The Effect of the Item's Type and Cognitive Level on Its Difficulty Index: The Sample of TIMSS 2015

Mustafa İlhan∗

Department of Mathematics and Science Education, Ziya Gokalp Education Faculty, Dicle University, Diyarbakır, Turkey, ORCID: 0000-0003-1804-002X

Nagihan Boztunç Öztürk

Lifelong Learning Center, Hacettepe University, Ankara, Turkey, ORCID: 0000-0002-2777-5311

Melek Gülşah Şahin

Department of Educational Sciences, Gazi Education Faculty, Gazi University, Ankara, Turkey, ORCID: 0000-0001-5139-9777

	In this research, the effect of an item's type and cognitive level on its difficulty index was investigated. The data source of the study consisted of the responses of the 12535 students in the Turkey sample (6079 and 6456 students from eighth and fourth grade respectively) of TIMSS 2015. The responses were a total of 215 items at the eighth-grade level (115 multiple-choice and 100 open-ended) and 178 items at the fourth grade level (93 multiple-choice and 85 open-ended) in the mathematics test. In the present study, the difficulty indices for the items of different types and cognitive level were calculated firstly. Then, the item type was defined as a dummy variable and multiple regression analysis was applied to test for the predictive effect of the type and cognitive level of the items on its difficulty index. In the study, it was determined that both the type and the cognitive level of the item had a statistically significant effect on its difficulty index. Compared to multiple choice items, it was detected that students had more difficulty in open-ended items. The effect of item type on difficulty index was found to be higher in eighth grade items than the fourth-grade ones. It was ascertained that difficulty index moved toward zero as its cognitive level increased at both grade levels. However, the effect of cognitive level on the difficulty index was higher in the fourth-grade items compared to the eighth grade items.

Introduction

International exams play an important role in confirming students' success and shaping the educational policies of countries. These exams provide the opportunity to make a

---

∗ Correspondence: mustafailhan21@gmail.com

comparison between the participant countries and they present important data regarding the change in a certain country's success over the years. Moreover, the fact that international exams consist of student and school questionnaires besides achievement tests turns these exams into a crucial data source for studies that aim to specify the variables which influence success.

One of the international exams in which Turkey participates is TIMSS (Trends in International Mathematics and Science Study), which is directed once in every four years by IEA (International Association for the Evaluation of Educational Achievement). TIMSS aims at assessing student achievement in mathematics and science and it focuses on the basic skills included in the curriculum (MoNE, 2015). In this respect, the results obtained through TIMSS provide precious feedback for the education policy makers, curriculum developers and teachers as to the functioning of the educational system. TIMSS is conducted at fourth and eighth grades, and this makes it possible to carry out quasi-longitudinal studies by means of using the relevant data (see https://www.iea.nl/studies/iea/timss).

One of the most important aspects of TIMSS is that it contains items pertaining to different cognitive levels and types in achievement tests. TIMSS include items at three different cognitive levels, which can be listed as knowing, applying and reasoning, all of which are parallel to the cognitive levels in Bloom's taxonomy (Delil, 2006). The first level includes concepts and principles that students should know. The second level focuses on students' ability to apply what they know while they are solving the problems that they encounter and while giving answers to the questions. The third level is about students' ability to create a solution to unfamiliar, stepwise and complex problems that go beyond routine issues. These cognitive levels listed above are common for both grades (Mullis & Martin, 2013). While the tests are being created for TIMSS, strict attention is paid to ensure they all include items at these three levels for each subject field.

There is a variety of items in TIMSS not only in terms of cognitive levels but also regarding format. TIMSS tests include multiple choice and open-ended questions. Multiple choice items require students to choose the correct answer among the given alternatives, whereas students construct their own answers for the open-ended questions. Both multiple choice items and open-ended questions have some advantages as well as constraints. The prominent advantages of multiple choice items are that they can be scored in an objective, easy and quick way (Haladyna, 2004), and content validity and reliability can be ensured easily as a relatively large number of items can be administered in a short time (Linn, 2006). The restrictions of multiple choice items can be enlisted as flows: they are not suitable for assessing high cognitive levels, it takes quite a lot of time to prepare them, examinees can choose the correct answer by guessing (Bennett, Ward, Rock & LaHart, 1990) and some factors that fall out of the relevant ability focused on such as reading speed and reading comprehension can interfere with the actual assessment results (Doğan, 2009). On the other hand, open-ended questions are advantageous in that they are suitable for assessing high- cognitive levels (Bahar, Nartgün, Durmuş & Bıçak, 2010, Tan, 2016), they cannot be answered correctly only through guessing (Turgut & Baykul, 2012) and they can be prepared in a short time (Başol, 2013). That being said the restrictions involve the fact that asking few open-ended questions decreases content validity (Özçelik, 2011), and scoring them is difficult; time-consuming at the same time relying on rater judgements (Tekin, 2009). One of the most effective methods to be employed in order to take the most advantage of multiple choice and open-ended questions and thusly minimize their restrictions is to make use of these two types of items together. In parallel with this, multiple choice and open-ended items are used together in TIMSS as is the case in other national and international examinations such as PISA (Program for International Student Assessment) and ABİDE [Monitoring and

Assessing Academic Skills (Akademik Becerilerin İzlenmesi ve Değerlendirilmesi)]. In TIMSS, multiple choice items are scored with 1 (0–1). On the other hand, open-ended questions are scored with 1 (0–1) or 2 (0–1–2) depending upon the content and the skill that is measured (see https://timssandpirls.bc.edu/timss2015/international-database/assessment-items.html)

Aim of the Study

This study aims at identifying the effect of the type and cognitive level of the item on its difficulty index by making use of the data gathered from TIMSS 2015. In line with this aim, first of all, the difficulty indices of the items of different cognitive level and type will be calculated. Then, the effect of the type and cognitive level of the item on the difficulty index was explored. The literature review shows that there are studies which analyse Turkey's success in international exams within the context of the type of the item (Demir, 2010, Köklü, 2017, Özer Özkan & Özaslan, 2018, Yılmaz Koğar & Koğar, 2019). This study is different from the previous studies in the literature in that it aims to test the predictive effect of the item's type on its difficulty index instead of depicting students' success in different types of items. Another aspect that differentiates this study from the others in the literature is that it intends to identify the effect of not only the type of the item but also the cognitive level of the item on the difficulty index. During the literature review, the researchers have not come across any study that analyses Turkey's success in international exams within the context of the cognitive level of the item. Moreover, this study is believed to be different from the other studies in the literature in that it attempts to identify whether the effect of the type and cognitive level of items on the difficulty index differs in TIMSS items at fourth and eighth grade level. When all these points are considered, it is thought that the results of the study will shed light to the ongoing discussions on multiple choice items and open-ended questions in high stake exams.

Method

Research Design

This study aims at depicting the facts without any intervention. Therefore, it can be accepted to be a descriptive research (Coe, 2017).

Source of Data and Arranging the Data

The data source of this study consists of the responses of the 12535 students in the Turkey sample (6079 and 6456 students from eighth and fourth grade respectively) of TIMSS 2015 to the items in the mathematics test. The data source was reached via the official international data base of TIMSS (https://timssandpirls.bc.edu/timss2015/international-database/). The parts of the data base that were specifically considered for the purpose of this study were "*Item Information Table*" and "*Item Percent Correct Statistics*" located under the *"Items"* section.

Item information tables include information regarding item numbers (ID), the topic/content that they are related to, their type and cognitive level. Item percent correct statistics, on the other hand, give information on the scores that the students get for each item. These pieces of information represent the proportion of students that give correct answers to multiple choice questions and open-ended questions that are scored as 0-1 to the total number of students that sit the exam. When it comes to open-ended questions that are scored as 0–1–2, they express the

ratio of students who give a fully correct answer and get 2, who give a partially correct answer and get 1, and who give an incorrect answer or does not provide any answer and get 0.

Item information tables and item percent correct statistics are shared as two separate sets of data on TIMSS official website. These two data sets were united with a view to realising the analysis within the scope of the study. To that end, the variables related to the type and cognitive level were taken from the data set of item information tables; the percentages of correct answers of students in Turkey sample were taken from the data set of items percent correct statistics. Afterwards, they were united in a single file on the basis of the item numbers. When the obtained data set was examined, it was seen that the percentage of correct answers given by the students belonging to five questions at eighth grade was not included in the item percent correct statistics, and hence these items were not included in the data set. Together with this, when the variable of alternatives regarding the multiple-choice items were examined, four items at eighth grade and seven items at fourth grade appeared to have two alternatives. As the proportion of these items to the total number of items was quite low, it was decided to take these items out of the data set. Consequently, 393 items in TIMSS 2015 mathematics test were included in the data analysis. The distribution of these items in accordance with their type and cognitive level are given in Table 1.

Table 1. Distribution of TIMSS 2015 mathematics test items in accordance with their type and cognitive level

Cognitive Level	Eighth Grade		Fourth Grade	
	Multiple choice	Open-ended	Multiple Choice	Open-ended
Knowing	53	21	42	32
Applying	47	48	36	36
Reasoning	15	31	15	17
Total	115	100	93	85

Data Analysis

The analysis process started with figuring out the difficulty indices of open-ended items by using Formula 1 (İlhan, 2019) below. When the multiple-choice items are in the question in TIMSS data set, the percent of correct answers represent the difficulty index of these items, so there was no need to calculate difficulty index once more for multiple choice items.

$$\text{p} = \frac{\left(\begin{array}{c}\textit{Mean of scores} \\ \textit{obtained from the item}\end{array}\right) - \left(\begin{array}{c}\textit{The lowest score to be} \\ \textit{obtained from the item}\end{array}\right)}{\left(\begin{array}{c}\textit{The highest score to be} \\ \textit{obtained from the item}\end{array}\right) - \left(\begin{array}{c}\textit{The lowest score to be} \\ \textit{obtained from the item}\end{array}\right)} \qquad \text{Formula 1}$$

After getting the difficulty indices regarding each item, means of difficulty indices concerning items of different type and cognitive level separately for eight and fourth grade, along with figures that include these means were created. After that, stepwise multiple regression analysis was applied in order to be able to test the predictive effect of the item's type and cognitive level on its difficulty index. As the item type is a variable at a nominal scale with two categories, it was transformed into dummy variable prior to the regression analysis. Right after the transformation process, it was checked if the assumptions of the regression analysis were met or not. The first assumption that was tested in the study was compliance of the data with the normal distribution. For that purpose, coefficients of skewness and kurtosis of eighth and fourth grade items were examined. The obtained coefficients are given in Table 2.

Table 2. Coefficients of skewness and kurtosis calculated for the data

Grade	Skewness		Kurtosis	
	Statistics	Standard Error	Statistics	Standard Error
Eight	.24	.17	−.28	.33
Four	.13	.18	−.69	.36

As is clear in Table 2, coefficients of skewness and kurtosis are all within the interval of ±1. According to Büyüköztürk, Çokluk and Köklü (2011), having coefficients of skewness within the interval of ±1 shows that the data do not present an important deviation from normal distribution. Having a *Z* value between the interval of ±2 after dividing coefficients of skewness and kurtosis by their standard error is another criterion that is considered acceptable for normal distribution in the literature (Tan, 2016). The values given in Table 2 meet this criterion. Therefore, it can be stated that the study data meet the normality distribution, which is the first assumption of regression analysis.

The second assumption of regression analysis is having a linear relationship between the dependent variable and independent variables. This assumption can be tested through correlation coefficients. When the correlation coefficients belonging to the relationship between the independent variables and dependent variable exceed the value of .30, it can be inferred that the related assumption is met (Pallant, 2005). In the current study, the dependent variable is the item difficulty index, whereas the independent variables are the type and cognitive level of the item. The correlation coefficients between the listed variables are presented in Table 3.

Table 3. Correlations between the type and cognitive level of the items and their difficulty index

	Eighth Grade			Fourth Grade		
	Difficulty index	Type of the item	Cognitive level of the item	Difficulty index	Type of the item	Cognitive level of the item
Difficulty index	1.00			1.00		
Type of the item	.50[*]	1.00		.31[*]	1.00	
Cognitive level of the item	−.32[*]	−.29[*]	1.00	−.35[*]	-.08	1.00

[*]*p* < .001

As can be witnessed in Table 3, the correlations between the type and cognitive level of the item, and difficulty index are statistically significant for both eighth and fourth grade. In light of this, it can be announced that the assumption regarding the regression analysis that there is a linear relationship between the dependent variable and independent variables is met. The third assumption of the multiple regression prescribes that there is no multicollinearity between the independent variables. Having a correlation value of over .80 between the independent variables shows that there might be a multicollinearity problem, while correlation coefficients that are equal to or over .90 show that there is a serious multicollinearity problem (Büyüköztürk, 2010). When the correlation coefficients given in Table 3 are considered in accordance with these criteria, it is clearly seen that the data set of this study does not have a multicollinearity problem. After confirming that the necessary assumptions are met, multiple regression analysis was carried out. In the study, independent samples *t*-test and one-way ANOVA were also performed to see the effect of the item's type and cognitive level on its difficulty index more

clearly. The data analysis was carried out through Microsoft Excel and SPSS 21.0 packaged software.

Results

This section starts with the figures that belong to the difficulty index in accordance with the type and cognitive level of items. There are two separate figures for TIMSS items at eighth and fourth grade level. Figure 1 given below shows the difficulty indices of the TIMSS items at the eighth-grade level.
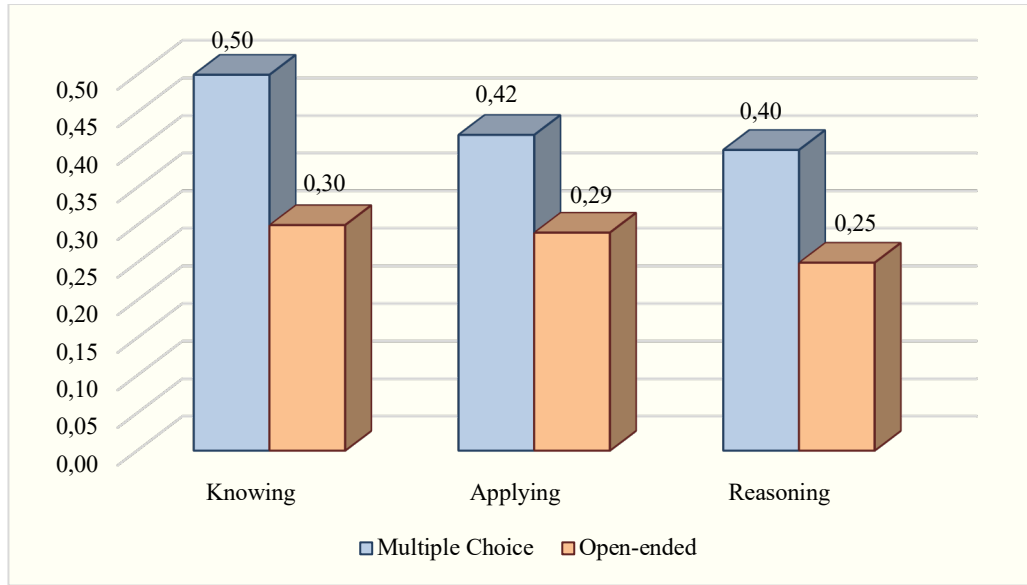

Figure 1. Difficulty indices of TIMSS items at eighth-grade level in accordance with the type and cognitive level of the items

According to Figure 1, difficulty indices of open-ended items are closer to zero when compared to multiple choice items at all cognitive levels. In other words, students seemed to have more difficulty in open-ended items when compared to the multiple-choice items. In fact, when the differences between difficulty indices of multiple choice and open-ended items were scrutinised through independent samples *t*-test, the result was found to be significant [$\bar{p}$ values for multiple choice and open ended items are .45 and .28 respectively, $t_{(213)}=8.50$, $p<.001$]. Over and above, the calculated difficulty indices show that an open-ended item at the knowing level is more difficult for students than a multiple-choice item at the reasoning level. Another point that draws attention in Figure 1 is that the higher cognitive level of an item is, the closer the difficulty index is to zero for both multiple-choice and open-ended items. In fact, when the differences between the difficulty indices of the items with different cognitive levels are tested by means of one-way ANOVA, the result is statistically significant [$\bar{p}$ values of the items at the level of knowing, applying and reasoning are .45, .35 and .30 respectively; $F_{(2-212)}=12.48$, $p<.001$]. On the other hand, Figure 2 given below shows the difficulty indices of the TIMSS items of different types and cognitive levels at the fourth-grade level.
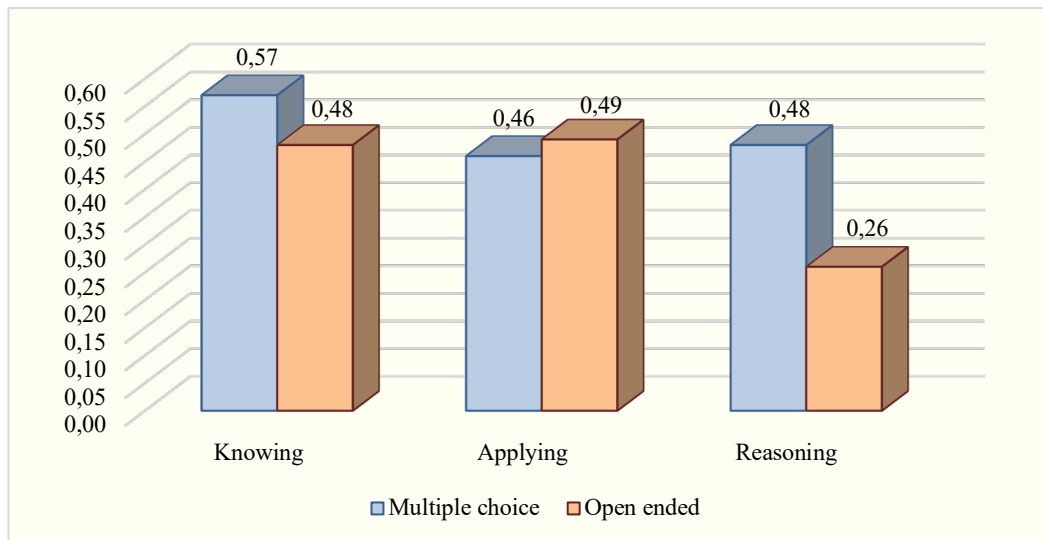
Figure 2. Difficulty indices of TIMSS items at fourth-grade level in accordance with the type and cognitive level of the items

According to Figure 2, the items whose difficulty indices are close to 1, which means that they are the easiest items for students, are multiple-choice questions at the knowing level, whereas the most difficult items are open-ended questions at the reasoning level. Furthermore, it is remarkable that the cognitive level at which the difference between difficulty indices of multiple-choice and open-ended items is the most is knowing, whereas the cognitive level at which the difference is found to be the least is applying. The difficulty indices of the items at the cognitive levels of applying and reasoning are close to each other in multiple choice questions, whereas the difficulty indices of items at the levels of knowing and applying are close to each other when it comes to the open-ended questions. As a matter of fact, the differences between the difficulty indices of the items in terms of both cognitive levels [$\bar{p}$ values of the items at the level of knowing, applying and reasoning are .53, .42 and .36 respectively; $F_{(2-175)}=12.79$, $p<.001$] and types [$\bar{p}$ values for multiple choice and open ended items are .51 and .40 respectively, $t_{(176)}=4.29$, $p<.001$] are statistically significant.

Figure 1 and Figure 2 point out that the item's type and cognitive level affect its difficulty index and that this effect changes in accordance with the grade levels. Having said that, the predictive effect of the type and cognitive level of an item on difficulty index was tested by carrying out a regression analysis to come to a definite conclusion about that. Regression analysis was done in line with the stepwise method, and the item type was defined as dummy variable by coding as "*multiple-choice:* 1, *open-ended*: 0." The results of the analysis for TIMSS items at the eighth and fourth grade levels are demonstrated in Table 4.

Table 4. Results of the regression analysis regarding the effect of the type and cognitive level of an item on its difficulty index

Grade	Step	Model	Unstandardized coefficients		Standardized coefficients	t	R	R^2
			β	Std error	Beta			
Eighth Grade	1	Constant	.28	.02		19.42[*]	.50	.25
		Type	.17	.02	.50	8.50[*]		
	2	Constant	.38	.03		11.54[*]	.54	.28
		Type	.15	.02	.45	7.38[*]		
		Cognitive Level	−.04	.01	−.19	−3.12[*]		
Fourth Grade	1	Constant	.62	.04		17.75[*]	.35	.12
		Cognitive Level	−.09	.02	−.35	−4.98		
	2	Constant	.55	.04		14.97[*]	.45	.19
		Cognitive Level	−.09	.02	−.33	−4.87[*]		
		Type	.11	.03	.28	4.16[*]		

For eighth grade F_{Type}=72.20[*], $F_{Type, Cognitive Level}$=42.45[*]
For fourth grade $F_{Cognitive Level}$ =24.83[*], $F_{Cognitive Level, Type}$=22.24[*]; [*]$p$<.01

As Table 4 indicates, the regression model that was created to predict the difficulty indices of TIMSS items is statistically significant for both eighth [F=42.45, p<.01] and fourth [F=22.24, p<.01] grade levels. At the first step of the regression analysis for eighth grade level TIMSS items, the variable of the item type took part and it could predict 25% of the variance regarding the item difficulty [R^2=.25]. At the second step of the analysis, the variance that could be predicted increased to 28% as the variable of the cognitive level was included in the model [R^2=.28]. In the present case, it is clear that the cognitive level of the item contributes to the model with a value of 3%. When Table 4 is examined, it is seen that β coefficients are positive for item type, whereas they are negative for cognitive level. It is necessary to interpret the results of the analysis in compliance with the multiple-choice questions as multiple choice items were coded as 1 and open-ended items were excluded after item type was transformed into dummy variable. In conformity with this, the fact that β coefficients regarding the multiple-choice items are positive stems from the fact that the difficulty indices of these items are closer to 1 when compared to the ones of open-ended items. The fact that coefficients of β regarding the variable of cognitive level are negative displays that the higher the cognitive level of an item is, the closer its difficulty index to zero, which means that the item gets more difficult for students.

The order of predictive variables included in the regression model for the fourth grade is different from the eighth grade. According to Table 4, the regression analysis respecting fourth grade TIMSS items firstly included the variable of cognitive level and it could predict 12% of the variance regarding the item difficulty index [R^2=.12]. In the second step, the variable of the item type as well as cognitive level were included in the regression model and the predicted variance increased to 19% [R^2=.19]. Thereupon, it can be articulated that the variable of item type contributes to the regression model with a value of 7%. What is more, β coefficients exhibit that the higher the cognitive level of an item gets, the closer its difficulty index gets to zero and that the difficulty indices are closer to zero in open-ended items when compared to multiple-choice items. That is, just like eighth grade students, fourth grade students had more difficulty in open-ended items alongside items that were at a high cognitive level.

Discussion and Conclusion

In this study, the effect of the type and cognitive level of an item on its difficulty index was analysed with the data set obtained from TIMSS 2015 mathematics test. The findings of the study clearly reveal that students are more successful in multiple-choice questions than they are at open-ended questions no matter what the cognitive level of the item is. As a result of the regression analysis, type of the item was found to be a significant predictor of the difficulty index at both eighth and fourth grade level. These findings are parallel to the results of the studies by Demir (2010) and Yılmaz Koğar and Koğar (2019). Demir (2010) conducted a study that compared the students' success in Turkey in different item types included in the cognitive field test in PISA 2003 and PISA 2006. At the end of this study, it was determined that the type of the item in which the students in Turkey had the greatest success in the sub-fields of reading skills and science literacy was multiple-choice questions. It was also clear that for the sub-fields of maths literacy and problem-solving skills, students' success in structured item types (multiple choice, complex multiple choice and semi-structured questions) was higher than it was in item types in which students were expected to construct the answer themselves (short-answer, open-ended). Similarly, Yılmaz Koğar and Koğar (2019) figured out the difficulty index of the items included in the science literacy cognitive test in PISA 2015 and detected that open-ended questions were more difficult than multiple-choice questions. Pepple, Young and Carroll (2009); Temizkan and Sallabaş (2011); Temel, Dinçol Özgür and Yılmaz (2012); Gayef, Oner and Telatar (2014); Delaram and Safiri (2014); Thawabieh (2016) as well as Öksüz and Güven Demir (2019) carried out studies in which they compared students' success in multiple-choice and open-ended questions not with the data obtained from national or international exams but with the tests conducted in class, and they got similar findings.

There are also some studies in the literature that have different results from the results of this study in terms of the effect of the item type on the difficulty index. Hastedt and Sibberns (2005) analysed the differences in success between open-ended and multiple-choice questions with the data set obtained from TIMSS 1995 and TIMSS 1999, and they found out that there was no statistically significant difference between students' success in different types of items included in maths test. The difference between the study results obtained by Hastedt and Sibberns (2005) and the findings of the present study can be explained by the fact that a country's education system influences students' success in parallel with their performance in different item types. Indeed, Hastedt and Sibberns (2005) pinpointed that the effect of item type on students' success differs from country to another and reported that student's academic achievement is lower in open-ended questions when compared to multiple-choice ones in Eastern European countries.

According to one of the remarkable findings of this study, the effect of the item type on the difficulty index was found to be 25% at eighth grade while it was 7% at fourth grade although the effect was seen to be statistically significant at both grade levels. This difference might be a reflection of high-stakes exams that are a crucial part of Turkish education system. The reason is that almost all of the high-stakes exams held in Turkey are composed of multiple-choice items, and open-ended questions are not included in these exams. Such exams inevitably impact the overall educational system (Çıkrıkçı Demirtaşlı, 2010). Especially high-stakes exams applied within the scope of transition to high school turn in-class activities, homework given to students and students' study habits into a test-focused nature as grade level proceeds from primary school to secondary school. Correspondingly, students at higher grade levels come across open-ended items less, and so questions in open-ended format turn out to be more difficult for students. Thence, the effect of the item's type on its difficulty index is more limited at lower grade levels, whereas it is more significant at higher grade levels. Consistent with this,

at the end of the study carried out by O'Leary (2001) to investigate countries' rankings in TIMSS in terms of item type, it was discovered that the item type that students mostly come across leads students to have different levels of success in items that are of different types. According to O'Leary's (2001) study, students who live in countries where multiple-choice items are frequently used are more successful in multiple-choice questions, whereas students who mostly come across open-ended items have a higher academic achievement in this type of questions. Eventually, item type turns out to be the most significant determinants of the item difficulty index at higher grade levels where students mostly encounter multiple-choice questions. In this respect, the steps taken by the Ministry of National Education and the Student Selection and Placement Centre to include open-ended questions as well as multiple-choice items in high stakes exams are important in terms of distancing student achievement from being a function of item type.

It was found out at the end of the study that the cognitive level of an item is a significant predictor of its difficulty index at both eighth and fourth grade level and that the difficulty index gets closer to zero as the cognitive level of the item increases. It was figured out that students had more difficulty in items with high cognitive level. This finding is in parallel with the study results of Kim, Patel, Uchizono and Beck (2012), Nevid and McClelland (2013), Veeravagu, Muthusamy, Marimuthu and Subrayan, (2010), and Koçdar, Karadağ and Şahin (2016), whereas it is in contradiction with the results of the studies carried out by Kibble and Johnson (2011), and Momsen et al. (2013). The fact that there is not a consistency between the findings of different studies that concentrated upon the effect of an item's cognitive level on its difficulty index can be attributed to differences in the content of the items that were included in these studies. To put it in a different way, an item which is at the remembering/knowing level can sometimes be more difficult than an item at the applying or a higher-level one because of its content. Momsen et al. (2013) exemplified this case with an item included in the study carried out by Nehm and Schonfeld (2008). In the study conducted by Nehm and Schonfeld (2008), it was expected that an item about natural selection could be easy for the participants as it was at the remembering level. Nevertheless, it was seen that the participants had difficulty with the item just mentioned. In this sense, the fact that different studies' coming to contradictious conclusions about the effect of an item's cognitive level on its difficulty index can be explained by the fact that qualities such as the content of the item, the subject field that it is related to, the way it is expressed, and others are also amongst the factors affecting the item difficulty index. This can also be the reason why the predictive effect of an item's cognitive level on its difficulty index is different for eighth and fourth grade TIMSS data set.

Recommendations

When it is considered that students have more difficulty in open-ended questions when compared to multiple-choice questions, it can be suggested that measurement and evaluation studies within the educational process should include more open-ended questions. When students come across open-ended items more, this may help them to experience less difficulty in dealing with this type of items. Considering that students have more difficulty in items at the cognitive level of reasoning, it should be noted that the process of learning and teaching should employ methods and techniques that will contribute to the development of high level mental processes and measurement and evaluation activities should also include items that focus on assessing high level cognitive skills. Lastly, as this study was carried out with the data set obtained from the items included in TIMSS 2015 maths test, a similar study can be carried out with the items included in science test or with the items included in ABİDE, which is accepted to be the national counterpart of TIMSS and PISA.

References

Bahar, M., Nartgün, Z., Durmuş, S., & Bıçak, B. (2010). *Geleneksel-tamamlayıcı ölçme ve değerlendirme teknikleri [Traditional-complementary measurement and evaluation techniques].* Ankara: Pegem.

Başol, G. (2013). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education].* Ankara: Pegem.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items.* Princeton, NJ: Educational Testing Service.

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı [Handbook of data analysis for social sciences].* Ankara: Pegem.

Büyüköztürk, B., Çokluk, Ö., & Köklü, N. (2011). *Sosyal bilimler için istatistik [Statistics for social sciences].* Ankara: Pegem Academy.

Coe, R. J. (2017). *The nature of educational research.* In R. Coe, M. Waring, L. V. Hedges, & J. Arthur (Eds.), Research methods & methodologies in education (pp. 5–14). Los Angeles: SAGE.

Çıkrıkçı Demirtaşlı, N. (2010). Açık uçlu soru formatı ve öğrenci izleme sistemi (ÖİS) akademik gelişimi izleme ve değerlendirme (AGİD) modülündeki kullanımı [Open-ended question format and its use in the module of student monitoring system (SMS) for monitoring and evaluating academic development (MEAD)]. *Cito Education Theory and Practice*, 8, 22–30. Retrieved from http://www.ogretmen.info/makale/33.pdf.

Delaram, M., & Safiri, A. (2014). A comparison of students' scores in multiple-choice and essay questions in mother and child health examination. *Future of Medical Education Journal,* 4(2), 15–18. https://doi.org/10.22038/fmej.2014.2598

Delil, H. (2006). *An analysis of geometry problems in 6 - 8 grades Turkish mathematics textbooks.* (Unpublished Master Thesis), Middle East Technical University, Institute of Social Sciences, Secondary Science and Mathematics Education, Mathematics Education, Ankara.

Demir, E. (2010). *Uluslararası öğrenci değerlendirme programı (PISA) bilişsel alan testlerinde yer alan soru tiplerine göre Türkiye'de öğrenci başarıları [The students' achievement in Turkey, according to the question types used in program for international student assessment (PISA) cognitive domain tests].* (Unpublished Master Thesis), Hacettepe University, Institute of Social Sciences, Educational Sciences, Measurement and Evaluation, Ankara.

Doğan, N. (2009). *Çoktan seçmeli testler [Multiple choice tests].* In H. Atılgan (Eds.), Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education] (pp. 223–268). Ankara: Anı.

Gayef, A., Oner, C., & Telatar, B. (2014). Is asking same question in different ways has any impact on student achievement? *Procedia - Social and Behavioral Sciences,* 152, 339–342. https://doi.org/10.1016/j.sbspro.2014.09.206

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the LEA TIMSS surveys. *Studies in Educational Evaluation*, 31, 145–161. https://doi.org/10.1016/j.stueduc.2005.05.007

İlhan, M. (2019). *Madde analizi ve madde ile test istatistikleri arasındaki ilişki [Item analysis and relationship between item and test statistics].* In B. Çetin (Eds.), Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education] (s. 269–298). Ankara: Anı.

Kibble, J. D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in Physiology Education,* 35(4), 396–401. https://doi.org/10.1152/advan.00062.2011

Kim, M. K., Patel, R. A., Uchizono, J. A., & Beck, L. (2012). Incorporation of Bloom's taxonomy into multiplechoice examination questions for a pharmacotherapeutics course. *American Journal of Pharmaceutical Education*, 76(6), 1–8. https://doi.org/10.5688/ajpe766114

Koçdar, S., Karadağ, N., & Şahin, M. D. (2016). Analysis of the difficulty and discrimination ındices of multiple-choice questions according to cognitive levels in an open and distance learning context. *The Turkish Online Journal of Educational Technology,* 15(4), 16–24. Retrieved from https://files.eric.ed.gov/fulltext/EJ1117619.pdf

Linn, R. L. (2006). *The standards for educational and psychological testing: Guidance in test development.* In S. M. Downing, & T. M. Haladyna (Eds.), Handbook of test development (pp. 27–38). Mahwah, NJ: Lawrence Erlbaum.

Milli Eğitim Bakanlığı [Ministry of National Education]. (2015). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. Sınıflar [National mathematics and science preliminary report of grade 4 and 8 for TIMSS 2015].* Retrieved from http://timss.meb.gov.tr/wp-content/uploads/TIMSS_2015_Ulusal_Rapor.pdf.

Mullis, I. V. S. & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Momsen, J., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE—Life Sciences Education*, 12(2), 239–249. https://doi.org/10.1187/cbe.12-08-0130

Nehm, R. H., & Schonfeld, M. (2008). Item feature effects in evolution assessment. *Journal of Research in Science Teaching,* 48(3), 237–256. https://doi.org/10.1002/tea.20400

Nevid, J. S., & McClelland, N. (2013). Using action verbs as learning outcomes: Applying Bloom's taxonomy in measuring instructional objectives in introductory psychology. *Journal of Education and Training Studies,* 1(2), 19–24. http://dx.doi.org/10.11114/jets.v1i2.94

O'Leary, M. (2001). Item format as a factor affecting the relative standing of countries in the third international mathematics and science study (TIMSS). *Irish Educational Studies*, 20(1), 153–177, https://doi.org/10.1080/0332331010200114

Öksüz, Y., & Güven Demir, E. (2019). Açık uçlu ve çoktan seçmeli başarı testlerinin psikometrik özellikleri ve öğrenci performansı açısından karşılaştırılması [Comparison of open ended questions and multiple choice tests in terms of psychometric features and student performance]. *Hacettepe University Journal of Education*, 34(1), 259–282. https://dx.doi.org/10.16986/HUJE.2018040550

Özer Özkan, Y., & Özaslan, N. (2018). Student achievement in Turkey, according to question types used in PISA 2003-2012 mathematic literacy tests. *International Journal of Evaluation and Research in Education (IJERE)*, 7(1), 57–64. Retrieved from https://files.eric.ed.gov/fulltext/EJ1174883.pdf

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows (Version 12)*. New York: Open University Press.

Pepple, D. J., Young, L. E., Carroll, R. G. (2009). A comparison of student performance in multiple-choice and long essay questions in the MBBS stage I physiology examination at the University of the West Indies (Mona Campus). *Advances in Physiology Education*, 34(2), 86–89. https://dx.doi.org/10.1152/advan.00087.2009

Tan, Ş. (2016). *Öğretimde ölçme ve değerlendirme KPSS el kitabı [Measurement and evaluation in teaching handbook for public personnel selection exam].* Ankara: Pegem.

Tan, Ş. (2016). *SPSS ve Excel uygulamalı temel istatistik–I [Basic statistics-I with SPSS and Excel]*. Ankara: Pegem.

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Ankara: Yargı.

Temel, S., Dinçol Özgür, S., & Yılmaz, A. (2012). The effect of different types of test on preservice chemistry teachers' achievement related to "chemical bondıng". *Problems of Education in the 21st century,* 41, 123–129. Retrieved from http://oaji.net/articles/2014/457-1409840433.pdf

Temizkan, M., & Sallabaş, M. E. (2011). Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması [Comparing multiple choice tests and open ended questions while evaluating reading comprehension skills]. *Dumlupınar University Journal of Social Sciences*, 30, 207–220. Retrieved from https://birimler.dpu.edu.tr/app/views/panel/ckfinder/userfiles/17/files/DERG_/30/207-220.pdf.

Thawabieh, A. M. (2016). Comparison between two test item formats: Multiple-choice items and completion items. *British Journal of Education*, 4(8), 63–74. Retrieved from http://www.eajournals.org/wp-content/uploads/A-Comparison-between-Two-Test-Item-Formats-Multiple-Choice-Items-and-Completion-Items.pdf

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Ankara: Pegem.

Veeravagu, J., Muthusamy, C., Marimuthu, R., & Subrayan, A. (2010). Using Bloom's taxonomy to gauge students' reading comprehension performance. *Canadian Social Science,* 6(3), 205–212. https://dx.doi.org/10.3968/j.css.1923669720100603.023

Yılmaz Koğar, E., & Koğar, H. (2019). Investigation of scientific literacy according to different item types: PISA 2015 Turkey sample. *Bolu Abant İzzet Baysal University Journal of Faculty of Education*, 19(2), 695–709. https://doi.org/10.17240/aibuefd.2019.19.46660-467271