95-112.

Day, R. R., & Park, J. (2005). Developing reading comprehension questions. *Reading in a Foreign Language*, *17*(1), 60-73.

Demiral, H., & Menşan, N. (2017). Sekizinci sınıf Türkçe dersinin PISA okuma becerilerine göre değerlendirilmesi [Evaluation of the eighth grade Turkish lesson according to PISA reading skills]. Küreselleşen dünyada eğitim (Edt: Özcan Demirel, Serkan Dinçer). Pegem Yayıncılık.

Dillon, J.T. (2006). Effect of questions in education and other enterprises. In *Westbury, I.*& *Milburn, G.* (Eds.), *Rethinking Schooling* (pp.145-174). Routledge. https://doi.org/10.4324/9780203963180

Downing, S. M. (2006). Twelve steps for effective test development. In Downing, S.M,& Haladyna, T. M. (Eds.), *Handbook of test development*, (pp.3-25). Routledge.

Durukan, E. (2009). 7. sınıf Türkçe ders kitaplarındaki metinleri anlamaya yönelik sorular üzerine taksonomik bir inceleme [A taxonomic study on questions to understand texts in 7th grade Turkish Textbooks]. *Journal of National Education, 181*, 84-93.

Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt Am.

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. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Erdoğan, T. (2017). İlkokul dördüncü sınıf öğrencilerinin ve öğretmenlerinin Türkçe dersine ilişkin sordukları soruların Yenilenmiş Bloom Taksonomisi açısından görünümü [The view of primary school fourth grade students and teachers' questions about Turkish language lessons in the terms of The Revised Bloom Taxonomy]. *Education and Science*, *42*(192). http://dx.doi.org/10.15390/EB.2017.7407

Eroğlu, D., & Kuzu, T.S. (2014). Türkçe ders kitaplarındaki dilbilgisi kazanımlarının ve sorularının Yenilenmiş Bloom Taksonomisine göre değerlendirilmesi [The evaluation of the grammar a cquisitions and questions in Turkish course books with respect to New Bloom Taxonomy]. *Başkent University Journal of Education*, *1*(1), 72-80.

Garden, R. A., & Orpwood, G. (1996). Development of The TIMSS Achievement Tests. *Third International Mathematics and Science Study. Technical Report, 1*.

Govindasamy, P., del Carmen Salazar, M., Lerner, J., & Green, K. E. (2019). Assessing the reliability of the framework for equitable and effective teaching with the many-facet rasch model. *Frontiers in Psychology*, *10*, 1363. https://doi.org/10.3389/fpsyg.2019.01363

Göçer, A. (2008). İlköğretim Türkçe ders kitaplarinin ölçme ve değerlendirme açisindan incelenmesi [Analysis of Turkish course books for measurement and evaluation]. *Atatürk University Journal of Social Sciences Institute*, *11*(1), 197-210.

Göçer, A. (2016). Lisansüstü eğitim gören Türkçe öğretmenlerinin yazılı sınav sorularının incelenmesi [Investigation of written exam questions of Turkish teachers who upper graduate education]. *Uşak University Journal of Social Sciences Institute*, *9*(27/3), 22-37.

Güfta, H., & Zorbaz, K. Z. (2008). İlköğretim ikinci kademe türkçe dersi yazılı sınav sorularının düzeyleri üzerine bir değerlendirme [A review regarding levels of written examination questions for Turkish courses of the secondary school]. *Çukurova University Journal of Social Sciences Institute*, *17*(2), 205-218.

Güneş, F. (2007). *Türkçe öğretimi ve zihinsel yapılandırma* [Turkish teaching and mental structuring]. Nobel Yayın Dağıtım.

Güneş, F. (2011). Dil Öğretim yaklaşımları ve Türkçe öğretimindeki uygulamalar [Language teaching approaches and their applications in teaching Turkısh]. *Mustafa Kemal University Journal of Social Sciences Institute*, *8*(15), 123-148.

Güneş, F. (2012). Testlerden etkinliklere türkçe öğretimi [Teaching Turkish from tests to activities]. *Journal of Language and Literature*, *1*(1), 31-42.

Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 2(*1), 37-50. https://doi.org/10.1207/s15324818ame0201_3

Haladyna, T.M., & Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, *2*(1), 51-78. https://doi.org/10.1207/s15324818ame0201_4

Haladayna, T.M. (1997). *Writing test ıtems to evaluate higher order thinking.* Viacom Company.

Hetherman, S.C. (2004). *An Application of Multi Faceted Rasch Measurement to Monitör Effectiveness of the Written Composition in English in The New York City Department of Education* [Doctoral dissertation, Colombia University]. Colombia University Libraries.

Hill, J.B. (2016). Questioning techniques: A study of instructional practice. *Peabody Journal of Education, 91*(5), 660-671. https://doi.org/10.1080/0161956X.2016.1227190

Karadüz, A. (2010). Dil becerileri ve eleştirel düşünme [Language skills and the critical thinking]. *Turkish Studies, 5*(3), 1566-1593. http://dx.doi.org/10.7827/TurkishStudies.1572

Karasar, N. (1998). *Araştırmalarda rapor hazırlama yöntemi* [*Research Report Preparation Method*]. Ankara: Pars Matbaacılık Sanayi.

Kardeş-Birinci, D. (2014). Merkezi sistem ortak sınavlarında ilk deneyim: Matematik dersi [The first experıence in central system common exams: mathematics]. *Journal of Research in Education and Teaching, 3*(2), 8-16.

Kavruk, H., & Çeçen, M.A. (2013). Türkçe dersi yazılı sınav sorularının bilişsel alan basamakları açısından değerlendirilmesi [Evaluation of Turkish language class exam questions in point of cognitive field levels]. *Journal of Mother Tongue Education, 1*(4), 1-9. https://doi.org/10.16916/aded.15990

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly, 29*, 346-365.

Kocaarslan, M., & Yamaç, A. (2018). Sınıf öğretmenlerinin Türkçe dersi sınavlarında sordukları metne dayalı anlama sorularının incelenmesi [Investigating text-based comprehension questions primary school teachers ask in exams of Turkish course]. *Trakya Journal of Education, 8*(2), 431-448. https://doi.org/10.24315/trkefd.356769

Kurudayıoğlu, M., & Çetin, Ö. (2015). Temel beceriler ve Türkçe öğretimi [Basic skills and Turkish education]. *Journal of Mother Tongue Education, 3*(3), 1-19. https://doi.org/10.16916/aded.65619

Lane, S., Raymond, M.R., & Haladyna, T.M. (2015). *Handbook of test development*. Routledge.

Linacre, J.M. (1989). *Many-facet Rasch Measurement* [Doctoral dissertation, University of Chicago]. Chicago University Libraries.

Linacre, J. M. (2007). *A User's Guide to FACETS: Rasch Model Computer Programs.* Chicago, IL.

Maden, S. (2011). Türkçe dersi öğretmenlerinin ölçme değerlendirmeye ilişkin algıları [Turkish course teachers' perceptions on measurement and evaluation]. *Journal of National Education, 41*(190), 212-233.

Marzano, R.J., & Kendall, J.S. (2007). *The new taxonomy of educational objectives.* Sage.

McMillan, J. H. (2013). *Classroom assessment: principles and practice for effective standards-based instruction* (6th Edition). Pearson.

Mislevy, R.J., & Riconscente, M.M. (2006). Handbook of test development. In Downing, S.M,&Haladyna,T.M. (Eds.), *Evidence-centered assessment design* (pp.61-90). Routledge.

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.

Noddings, N. (2018). *Philosophy of education*. Routledge.

Önalan, K., & Nesrin, Z. (2015). Türkçe ders kitaplarındaki metin altı soruların aşamalı sınıflandırmaya göre incelenmesi [Examining text-based comprehension questions in turkish textbooks of the 1 st - the 8 st Graders]. *International Journal of Languages' Education and Teaching*, 1527-1533.

Özçelik, D.A. (2009). *Test hazırlama klavuzu* [Test Guide]. (4. Baskı). Pegem Akademi.

Özdemir, M., Özdemir, O., & Çetinkaya, Ç. (2007, 15-17 November). *Analysis of the questions in the primary Turkish course workbooks* [Conference presentation]. 1. Ulusal İlköğretim Kongresi, Ankara.

Peterson, D.S., & Taylor, B.M. (2012). Using higher order questioning to accelerate students' growth in reading. *The Reading Teacher*, *65*(5), 295-304. https://doi.org/10.1002/TRTR.01045

Republic of Turkey Ministry of National Education (2006). Turkish Course Curriculum.

Republic of Turkey Ministry of National Education (2019). Turkish Course Curriculum. http://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=663

Republic of Turkey Ministry of National Education (2020). Turkish Language Exam in Four Skills. https://www.meb.gov.tr/meb_iys_dosyalar/2020_01/20094146_Dort_Beceride_Turkce_Dil_Sinavi_Ocak_2020.pdf

Sarar Kuzu, T. (2013). Türkçe ders kitaplarındaki metin altı sorularının Yenilenmiş Bloom Taksonomisindeki hatırlama ve anlama bilişsel düzeyleri açısından incelenmesi. [Investigation of the text following questions in Turkish course books with respect to their remembering and understanding cognition levels of The Revised Bloom Taxonomy]. *Sivas Cumhuriyet University Faculty of Letters Journal of Social Sciences*, *37*(1), 58-76.

Shaunessy, E. (2000). Questioning techniques in the gifted classroom. *Gifted Child Today*, *23*(5), 14-21. https://doi.org/10.4219/gct-2000-752

Stenner, A. J. (1990). Objectivity: Specific and General. *Rasch Measurement Transactions*, *3*(4), 111.

Topçu, E. (2017). TEOG Tarih sorularının Yenilenmiş Bloom Taksonomisine göre analizi [Analysis of History questions asked in the transition from primary to secondary education according to The Renewed Bloom Taxonomy]. *International Journal of Turkish Education Sciences*, *9*, 321-335.

Turkish Course Common Exam. (2018). The Measurement and Evaluation Center. http://sakarya.odm.meb.gov.tr

Uto, M. (2020). Accuracy of performance-test linking based on a many-facet Rasch model. *Behavior Research Methods*, 1-15.https://doi.org/10.3758/s13428-020-01498-x

Webb, N. L. (2007). Issue related to judging the alignmnet of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7-25.

Wilen, W.W. (1991). *Questioning skills for teachers*. National Education Association.

Wright, B. D., & Linacre, J. M. (1994). Reasonable Mean-Square Fit Values. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 8(3),* 370.

Yeşilyurt, E. (2012). Öğretmen adaylarının bilişsel alanla ilgili sınama durumu soruları yazma yeterliklerinin değerlendirilmesi [Evaluating teacher candidates' competencıes on writıng testing situation questions related to cognitive domain]. *Kastamonu University Journal of Education*, *20*(2), 519-530.

Yıldırım, K. (2012). Öğretmenlerin öğrencilerin okuduğunu anlama becerilerini değerlendirmede kullanabilecekleri bir sistem: Barrett Taksonomisi [A system to be used by teachers to evaluate students' reading comprehensıon skills: Barrett Taxonomy]. *Mustafa Kemal University Journal of Social Sciences Institute*, *9*(18), 45-58.

Yıldız, D. Ç. (2015). Türkçe dersi sınav sorularının yeniden yapılandırılan Bloom Taksonomisine göre analizi [The analysis of Turkish course exam questions according to re-constructed Bloom's Taxonomy]. *Gaziantep University Journal of Social Sciences*, *14*(2), 479-497.