Yılmaz Koğar, E. & Koğar, H. (2019). Investıgatıon of scıentıfic lıteracy accordıng to dıfferent ıtem types: Pısa 2015 Turkey sample. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 19* (2), 695-709.

Makalenin Türü / Article Type : Araştırma Makalesi / Researh Article
Geliş Tarihi / Date Received : 04.10.2018
Kabul Tarihi / Date Accepted : 28.05.2019
Yayın Tarihi / Date Published : 01.07.2019

https://dx.doi.org/10.17240/aibuefd.20xx.xx.xxxxx-xxxxxx

INVESTIGATION OF SCIENTIFIC LITERACY ACCORDING TO DIFFERENT ITEM TYPES: PISA 2015 TURKEY SAMPLE

Esin YILMAZ KOĞAR[1], Hakan KOĞAR[2]

ABSTRACT

The aim of the present study was to reveal the psychometric properties of the items in the cognitive test of PISA 2015 assessing scientific literacy according to different item types and to examine scientific literacy in relation to different independent variables. In the sample of PISA 2015 Turkey, 175 students from 5895 students were included in the study with the aim of researching. Descriptive statistics and various hypothesis tests were used to obtain the findings of the research. When scientific literacy and item difficulty averages for all three item types were examined, students with a high level of scientific literacy level were more successful at responding to constructed response (CR) items, while students with a low level of scientific literacy were more successful at answering multiple choice (MC) items. Male students were more successful than female students in responding to MC and complex multiple choice (CMC) items, while female students were more successful than male students in answering CR items. It was found that students with a high level of economic, social and cultural status were more successful than those with a low level of economic, social and cultural status.

Keywords: Item types, PISA, scientific literacy, multiple choice items, complex multiple choice items, constructed response items

FARKLI MADDE TÜRLERİNE GÖRE FEN OKURYAZARLIĞININ İNCELENMESİ: PISA 2015 TÜRKİYE ÖRNEĞİ

ÖZ

Bu araştırmanın amacı, farklı madde türlerine göre PISA 2015 fen okuryazarlığı bilişsel testinde yer alan maddelerin psikometrik özelliklerini ortaya koymak ve farklı madde türlerinden elde edilen fen okuryazarlığı puanlarını farklı bağımsız değişkenlere göre incelemektir. PISA 2015 Türkiye örnekleminde yer alan 5895 öğrenciden, araştırmanın amacı doğrultusunda 175 öğrenci çalışmaya dâhil edilmiştir. Araştırmaya iki fen kümesinde yer alan toplam 35 fen okuryazarlığı maddesi alınmıştır. Araştırmanın bulgularını elde edebilmek için betimsel istatistikler ve çeşitli hipotez testleri kullanılmıştır. Çoktan seçmeli (ÇS) maddelerin daha kolay olmasının yanında ayırt ediciliğinin oldukça düşük olduğu; cevabı yazılan (CY) maddelerin ise diğer madde türlerine göre zor ve ayırt ediciliğinin daha yüksek olduğu belirlenmiştir. Karmaşık çoktan seçmeli (KÇS) maddelerin ise ÇS maddeleri ile benzer bir madde güçlüğüne sahip olmasıyla birlikte, ayırt ediciliğinin ÇS maddelerine göre daha yüksek olduğu sonucuna varılmıştır. Üç ayrı madde türündeki fen okuryazarlığı ve madde güçlük ortalamaları incelendiğinde, fen okuryazarlık düzeyi yüksek olan öğrencilerin CY maddelerinde; fen okuryazarlık düzeyi düşük olan öğrencilerin ise ÇS maddelerinde daha başarılı olduğu belirlenmiştir. Bu çalışmada kullanılan madde türü sınıflandırması PISA 2015 uygulamasına aittir. Farklı madde türü sınıflandırmaları ile bu araştırma tekrar edilebilir. Ayrıca fen okuryazarlık düzeyi düşük, orta ve yüksek olan ülkelerden elde edilecek örneklemler üzerinden benzer bir çalışma yürütülebilir. Bu araştırma doğrultusunda Türkiye'de, özellikle merkezi sınavlarda farklı soru türlerinin birlikte kullanılması yönünde atılan adımların olumlu sonuçlar verebileceği düşünülmektedir. Farklı madde türlerinin fen başarısının belirlenmesinde birlikte kullanılması önerilmektedir.

Anahtar Kelimeler: Madde türü, PISA, fen okuryazarlığı, çoktan seçmeli maddeler, karmaşık çoktan seçmeli maddeler, cevabı yazılan maddeler

[1]*Dr. Öğr. Üy., Niğde Ömer Halisdemir Üniversitesi, Eğitim Fakültesi, esinyilmaz@ohu.edu.tr,* https://orcid.org/0000-0001-6755-9018
[2]*Doç. Dr., Akdeniz Üniversitesi, Eğitim Fakültesi, hkogar@gmail.com,* https://orcid.org/0000-0001-5749-9824

1. INTRODUCTION

Large-scale tests are used in many countries to determine student success at national and international levels. The items in these tests can be in various formats, such as multiple choice, true false, blank-filling, short response, and open-ended. These different types of items are generally categorized based on the type of response required: items necessitating the selection of the correct response (selected-response - SR) and items requiring a written response (constructed-response - CR) (Haladyna & Rodriguez, 2013; Osterlind, 1998). In SR items correct and incorrect alternatives are given and the test taker is required to select one of the alternatives. SR is the most frequently used item type in large-scale achievement tests (Bleske-Recheck, Zeug, & Webb, 2007; Osterlind, 1998). On the other hand, in CR items no alternatives are given; the test taker is required to express his/her response by means of words, statements, figures etc. CR items may entail tasks such as solving problems, writing essays, or drawing diagrams or graphs (Lukhele, Thissen, & Wainer, 1994).

The most widely used item types among SR items are multiple choice (MC) and true-false (TF) items (Osterlind, 1998). MC items include a question or an incomplete statement generally accompanied with a correct alternative and various numbers of incorrect alternatives called distractors (Haladyna & Rodriguez, 2013). On the other hand, TF items, which are a popular type of item among SR items, require selecting between two alternatives (Haladyna, 1992) or consist of a root and a series of alternative responses (Frisbie & Druva, 1986). Although TF items substantially resemble MC items in format, the fundamental difference between the two item types is that while there is only one correct response in MC items, TF items may include varying numbers of correct responses or true-false patterns (Dudley, 2006). TF items are quite often called complex multiple choice (CMC) items (Haberkorn, Pohl, & Carstensen, 2016; Frisbie & Sweeney, 1982). While SR items are generally scored in two categories (correct: 1; incorrect: 0), CR items include numerous categories and are scored ractionally (Kinsey, 2003).

Each item type has its own advantages and disadvantages. The advantages that multiple choice items provide can be listed as follows: (a) scoring the items and analysing these scores easily (Bible, Simkin, & Kuechler, 2008), (b) preventing students from losing points based on language deficiencies such as grammar, writing or punctuation errors (Zeidner, 1987), (c) being free of scorer bias (Walstad, 1998), (d) being able to construct the test with empirical proof (item analyses etc.) (Ben-Simon, Budescu, & Nevo, 1997), (e) being able to collect data from a large scale in an effective and easy way (Dufresne, Leonard, & Gerace, 2002). However, it also has disadvantages, such as the requirement of a large sample size to develop a MC test with a high degree of reliability (Bacon, 2003), the possibility of arriving at the correct answer by eliminating the other alternatives (Bush, 2001; Hobson & Ghoshal, 1996), the difficulty of preparing the test when there is no test bank at reach (Brown, Bull, & Pendlebury, 1997) and the fact that items of MC tests that are not written in accordance with test writing principles conceal students' knowledge rather than disclose it (Dufresne et al., 2002). As for the advantages and disadvantages of TF items, measuring educational success in a explicit, simple and effective way can be regarded as the main advantage, while the disadvantages can be stated as the categorical responses that necessitate justification but do not require verification and the possibility of arriving at the correct response by chance (Ebel, 1970). Furthermore, Frisbie and Sweeney (1982) state that students are more successful in responding to these items when compared with MC items, and that TF items are more advantageous than MC items as more responses are obtained in a shorter period of time with these items. However, the conception that CR items measure the ability to solve real life problems better than SR items (Bacon, 2003; Fenna, 2004; Hancock, 1994; Rogers & Hartley, 1999) has led to the use of CR items in large-scale items as well. On the other hand, the disadvantageous aspects can be listed as SR items' requiring great effort in ensuring objectivity (training of the raters/scorers in how to score each item etc.) (Bennett et al., 1990) and their having a lower level of content validity when compared to that of CS items.

Every kind of response type has both advantages and disadvantages. Hence, to eliminate the disadvantages inherent in each type of item, Martinez (1999) recommends that tests should include a combination of different types of items. Today, different types of items are commonly employed in many large-scale tests, and it is considered that the validity of assessments is thus increased (Griffo, 2011). Examples of these tests are National Assessment of Educational Progress (NAEP), Test of English as a Foreign Language (TOEFL), Trends In International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA). The present study focuses on PISA − a project of the Organisation for Economic Co-operation and Development (OECD), which is administered every three years to 15-year-old students who have received at least seven years of education. In PISA 2015, the focus of assessment was on scientific literacy, with reading skills, math literacy and collaborative problem solving as minor areas of assessment. In addition, PISA 2015 included the assessment of financial literacy, which was optional for countries and economies (OECD, 2016a). In PISA, either math or scientific literacy is identified as the primary domain and the remaining skills in the two domains are assessed comparatively and in a less detailed manner (OECD, 2009). In PISA 2015, the primary domain was scientific literacy. PISA 2015 consisted of a combination of multiple choice items and items requiring constructed responses.

Three classes of items are used to assess the competencies and scientific knowledge identified in the framework (OECD, 2016a, p.41):

- simple multiple choice: items calling for
  - selection of a single response from four options
  - selection of a "hot spot", an answer that is a selectable element within a graphic or text
- complex multiple choice: items calling for
  - responses to a series of related "Yes/No" questions that are treated for scoring as a single item (the typical format in 2006)
  - selection of more than one response from a list
  - completion of a sentence by selecting drop-down choices to fill multiple blanks
  - "drag-and-drop" responses, allowing students to move elements on screen to complete a task of matching, ordering or categorising
- constructed response: items calling for written or drawn responses: Constructed-response items in scientific literacy typically call for a written response ranging from a phrase to a short paragraph (e.g. two to four sentences of explanation). A small number of constructed-response items call for drawing (e.g. a graph or diagram). In a computer-based assessment, any such item is supported by simple drawing editors that are specific to the response required.

There are numerous studies in literature that investigate item type with respect to the gender variable. In some of these studies, it was found that males were more successful than females in responding to multiple choice items (Bell & Hay, 1987; Bolger & Kellaghan, 1990; Breland et al., 1994; Bridgeman & Lewis, 1994; Lumsden & Scott, 1987, 1995; Murphy, 1982; Walstad & Robinson, 1997; Zhang & Manon, 2000). However, in a study by Ghorpade and Lackritz (1998), no significant difference between the results of females and males in multiple choice items were found. As for items requiring constructed responses, in most studies that were conducted, it was found that females were more successful when compared to males (Bible et al., 2008; Bolger, 1984; Ghorpade & Lackritz, 1998; Harris & Kerby, 1997; Lumsden & Scott, 1987; Murphy, 1980, 1982; Schwabe, McElvany, & Trendtel, 2015; Zhang & Manon, 2000). However, some studies reported an insignificant impact of gender upon student success with respect to type of item (Chan & Kennedy, 2002; Bacon, 2003; Greene, 1997; Walstad & Becker, 1994). Arthur and Everaert (2011) examined the success level of bachelor and master's students' in a financial accounting exam by gender. In this study, it was found that among bachelor students, females were more successful than males in multiple choice and open-ended questions. On the other hand, in the master's group, it was found that, when compared to males, females showed a higher level of performance only in the open-ended questions. Beller and Gafni (2000) examined whether or not the performances of female and male students varied across multiple choice and open-ended responses in the math test of the International Assessment of Educational Progress (IAEP), conducted in the years 1988 and 1991. It was found that in the 1988 test the impact of gender was higher upon multiple choice items, whereas in the 1991 test the impact of gender was higher upon the open-ended items. Owing to these varying results, researchers suggested that the difference between the math performance in accordance with item type should not be accounted for solely by gender differences. According to this study in literature, there are no firm results regarding the relationship between item type and gender variables. In the present study, gender is treated as a variable. In addition, as stated by Beller and Gafni (2000), with the conception that students' performances in different types of items can be affected by other variables, students' economic, social and cultural status could be added as another variable since the findings of numerous studies indicated that this variable affected student success (Baker, Goesling, & LeTendre, 2002; Berberoğlu et al., 2003; Chiu, 2007; Klein, 1971; Koutsoulis & Campbell, 2001; Topçu & Yılmaz-Tüzün, 2009). However, no study examining both students' success regarding item type and their economic, social and cultural status were encountered in the related literature. For this reason, assuming that this variable can also have an impact on success in relation to item type, the variable of economic, social and cultural status was also included within the scope of the present study.

1.1. The Aim of the Study

The aim of the present study was to reveal the psychometric properties of the items in the cognitive test of PISA 2015 assessing scientific literacy according to different item types and to examine scientific literacy in relation to different independent variables. Thus, in this study the responses to the following research questions were sought:

1- What are the descriptive statistics of different item types used in the cognitive tests for science and math literacy and reading skills?

2- What are the descriptive statistics of different item types used to assess science content knowledge for scientific literacy?
3- Is there a significant difference between item difficulty indices and item discrimination indices of different items in the scientific literacy test?
4- What are the mean, standard deviation and item difficulty averages of students' scientific literacy levels according to three different item types?
5- Do students' mean scores and standard deviations of different item types vary by gender?
6- Do students' mean scores and standard deviations of different item types vary by the economic, social and cultural status (ESCS) index?

1.2. The Importance of the Study

There are numerous different reasons underlying the interest in studying different types of items. A study by Bible et al. (2008) lists these reasons as follows: reliable and valid assessment of students' knowledge, the access to course materials that ensure the designation of fair marks to students, the desire to develop tests that assess student expertise in an effective and objective way, the desire to receive feedback about the effectiveness of educational methods and pedagogies, and the preparation of new graduates for their fields of expertise. Furthermore, it is believed that different types of items activate different cognitive processes (Becker & Johnston, 1999; Bridgeman & Rock, 1993; Kennedy & Walstad, 1997; Kuechler & Simkin, 2003; Walstad & Becker, 1994). Hence, tests prepared using different types of items have gained increasing importance. In Turkey numerous exams are administered by ÖSYM (Student Selection and Placement Center). One of the these exams is LYS (University Entrance Exam). It has been declared in the ÖSYM 2017 guide that short response questions will be included in the exam as of the following year. But this approach has not been continued in the following years. The notion that using a single type of item in tests prevents gaining sufficient knowledge about the test taker has started to have an impact on many achievement tests. For these reasons, this study, which is based on different item types in use in Turkey, is regarded to be of significance for educational sciences.

2. METHOD

This study, which aimed to examine the distribution of different item types in the cognitive tests and the success rates of the different item types in the primary scientific literacy cognitive test in relation to the variables of gender and economic, social and cultural status, employed a survey design.

2.1. Population and Sample

540 students participated in PISA 2015 (OECD, 2016b) as representatives of 29 million 15-year-old students receiving education at schools across 72 participant countries and economies. A total of 5895 students from Turkey, who were selected according to the Classification of Statistical Region Units (SRU), participated in PISA 2015. These students, who were grouped based on the type of school they were enrolled in, were selected by the PISA International Center via random sampling from 187 schools representing 12 regions and 57 cities. In order to examine both the old and new scientific literacy items together, the first 18 science sets were taken into consideration. A set selected randomly from these 18 sets formed the sample of this study. Science set number 2 was included in the sample of the research study. The forms included in this set were Form-32, Form-41, Form-44, Form-53, Form-56, Form-65, Form-77, Form-80, Form-89 and Form-92. The total number of students included in the sample was 175. As sample selection was done based on test design, detailed information is presented in the section on data collection tools.

2.2. Data Collection Tools

In PISA 2015, the cognitive tests were administered via either paper-and-pencil or computer-based testing. While the paper-and-pencil assessments included reading, science and math items that were formerly used, the computer based assessments included new items. PISA 2015 was administered as either paper-and-pencil or computer-based testing depending on the country's preference. There was a total of 96 different forms – 30 different forms for the paper-and-pencil test and 66 different forms for the computer-based test. As the administration for the sample from Turkey was made via the computer, from hereby information about the computer-based administration will be given.

The computer-based administration was designed as a two-hour test. As the primary domain was scienc literacy in the 2015 PISA administration, the students completed the test by allocating one hour to the assessment of science and another hour − 30 minutes each − to any two of math, reading, or collaborative problem solving assessments. In the PISA 2015 cognitive test administration, a different test design was employed from the previous implementations. The items were placed in the forms of the computer-based implementation as follows:

1. The forms included four groups of items. While two of these groups definitely included items related to scientific literacy, the other two groups consisted of items related to math literacy, reading literacy and/or collaborative problem solving skills.

2. Two sets making up the scientific literacy was subdivided into 12 different sets (F1-F2). These 12 sets formed 36 possible science sets with varying pairwise rotations.

Table 1 presents information regarding combination of science sets of scientific literacy items (OECD, 2016c, pp.16-17). In the first 18 science sets, there were both old and new scientific literacy items, in sets between 19-33, there were only new scientific literacy items and in sets 34 and 36, there were only old scientific literacy items.

Table 1.
Combinations of Science Clusters Belonging to Scientific Literacy Items

N	**S**	**S**	**N**	**S**	**S**	**N**	**S**	**S**	**N**	**S**	**S**	**N**	**S**	**S**	**N**	**S**	**S**
1	01	07	**7**	04	10	**13**	09	02	**19**	07	08	**25**	09	11	**31**	12	07
2*	**01**	**10**	**8**	05	11	**14**	09	06	**20**	08	09	**26**	10	07	**32**	12	09
3	02	08	**9**	06	12	**15**	10	03	**21**	07	11	**27**	10	09	**33**	12	11
4	03	09	**10**	07	06	**16**	11	02	**22**	08	10	**28**	10	12	**34**	02	04
5	03	12	**11**	08	01	**17**	11	04	**23**	08	12	**29**	11	08	**35**	05	01
6	04	07	**12**	08	05	**18**	12	05	**24**	09	08	**30**	11	10	**36**	06	03

*Selected combination in research

3. There were six different groups in each of the test forms in these 36 science sets with 66 test forms placed randomly. Every student was randomly assigned a number between 1 and 6. By taking into consideration the student's form number and the randomly assigned number, the two sets of science items the student was to answer was determined. For more and detailed information, the OECD document (2016c) can be referred to.

The least administered form in Turkey was Form-75, which was administered to 14 students, while the most administered form was Form-93, which was administered to 230 students.

The properties of the items in the F1 and F10 science sets, selected for this study, are presented in Table 2. In the two science sets selected for this study, there were eight (five old, three new) multiple choice items, 16 (eight old, eight new) complex multiple choice items and 11 (five old and six new) open-ended items. In these science sets, there was a total of 35 scientific literacy items.

Table 2.
Distribution of Items in Selected Set Combination for Research

Item Type	**F01 (old)**		**F10 (new)**		**Total**	
	f	**%**	**f**	**%**	**f**	**%**
Multiple Choice (MC)	5	27.78	3	17.65	8	100.00
Complex Multiple Choice (CMC)	8	44.44	8	47.06	16	100.00
Constructed Response (CR)	5	27.78	6	35.29	11	100.00
Total	18	100.00	17	100.00	35	100.00

2.3. Analysis of Data

Different descriptive statistics were computed for the total scores and items obtained from the different item types of the PISA 2015 scientific literacy. Whether or not there was a significant difference between item difficulty indices and item discrimination indices of the items in the scientific literacy tests of different item types was examined using the Kruskal Wallis H test. This test was preferred since the number of items was small. Moreover, whether average scores obtained from different test types varied by gender and economic, social and cultural status was examined by means of t tests for independent sample groups. Whether average scores obtained from each item type varied depending on the level of the student was examined via one-way ANOVA. The significance level was set at .05. For significant differences, effect sizes were calculated.

In order to compare the scores obtained from different types of items, the responses given to MC, CMS and CR items were rescored, assigned 1 to completely correct answers and 0 to all other possible responses. All the scores and average scores for each item type was calculated in this way. The economic, social and cultural status index for Turkey was -1.448 on average. Those students who received an index value equal to or below this value was defined as the "lower" status group and those who scored a higher value was defined as the "higher" status group.

3. FINDINGS

The descriptive statistics for the different item types used in the cognitive tests for scientific literacy, math literacy and reading skills are presented in Table 3. There were 103 items assessing reading skills in PISA 2015. 34.95% of these items were MC, 11.65% were CMC and 53% were CR items. In the cognitive test where reading skills were assessed, it was observed that the highest number of items were CY and the lowest number of items were CMC items. As for the 81 items assessing math literacy, 24.70% consisted of MC items, 17.28% consisted of CMC items and 58.02% consisted of CR items. In the math literacy cognitive test, the highest number of items were SR type of items, then followed MC and the lowest number of items were CMC. The primary domain of scientific literacy in PISA 2015 consisted of 184 items. Of these, 29.35% were MC, 35.87% were CMC and 34.78% were CR items. In the scientific literacy test, the highest number of items were CMC and the lowest number was MC items. In PISA 2015 there was a total of 368 items assessing reading skills, and science and math literacy. Of these, 29.89% were MC items, 25.00% were CMC items and 45.11% were CR items. In the cognitive tests assessing these three domains, the highest number of items were MC and the lowest number of items were CMC items. In all three of the cognitive tests, that the number of CR item type was the highest was an interesting finding.

Table 3.
Distribution of Items Used in PISA 2015 Cognitive Test

Item Type	Reading		Mathematic		Science		Total	
	f	%	f	%	f	%	f	%
Multiple Choice	36	34.95	20	24.70	54	29.35	110	29.89
Complex Multiple Choice	12	11.65	14	17.28	66	35.87	92	25.00
Constructed Response	55	53.40	47	58.02	64	34.78	166	45.11
Total	103	100.00	81	100.00	184	100.00	368	100.00

The distribution of the item types used to assess the science content knowledge for scientific literacy is presented in Table 4. When Table 4 is examined, it can be observed that in the cognitive test for scientific literacy, the "Living Systems" knowledge of science category had the highest number of items. Of these, 31.08% were MC and CMC items, and 37.84% consisted of CR items. Of the 49 items assessing earth and space systems, 12.24% were MC items, 53.06% were CMC items and 34.69% consisted of CR items. Of the 61 items assessing physical systems, 40.98% were MC items, 27.87% were CMC items and 31.15% consisted of CY items. In the cognitive tests assessing these three science categories of knowledge, the highest number of items was CMC and the lowest number was MC type of items.

Table 4.
Distribution of Items Used in Content Knowledge of Science

Item Type	Living Systems		Earth and Space Systems		Physical Systems		Total	
	f	%	f	%	f	%	f	%
Multiple Choice	23	31.08	6	12.24	25	40.98	54	29.35
Complex Multiple Choice	23	31.08	26	53.06	17	27.87	66	35.87
Constructed Response	28	37.84	17	34.69	19	31.15	64	34.78
Total	74	100.00	49	100.00	61	100.00	184	100.00

The findings obtained from the examination of the item difficulty indices of the total 35 items in the two science groups selected for this study are presented in Table 5. According to the responses given to the eight MC items assessing scientific literacy, the average item difficulty index was .42; according to the responses given to the 16 CMC items, the average item difficulty index was .36 and according to the responses given to the 11 CR items, the item difficulty average was .23. The most difficult type of item in the test was the MC, while the easiest type of item was the CMC. While the item difficulty indices of CMC and CR items showed a normal distribution, the item difficulty indices for MC did not portray a normal distribution. According to the findings of the Kruskal Wallis H test, it was found that the item difficulty indices of the items in the scientific literacy test varied across item types ($p<.05$). It was found that the item difficulty indices for the MC and CMC items were higher than those of CR items. MC and CMC items were easier than CR items.

Table 5.
Item Difficulty of Different Item Type

Descriptive Statistics	MC	CMC	CR	Scientific Literacy	Kruskal Wallis H	p	Significant Difference
N	8	16	11	35			
Mean	.42	.36	.23	.33			
Std Dev.	.21	.17	.10	.18			
Ranj	.69	.61	.37	.84			MC – CR
Variance	.05	.03	.01	.03	6.65	.036	CMC – CR
Minimum	.21	.06	.07	.06			
Maximum	.90	.67	.44	.90			
Skewness	1.84	-.04	.71	.99			
Kurtosis	4.09	-.68	.66	1.66			

The findings obtained as a result of the examination of the item difficulty indices of the total 35 items in the two science sets selected for this study are presented in Table 6. The item discrimination indices for MC, CMC and CR items were .13, .25 and .39, respectively. The highest level of item discrimination was observed for CR items, while the lowest level was for MC items. While the item discrimination indices for CMC and CR items showed a normal distribution, those for MC did not display a normal distribution. According to the findings of the Kruskal Wallis H test, it was observed that the item discrimination indices of the items in the scientific literacy test varied by item type ($p<.01$). It was found that the item discrimination indices for CR items were higher than those for MC and CMC items, and the item discrimination indices for CMC were higher than those for MC items. As such, the items that had the highest discrimination power were CR items and the ones with the lowest discrimination power were the MC items. This significant difference had a high effect size (eta-square=.15).

Table 6.
Item Discrimination of Different Item Type

Descriptive Statistics	MC	CMC	CR	Scientific Literacy	Kruskal Wallis H	p	Significant Difference
N	8	16	11	35			
Mean	.13	.25	.39	.27			
Std Dev.	.09	.10	.09	.13			
Ranj	.23	.33	.27	.55			CR – MC
Variance	.01	.01	.01	.02	18.74	.000	CR – CMC
Minimum	.02	.11	.29	.02			CMC – MC
Maximum	.25	.44	.57	.57			
Skewness	.23	.57	.87	.11			
Kurtosis	-1.97	-.32	.09	-.36			

The means, standard deviations and item difficulty averages of all three different types of items according to the students' scientific literacy levels are presented in Table 7. A large majority of the students included in the sample of this study were at level 1a or level 2 of scientific literacy. There was only one student who was above level 3 in scientific literacy. The students with the highest success levels in MC and CMC items were at level 2 and the student with the highest success level in CR items was at level 5. When MC items were examined, it was observed that the students at levels 1a and 2 in scientific literacy were more successful than the students at the other levels. Students at levels 2 and 5 in scientific literacy in CMC items were more successful than those at other levels. Students at levels 3 and 5 in scientific literacy in CR items were more successful than those at other levels. Among all the different types of items, the item type for which the highest number of correct answers were given was MC (45%) and these students were at level 2. On the other hand, the item type for which the lowest number of correct answers were given was CR (21%) and again these students were at level 2. The student at level 5 in scientific literacy had answered 83 of the MC and CMC items and 36% of the CR items correctly. This significant difference had a high effect size (eta-square=.52).

Table 7.
Means, Standard Deviations and Average of Item Difficulties of Different Item Types According to The Level of Scientific Literacy of Students

Level of Scientific literacy of Students	Item Type									N
	Multiple Choice			Complex Multiple Choice			Constructed Response			
	$\bar{X}$	SD	p	$\bar{X}$	SD	p	$\bar{X}$	SD	p	
1b	3.26	1.10	.41	5.95	2.39	.37	2.42	2.36	.22	19
1a	3.41	1.36	.43	5.41	2.69	.34	2.41	2.25	.22	68
2	3.56	1.43	.45	6.32	2.97	.40	2.34	2.39	.21	59
3	3.11	1.29	.39	5.89	2.41	.37	2.61	2.17	.24	28
4	-	-	-	-	-	-	-	-	-	-
5	3.00	-	.38	6.00	-	.38	4.00	-	.36	1
6	-	-	-	-	-	-	-	-	-	-

Whether or not the average scores obtained for each item type varied depending on scientific literacy level of students – excluding the student whose scientific literacy level was 5 – was examined via one-way ANOVA. According to the ANOVA findings, no significant difference was found among the MC ($F_{(3\text{-}170)=}$.78, p>.05), the CMC ($F_{(3\text{-}170)}$=1.19, p>.05) and the CR ($F_{(3\text{-}170)}$=.09, p>.05) types of items with respect to student levels.

The findings regarding the examination of average student scores obtained from different types of items with respect to the gender variable are presented in Table 8. Male students were found to be more successful than female students in responding to MC and CMC items. On the other hand, female students were found to be more successful than male students in responding to CR items. Whether or not average scores obtained from each type of item varied by gender was examined using t tests for independent samples. No significant difference was found by gender with respect to MC ($t_{(173)}$=1.00, p>.05), CMC ($t_{(173)}$=.04, p>.05) and CR ($t_{(173)}$=.41, p>.05) types of items.

Table 8.
Average Student Scores Obtained from Different Types of Items with Respect to he Gender Variable

Item Type	Gender	N	$\bar{X}$	SD	t	df	p
MC	**Female**	86	3.29	1.41	1.00	173	.317
	Male	89	3.49	1.27			
CMC	**Female**	86	5.85	2.71	.04	173	.968
	Male	89	5.87	2.74			
CR	**Female**	86	2.50	2.42	.41	173	.684
	Male	89	2.36	2.13			

The findings obtained as a result of the examination of average student scores of different types of items in relation to the economic, social and cultural status index are presented in Table 9. Whether or not average scores obtained from each type of item varied by economic, social and cultural status was examined using t tests for independent samples. No significant difference was found economic, social and cultural status with respect to MC ($t_{(173)}$=2.43, p<.05), CMC ($t_{(173)}$=3.30, p>.01) and CR ($t_{(173)}$=3.52, p>.01) types of items. The students with a high economic, social and cultural status were found to be more successful when compared to those with a lower level of economic, social and cultural status. The significant findings as regards the CMC (eta-square=.06) and CR (eta-square=.07) items had a moderate effect size, while those regarding the MC (eta-square=.03) items had a small effect size.

Table 9.
Average Student Scores Obtained from Different Types of Items with Respect to the ESCS Variable

Item Type	ESCS	N	$\bar{X}$	SD	t	df	p
MC	**Low**	87	3.15	1.34	2.43	173	.016
	High	88	3.64	1.36			
CMC	**Low**	87	5.20	2.50	3.30	173	.001
	High	88	6.51	2.77			
CR	**Low**	87	1.84	2.09	3.52	173	.001
	High	88	3.01	2.31			

4. DISCUSSION and RESULTS

The present study aimed to reveal the psychometric properties of the items in the scientific literacy cognitive test of PISA 2015 according to different item type and examine scientific literacy based on different independent

variables. When the descriptive statistics of different types of items used in the science and math literacy and the reading skills cognitive tests were examined, it can be concluded that one of the most frequently used item types in all three cognitive tests was CR items. This can be accounted for with the recent growing interest laid upon item types in which students provide a written response in wide-scale exams. It can be observed that along with MC item types, the TIMSS 2015 implementation also included many item types where a written response was required.

The primary domain of assessment in PISA 2015 was scientific literacy. For this reason, distributions of the different types of items used within the scope of science content knowledge of scientific literacy were also taken into consideration. In the scientific literacy cognitive test, items related to the following content knowledge are included: "Living Systems," "Earth and Space Systems," and "Physical Systems". While the highest number of items used in the science content knowledge of living systems was CR items, those for the earth and space systems was CMC items and those for the physical systems was MC items. This may be accounted for with the fact that some item types may be more effective in assessing different scopes of science content knowledge.

In the light of the findings obtained from the analysis of difficulty indices of the items in the scientific literacy test for different item types, it was found that the type of item students were most successful at answering was MC, while the type of item they were least successful at answering was the CR. In a study by O'Leary (2001), in which the average scientific literacy yielded by the TIMSS 1995 was examined according to different item types, all the 12 countries which was included in the study yielded similar findings. In a study where MC and CR item types were used, Thawabieh (2016) found that MC items were easier than the CR type of item. It was found that the item type that had the highest discrimination power in the scientific literacy test was the CR items and those with the lowest discrimination power was the MC item types. Aydın (1996), Kan and Kayapınar (2006) and Uğurlu (1993) found that items to which written responses were required had a higher discrimination power than those whose answers were to be selected. It was found that MC items were easier and had a lower level of discrimination power, while CR items were more difficult and had a higher discrimination power. As for CMC items, even though they were found to have a similar level of item difficulty with that of MC items, it was found that they had a higher discrimination power than that of MC items.

The scientific literacy level of the students within the sample of the research study was found to be 3 or lower. When scientific literacy and item difficulty averages for all three item types were examined, students with a high level of scientific literacy level were more successful at responding to CR items, while students with a low level of scientific literacy were more successful at answering MC items. However, the success scores obtained by students for each item type did not vary by level of the student.

As a result of the examination of student average scores for each item type by gender, it was found that male students were more successful than female students in responding to MC and CMC items, while female students were more successful than male students in answering CR items. In related literature, there are studies reporting a higher success level of male students in MC items (Bell & Hay, 1987; Bolger & Kellaghan, 1990; Bridgeman & Lewis, 1994; DeMars, 2000; Zhang & Manon, 2000). In some other studies female students were found to be more successful than male students in items requiring written responses (Bible et al., 2008; Bolger, 1984; DeMars, 2000; Ghorpade & Lackritz, 1998; Harris & Kerby, 1997; Lumsden & Scott, 1987; Zhang & Manon, 2000). The reason why female students are more successful in open-ended items can be accounted for with female students' ability in verbal expression (Beller & Gafni, 2000). Moreover, the interaction between item type and gender can derive from different responding strategies or risk-taking tendencies (Beller & Gafni, 2000). However, in this study no significant difference was found among average scores by gender. Similar findings were also encountered in other studies (Chan & Kennedy, 2002; Bacon, 2003; Ghorpade & Lackritz, 1998; Greene, 1997; Walstad & Becker, 1994). Similarly, in a study by Bacon (2003), no significant difference was found by gender and this finding was accounted for with the difference between students' talents and that verbal ability, problem solving ability and test solving ability may be distributed equally across the two genders. A similar situation may be valid in the present study as well.

It was found that students with a high level of economic, social and cultural status were more successful than those with a low level of economic, social and cultural status. In PISA, economic, social and cultural status is an index value consisting of accommodation opportunities, and the educational level and occupation of the parents. It is reported in various studies that economic, social and cultural status has an impact on success in sciences (Baker et al., 2002; Berberoğlu et al., 2003; Chiu, 2007; Klein, 1971; Koutsoulis & Campbell, 2001; Topçu & Yılmaz-Tüzün, 2009).

In the present study it was identified that each item type has its own positive and negative aspects. Students who are unsuccessful at answering MC, but successful at responding to CR items reach a higher level of success. For this reason, it is recommended that different types of items be used in combination to determine success in science. When the findings of the present study is taken into consideration, it is considered that the steps taken in Turkey, particularly the combined use of different types of questions in standard exams, could yield positive outcomes.

This study focused on scientific literacy. A similar study can be carried out in the area of math literacy and reading skills. That the scientific literacy levels of the students in the present study, which was conducted on PISA 2015 with a sample from Turkey, was low may lead to a shrinkage in range. This could be given as the reason underlying the non-significant findings in the hypothesis tests. Hence, this study could be replicated in countries where samples with low, moderate and high levels of scientific literacy can be obtained. The item type classification used in the present study belongs to PISA 2015. However, this study could be replicated with the use of a different classification of item types.

REFERENCES

Arthur, N., & Everaert, P. (2011). Gender and performance in accounting examinations: exploring the ımpact of examination format. *Accounting Education*, *21*(5), 471-487.

Ayılın, B. (1993). *Seçme gerektiren test maddeleri ile kısa cevap gerektiren test maddelerinin psikometrik özellikleri ve öğrenci başarısı bakımından karşılaştırılması*. Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31-36.

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley effect" on mathematics and science achievement. *Comparative education review*, *46*(3), 291-312.

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record*, *75*(4), 348-357.

Bell, R., & Hay, J. (1987). Differences and biases in English language examination formats. *British Journal of Educational Psychology*, *57*(2), 212–220.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1-2), 1-21.

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*(1), 65-88.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement*, *14*(2), 151-162.

Berberoğlu, G., Çelebi, Ö., Özdemir, E., Uysal, E. ve Yayan, B. (2003). Üçüncü uluslararası matematik ve fen çalışmasında Türk öğrencilerinin başarı düzeylerini etkileyen etmenler. *Eğitim Bilimleri ve Uygulama, 2*(3), 3-14.

Bible, L., Simkin, M. G., & Kuechler, W. L. (2008). Using multiple-choice tests to evaluate students' understanding of accounting. *Accounting Education: An International Journal, 17*, 55-68.

Bleske-Rechek, A., Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment & Evaluation in Higher Education*, *32*(2), 89-105.

Bolger, N. (1984). *Gender differences in academic achievement according to method of measurement*. 92nd The Annual Meeting of the American Psychological Association, August, Toronto, Rio.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 165–174.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement history examination. *Journal of Educational Measurement*, *31*(4), 275-293.

Bridgeman, B. & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, *31*(1), 37–50.

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple–choice and open–ended analytical questions. *Journal of Educational Measurement*, *30*(4), 313-329.

Brown, G. A., Bull, J., & Pendlebury, M. (2013). *Assessing student learning in higher education*. London: Routledge.

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher education*, *25*(2), 157-163.

Chan, N., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, *68*(4), 957-971.

Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries: country-, school-, and student-level analyses. *Journal of Family Psychology*, *21*(3), 510.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true–false item type under norm-referenced conditions. *Language Testing*, *23*(2), 198–228.

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Marking sense of students' answers to multiple-choice questions. *The Physics Teacher*, *40*(3), 174-180.

Ebel, R. L. (1970). The case for true-false test items. *The School Review*, *78*(3), 373-389.

Fenna, D. S. (2004). Assessment of foundation knowledge: are students confident in their ability?. *European journal of engineering education*, *29*(2), 307-312.

Frisbie, D. A., & Druva, C. A. (1986). Estimating the reliability of multiple true-false tests. *Journal of Educational Measurement*, *23*(2), 99-105.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement*, *19*(1), 29-35.

Ghorpade, J., & Lackritz, R. (1998). Equal opportunity in the classroom: test construction in a diversity-sensitive environment. *Journal of Management Education*, *22*(4), 452–471.

Greene, B. (1997). Verbal abilities, gender, and the introductory economics course: A new look at an old assumption. *The Journal of Economic Education*, *28*(1), 13-30.

Griffo, V. (2011). *Examining NAEP: The effect of item format on struggling 4th graders' reading comprehension.* Unpublished Doctoral Dissertation in Special Education. University of California, Berkeley, USA.

Haberkorn, K., Pohl, S., & Carstensen, C. (2015). Incorporating different response formats of competence tests in an IRT model. *Psychological Test and Assessment Modeling*, *58*(2), 223-252.

Haladyna, T. M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education*, *5*(1), 73-88.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of experimental education*, *62*(2), 143-157.

Harris, R. B., & Kerby, W. C. (1997). Statewide performance assessment as a complement to multiple-choice testing in high school economics. *The Journal of Economic Education*, *28*(2), 122-134.

Hobson, A., & Ghoshal, D. (1996). Flexible scoring for multiple-choice exams. *The Physics Teacher*, *34*(5), 284-284.

Kan, A. ve Kayapınar, U. (2010). Yabancı dil eğitiminde aynı davranışları yoklayan çoktan seçmeli ve kısa cevaplı iki testin madde ve test özelliklerinin karşılaştırılması. *Eğitim ve Bilim*, *32*(142), 65-71.

Kennedy, P., & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied Measurement in Education*, *10*(4), 359-375.

Kinsey, T. L. (2003). *A comparison of IRT and rasch procedures in a mixed-item format test.* Doctoral dissertation, University of North Texas.

Klein, C. A. (1971). Differences in science concepts held by children from three social-economic levels. *School Science and Mathematics*, *71*(6), 550-558.

Koutsoulis, M. K., & Campbell, J. R. (2001). Family processes affect students' motivation, and science and math achievement in Cypriot high schools. *Structural Equation Modeling*, *8*(1), 108-127.

Kuechler, W. L., & Simkin, M. G. (2003). How well do multiple choice tests evaluate student understanding in computer programming classes?. *Journal of Information Systems Education*, *14*(4), 389-400.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected ıtems on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234-250.

Lumsden, K. G., & Scott, A. (1987). The economics student reexamined: Male-female differences in comprehension. *Research in Accounting Education*, *18*(4), 365–375.

Lumsden, K. G., & Scott, A. (1995). Economics performance on multiple-choice and essay examinations: a large-scale study of accounting students. *Accounting Education: an international journal*, *4*(2), 153–167.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207-218.

Murphy, R. J. L. (1980). Sex differences in GCE examination entry statistics and success rates. *Educational Studies*, *6*(2), 169–178.

Murphy, R. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology*, *52*(2), 213–219.

OECD (Organisation for Economic Cooperation and Development) (2009). *PISA 2006 technical report.*

OECD, (2016a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy.* Paris: OECD Publishing.

OECD, (2016b). *PISA 2015 PISA results in focus.* Paris: OECD Publishing.

OECD, (2016c). *PISA 2015 technical report. Chapter 2: Test design and development*. Paris: OECD Publishing.

Osterlind, S. J. (1998*). Constructing test items: Multiple-choice, constructed-response, performance, and other formats.* Dordrecht, Netherlands: Kluwer Academic.

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, *112*(4), 1137-1162.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*(2), 234-247.

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly, 50*(2), 219-232.

Thawabieh, A. M. (2016). A comparison between two test item formats: Multiple-choice items and completion items. *British Journal of Education*, *4*(8), 32-43.

Topçu, M. S. ve Yılmaz-Tüzün, Ö. (2009). Elementary students' metacognition and epistemological beliefs considering science achievement, gender and socioeconomic status. *Elementary Education Online, 8*(3), 676-693.

Uğurlu, N. (1996). *Aynı davranışı ölçmeye yönelik kısa cevaplı, çoktan seçmeli ve kompozisyon tipi maddelerden oluşan testlerin özelliklerinin incelenmesi.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Walstad, W. B. (1998). Multiple choice tests for the economics course. W. B. Walstad & P. Saunder (Eds.). In *Teaching Undergraduate Economics: A Handbook for İnstructors* (pp. 287-304), New York: McGraw-Hill.

Walstad, W. B., & Becker, W. E. (1994). Achievement differences on multiple-choice and essay tests in economics. *The American Economic Review*, *84*(2), 193-196.

Walstad, W. & Robinson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *Journal of Economic Education*, *28*(2), 155–171.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *The Journal of Educational Research, 80*(6), 352-358.

Zhang, L., & Manon, J. (2000) *Gender and achievement-understanding gender differences and similarities in mathematics assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, April 2000 (pp. 24-28). New Orleans, LA.

GENİŞLETİLMİŞ ÖZET

1. Giriş

Birçok ülkede ulusal ve uluslararası düzeyde öğrenci başarılarını belirlemek için geniş ölçekli testler kullanılmaktadır. Bu testlerde yer alan maddeler çoktan seçmeli maddeler, doğru-yanlış maddeleri, boşluk doldurma maddeleri, kısa cevaplı maddeler, açık uçlu maddeler gibi çeşitli türlerde olabilmektedir. Bu madde türleri genel olarak verilen cevaplar arasından öğrencinin doğru olanı seçmesini gerektiren maddeler – cevabı seçilen, CS (selected-response - SR) ve cevabı öğrencinin kendisinin yazmasını gerektiren maddeler – cevabı yazılan, CY (constructed-response - CR) olarak sınıflanmaktadır (Haladyna ve Rodriguez, 2013; Osterlind, 1998). CS maddelerinde doğru ve yanlış seçenekler bir arada bulunur ve bireyden bu seçeneklerden birini işaretlemesi istenir. CS, geniş ölçekli başarı testlerinde en çok kullanılan madde türüdür (Bleske-Recheck, Zeug ve Webb, 2007; Osterlind, 1998). CY maddelerinde ise herhangi bir seçenek bulunmaz, bireyden cevabını kelimelerle, cümlelerle, şekillerle vb. ifade etmesi beklenir. CY maddeleri çok adımlı bir matematik problemini çözme, deneme yazma, diyagram ya da grafik çizme gibi görevleri içerebilir (Lukhele, Thissen ve Wainer, 1994). Madde türü ile ilgilenilmesinin birçok farklı nedeni bulunmaktadır. Bible ve arkadaşları (2008) yaptıkları çalışmada bu nedenleri öğrenci bilgisini geçerli ve güvenilir bir şekilde değerlendirebilmek, öğrencilere adil notların verilmesini sağlayacak ders materyallerine ulaşılmasını sağlamak, etkili ve objektif bir tavırla öğrenci uzmanlığını değerlendirecek testler geliştirme isteği, eğitimsel yöntemlerin ve pedagojinin etkililiği hakkında geri bildirim alma isteği, yeni mezunların uzmanlık alanlarına girişlerine onları hazırlama amacı şeklinde ifade etmiştir. Ayrıca farklı madde türlerinin farklı bilişsel süreçleri aktif hale getirdiği de düşünülmektedir.

Bu araştırmanın amacı, farklı madde türlerine göre PISA 2015 fen okuryazarlığı bilişsel testinde yer alan maddelerin psikometrik özelliklerini ortaya koymak ve farklı madde türlerinden elde edilen fen okuryazarlığı puanlarını farklı bağımsız değişkenlere göre incelemektir.

2. Yöntem

PISA 2015 fen okuryazarlığı bilişsel testinde yer alan maddelerin psikometrik özelliklerini ortaya koymayı ve farklı madde türlerinden elde edilen fen okuryazarlığı puanlarını farklı bağımsız değişkenlere göre incelemeyi amaçlayan bu araştırma tarama türündedir. Türkiye'de PISA 2015 uygulamasına, İstatistikî Bölge Birimleri Sınıflaması (İBBS) Düzey 1'e göre belirlenen 12 bölgeyi temsil eden 57 ilden ve okul türlerine göre tabakalandırılarak PISA uluslararası merkez tarafından seçkisiz yöntemle belirlenen 187 okuldan, toplam 5895 öğrenci katılmıştır. Bu çalışmada eski ve yeni fen okuryazarlığı maddelerini beraber inceleyebilmek için ilk 18 fen kümesi dikkate alınmıştır. Bu 18 kümeden seçkisiz seçilen bir küme bu araştırmanın örneklemini oluşturmuştur. Araştırma için seçilen iki fen kümesinde, beşi eski üçü yeni olmak üzere sekiz çoktan seçmeli; sekizi eski ve sekizi yeni olmak üzere 16 karmaşık çoktan seçmeli ve beşi eski altısı yeni olmak üzere toplam 11 cevabı yazılan madde yer almaktadır. Bu fen kümelerinde toplam 35 fen okuryazarlığı maddesi yer almaktadır.

PISA 2015 fen okuryazarlığına ait farklı madde türlerinden elde edilen madde ve toplam puanlara ait betimsel istatistikler hesaplanmıştır. Farklı madde türlerine ait fen okuryazarlığı testinde yer alan maddelerin madde güçlük indeksleri ve madde ayırt edicilik indeksleri arasında anlamlı bir farklılık olup olmadığı, madde sayısının az olması nedeniyle Kruskal Wallis H testi ile incelenmiştir. Ayrıca, her bir madde türünden elde edilen puan ortalamalarının cinsiyet ile ekonomik, sosyal ve kültürel durum değişkenlerine göre farklılaşıp farklılaşmadığı bağımsız örneklemler için t testi ile incelenmiştir. Her bir madde türünden elde edilen puan ortalamalarının fen okuryazarlığı öğrenci düzeyine göre farklılaşıp farklılaşmadığı tek yönlü ANOVA ile incelenmiştir. Anlamlılık düzeyi 0,05'tir. Anlamlı bulunan farklılıklar için etki büyüklükleri hesaplanmıştır.

3. Bulgular, Tartışma ve Sonuç

Bu maddelerin %29,89'u çoktan seçmeli (ÇS) maddelerinden, %25,00'i karmaşık çoktan seçmeli (KÇS) maddelerinden ve %45,11'i cevabı yazılan (CY) maddelerinden oluşmaktadır. Bu üç alanın ölçüldüğü bilişsel testlerde en çok CY maddeleri en az ise KÇS maddeleri yer almaktadır. Bilişsel testlerin üçünde de CY madde türünün sayısının yüksek olması dikkat çekici bir bulgudur.

Fen okuryazarlığını ölçen sekiz ÇS maddesine verilen cevaplara göre madde güçlük indeksi ortalaması 0,42; 16 KÇS maddesine verilen cevaplara göre madde güçlük indeksi ortalaması 0,36 ve 11 CY maddesine verilen cevaplara göre madde güçlük indeksi ortalaması 0,23'tür. Testteki en kolay madde ÇS madde türünde, en zor madde ise KÇS madde türünde yer almaktadır. KÇS ve CY maddelerine ait madde güçlük indeksleri normal dağılım gösterirken ÇS maddelerinin güçlük indeksleri normal dağılmamaktadır. Kruskal Wallis H testine ait bulgulara göre fen okuryazarlığı testinde yer alan maddelerin madde güçlük indekslerinin madde türlerine göre farklılaştığı belirlenmiştir ($p<0,05$). Madde ayırt edicilik indeksi ortalamaları, ÇS maddeleri için 0,13, KÇS

maddeleri için 0,25 ve CY maddeleri için 0,39 olarak belirlenmiştir. Testteki en çok ayırt edici madde CY madde türünde, en az ayırt edici madde ise ÇS madde türünde yer almaktadır. KÇS ve CY maddelerine ait madde ayırt edicilik indeksleri normal dağılım gösterirken ÇS maddelerinin ayırt edicilik indeksleri normal dağılmamaktadır. Kruskal Wallis H testine ait bulgulara göre fen okuryazarlığı testinde yer alan maddelerin madde ayırt edicilik indekslerinin madde türlerine göre farklılaştığı belirlenmiştir ($p<0,01$).

ANOVA bulgularına göre, ÇS madde türünde ($F(3\text{-}170)=0,78$, $p>0,05$); KÇS madde türünde ($F(3\text{-}170)=1,19$, $p>0,05$) ve CY madde türünde ($F(3\text{-}170)=0,09$, $p>0,05$) öğrenci düzeyine göre anlamlı bir farklılık bulunmamıştır. ÇS madde türünde ($t(173)=1,00$, $p>0,05$); KÇS madde türünde ($t(173)=0,04$, $p>0,05$) ve CY madde türünde ($t(173)=0,41$, $p>0,05$) cinsiyete göre anlamlı bir farklılık bulunmamıştır. KÇS madde türünde ($t(173)=3,30$, $p<0,01$) ve CY madde türünde ($t(173)=3,52$, $p<0,01$) ekonomik, sosyal ve kültürel duruma göre anlamlı bir farklılık bulunmuştur.

Öğrencilerin madde türlerine ait puan ortalamalarının cinsiyet değişkenine göre incelenmesi sonucunda erkek öğrenciler ÇS ve KÇS maddelerinde kız öğrencilere göre daha başarılı bulunurken kız öğrenciler ise CY maddelerinde erkek öğrencilere göre daha başarılı bulunmuştur. Üst ekonomik, sosyal ve kültürel duruma sahip öğrencilerin tüm madde türlerinde alt ekonomik, sosyal ve kültürel duruma sahip öğrencilere göre daha başarılı olduğu belirlenmiştir.

Bu çalışmada her madde türünün olumlu ve olumsuz yönleri olduğu belirlenmiştir. ÇS maddelerinde başarısız öğrenciler, CY maddelerinde ise başarılı öğrenciler daha başarılı olmaktadır. Bu nedenle farklı madde türlerinin fen başarısının belirlenmesinde birlikte kullanılması önerilmektedir. Bu araştırmanın bulguları göz önüne alındığında, Türkiye'de, özellikle merkezi sınavlarda farklı soru türlerinin birlikte kullanılması yönünde atılan adımların olumlu sonuçlar verebileceği düşünülmektedir.

Bu araştırmada fen okuryazarlığı üzerinde durulmuştur. Matematik okuryazarlığı ve okuma becerileri için de benzer bir çalışma yapılabilir. PISA 2015 Türkiye örneklemi üzerinden yürütülen bu çalışmada, Türkiye örneklemindeki öğrencilerin fen okuryazarlığı düzeylerinin düşük olması, ranj daralmasına sebep olabilir. Hipotez testlerindeki anlamlı olmayan bulguların sebebi olarak bu durum gösterilebilir. Bu nedenle, fen okuryazarlığı düzeyi düşük, orta ve yüksek ülkelerden elde edilecek örneklemler üzerinden bu araştırma tekrar edilebilir. Bu araştırmada kullanılan madde türü sınıflandırması PISA 2015 uygulamasına aittir. Farklı madde türü sınıflandırmaları ile bu araştırma tekrar edilebilir.