TURJE
www.turje.org

Volume: 6 Issue: 4

DOI: 10.19128/turje.333687

Turkish Journal of Education

Research Article

Received	*09.08.2017*
Accepted	*10.10.2017*

Three, four, and none of the above options in multiple-choice items

Erkan H. Atalmış
Kahramanmaraş Sütçüimam University, Faculty of Education, Kahramanmaraş, Turkey, eatalmis@ksu.edu.tr
ORCID: orcid.org/0000-0001-9610-491X
Neal M. Kingston
University of Kansas, Achievement and Assessment Institute, Lawrence, KS 66047, USA, nkingsto@ku.edu
ORCID: orcid.org/0000-0003-1870-309X

ABSTRACT High-quality multiple-choice (MC) items are essential for creating efficient, valid assessments. Haladyna, Downing, and Rodriguez (2002) suggested that using plausible distractors is crucial to achieving this goal, although distractor creation can be time-consuming and challenging. Haladyna et al. (2002) provided two related test development guidelines: #18, “Write as many plausible distractors as you can,” and #25, “Use carefully None of the above.” This research aims to test the impact of these two guidelines on item difficulty (p), item discrimination (r), and test reliability for mathematics items empirically. The research findings have revealed that item discrimination and test reliability were not statistically different across MC items with four options, three options, and NOTA options while the means of item difficulty of four-option MC items was not statistically different from those of three-option and NOTA-option MC items. However, the mean of item difficulty of NOTA-option MC items was statistically lower than those of three-option.

Keywords *assessment, item-writing guidelines, multiple-choice items, number of options, none of the above,*

Çoktan seçmeli sorularda üç, dört ve hiçbiri seçenekleri

ÖZ Etkili ve geçerli değerlendirmeler yapabilmek için yüksek kaliteli çoktan seçmeli sorular (maddeler) oluşturmak gereklidir. Bu amaç doğrultusunda Haladyna, Downing, and Rodriguez (2002) çoktan seçmeli madde oluşturma sürecinde uygun çeldiricilerin kullanılmasının önemli olduğunu vurgulamaktadırlar. Ancak bu çeldirilerin oluşturulması zaman alıcı ve zahmetli bir süreçtir. Bu bağlamda Haladyna ve diğerleri (2002) test geliştirme ile ilişkili olan "Olabildiği kadar uygun ve mantıklı çeldiricileri yazınız" ve "Hiçbiri seçeneğini dikkatli kullanınız" ilkelerini rapor etmektedirler. Bu araştırmanın amacı bu iki ilke kullanımının matematik sorularınn zorluk derecesine, ayırt edicilik derecesine ve test güvenirliğine etkisini test etmektir. Araştırma sonucunda, madde ayırt edicilik derecesi ve test güvenirliği dört seçenekli, üç seçenekli ve “hiçbiri” seçeneğine sahip maddeler arasında istatistiksel olarak anlamlı bir farklılık göstermediği bulunmuştur. Benzer şekilde, dört seçenekli maddelerin madde zorluk dereceleri, üç seçenekli ve “hiçbiri” seçeneğine sahip maddelerin madde zorluk derecelerinden istatistiksel olarak anlamlı bir farklılık göstermediği bulunurken, “hiçbiri” seçeneğine sahip maddelerin madde zorluk dereceleri üç seçenekli maddelerin madde zorluk derecelerinden istatistiksel olarak düşük olduğu elde edilmiştir.

Anahtar Kelimeler *Değerlendirme, soru-yazma teknikleri, çoktan seçmeli sorular, seçenek sayısı, yukarıdakilerden hiçbiri,*

Cite This Article: Atalmış, EH & Kingston NM, (2017). Three, four, and none of the above options in multiple-choice items, *Turkish Journal of Education. 6*(4), 143-157. DOI:10.19128/turje.333687

INTRODUCTION

Multiple-choice (MC) items are commonly used in professionally developed assessments for various disciplines (Haladyna, Downing, & Rodriguez 2002; McCoubrie, 2004) due to their ability to obtain accurate and objective scores and efficient administeration and scoring tests. Creating plausible distractors (guideline 29 – "make all distractors plausible") is a crucial part of writing a well-constructed MC item (Haladyna et al., 2002) because distractors play an important role in increasing quality in terms of psychometric properties of the items and the test. However, writing plausible distractors is a difficult part of the item-writing process (Haladyna & Downing, 1989; Hansen & Dexter, 1997; Rich & Johanson, 1990) as plausible distractors should be written to reflect students' common errors (Haladyna & Downing, 1993; Haladyna et al., 2002); thus, creating plausible distractors can require a significant pedagogical background for item writers and considerable time for crafting each MC item. This increases the cost of item construction (Haladyna & Downing, 1993; Haladyna & Rodriguez, 2013) and decreases the frequency of using MC items (Burton, Sudweeks, Merrill, & Wood, 1991; Hansen& Dexter, 1997) compared to open-ended items.

Previous studies discussed the qualiy of distractors and revealed that more than 50% of these items have non-functioning distractors (Haladyna & Downing, 1993; Tarrant, Ware, & Mohammed, 2009). However, some alternative approaches can be used to make writing effective answer options easier and more straightforward. For instance, including fewer options or using "None of the Above" (NOTA) as a last option are strategies that could be employed to construct MC items. Haladyna et al. (2002) discussed these alternative methods in their guidelines, such as #18, "Write as many plausible distractors as you can," and #25, "Use carefully *None of the above*" (p. 341). Table 1 illustrates an MC item in three formats: the item with four options, with three options, and with an NOTA option. Each example item has the same content.

Table 1

Three formats of a multiple choice item

MC item with four options	MC item with three options	MC item with the NOTA option
What is $\frac{1}{3}+\frac{3}{4}$?	What is $\frac{1}{3}+\frac{3}{4}$?	What is $\frac{1}{3}+\frac{3}{4}$?
A. $\frac{13}{12}$ *	A. $\frac{13}{12}$ *	A. $\frac{13}{12}$ *
B. $\frac{4}{7}$	B. $\frac{4}{7}$	B. $\frac{4}{7}$
C. $\frac{13}{24}$	C. $\frac{13}{24}$	C. $\frac{13}{24}$
D. $\frac{13}{34}$		D. None of the Above

*Key for all three versions

Literature Review

Item-writing guidelines for valid item and test construction were addressed by a limited number of studies. In 1989, Haladyna and Downing suggested 43 item-writing guidelines based on measurement and evaluation textbooks. In 2002, Haladyna et al. redesigned the existing version identifying 31 valid item-writing guidelines mainly for classroom assessment, and classified them into five categories: content, formatting, style, forming the stem, and forming the choices.

The guidelines shown in Table 1 are not commonly used in large-scale and classroom test creation as few empirical researches are available to support them. Frey, Petersen, Edwards, Pedrotti, and Peyton (2005) analyzed the 20 best-known assessment textbooks. A total of 30% of the books presented an item writing guideline related to the number of options (Haladyna et al. 2002, guideline 18) while 75% presented a guideline related to NOTA items (Haladyna et al. 2002, guideline 25). Other texts stated that guidelines 18 and 25 are controversial, that inconsistent results have been found, and the empirical studies are limited.

None of the Above as an Answer Option

The number of empirical studies from the past 25 years about the NOTA option is limited, with only seven studies supporting the NOTA option with different ways. Various studies have investigated the impact of the NOTA option on item characteristics (item difficulty and item discrimination), while others have focused on test characteristics (test reliability). In other words, researchers tested item and test characteristics to determine how appropriate NOTA options are in MC item creation. Therefore, we will discuss the acceptability of NOTA items in large-scale standardized testing and classroom evaluation.

Replacement method

Replacement method is a method for embedding NOTA as an option in a conventional MC item, which consisted of a key and independent distractors (not including "None of the Above" or "All of the Above" option). This method is applied by choosing one answer option to replace with the NOTA option. For example, in the third item in Table 1, the NOTA option replaces the last option of the first item. In this example, the last option of the first item is randomly chosen to be replaced with the NOTA option. The other answer options could also be chosen and substituted with the NOTA option through different methods. Hence, various replacement methods are plausible for use in identifying a poorly functioning distractor to allow inserting the NOTA option.

Although studies over the past 25 years have explored different methods for embedding NOTA as an answer option in an MC item, some of the empirical studies explicitly provided the replacement methods that they used. To illustrate, the NOTA option was substituted for the most frequently chosen distractor (Tollefson, 1987), the least frequently chosen distractor (Crehan et al., 1993), and a randomly selected distractor (Frary, 1991).One study suggested that NOTA should be added as an alternative option (Odegard & Koen, 2007).

Item difficulty

Item difficulty is defined as the proportion of students who choose the correct answer. Most of the empirical studies over the past 25 years have investigated the impact of the NOTA option on item difficulty. Knowles and Welch (1992) performed a meta-analysis and found that the NOTA option did not significantly change item difficulty despite the fact that using NOTA options decreased test scores by 1.00 point compared to conventional item type. However, five studies indicated that using NOTA option made a statistically significant increase in difficulty compared to conventional MC items (Crehan et al., 1993; Frary, 1991; Kolstad & Kolstad, 1991; Rich & Johanson, 1990; Tollefson, 1987). The recent study by DiBattista and his colleagues (2014) found no difference between conventional MC items and NOTA items when using NOTA as a distractor; however, they found that items in which NOTA is used as a key were statistically more difficult than conventional items. To sum up, previous studies presented that the NOTA-option did not makes items easier. However, more research is needed to accurately confirm that there is a significant impact of the NOTA-option on item difficulty.

Item discrimination

Item discrimination is defined as how well the item differentiates students with high ability in the construct of interest from students with low ability. Several studies over the past 25 years have addressed the item discrimination of NOTA items. Some found no statistical difference between NOTA items and conventional MC items (Crehan & Haladyna, 1991; DiBattista et al., 2014; Frary, 1991; Knowles & Welch, 1992; Tollefson, 1987). However, Rich and Johanson (1990) found that NOTA items were more discriminating than conventional MC items. Therefore, previous studies showed that NOTA option did not harm item discrimination.

Test reliability (Internal consistency)

Test reliability for NOTA items was reported in only two out of the seven aforementioned empirical studies. The studies showed that reliability of tests composed of NOTA items and conventional MC items did not significantly differ (Kolstad & Kolstad, 1991; Rich & Johanson, 1990). Further studies are required to identify whether NOTA items have a significant impact on test reliability.

Items with Three Answer Options

The set of empirical studies conducted over the last 25 years have had controversial results regarding the impact of the number of answer options on item characteristics (item difficulty and item discrimination) and on test characteristics (test reliability and test validity).

Some studies reported the elimination methods the authors used to reduce the number of answer options in MC items. Different types of elimination methods were applied in these studies in order to reduce options. One common method was to delete the least-selected option (Abad, Olea, & Ponsoda, 2001; Dehnad, Nasser, & Hosseini, 2014; Delgado & Prieto, 1998; Landrum, Cashin, & Theis, 1993; Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006; Tarrant & Ware, 2010). Another study calculated the point-biserial correlation coefficient for each single option of an item and eliminated the option with the least discrimination (Trevisan, Sax, & Michael, 1991). Additionally, a recent study deleted one answer option at random to construct the item with fewer options (Baghaei & Amrahi, 2011).

Item difficulty

Limited number of empirical studies have compared item difficulty for MC items with four options and MC items with three options. Five of the eight studies found that item difficulty was not statistically different between four-option items and three-option items. However, the researchers applied a mix of elimination methods to construct three-option MC items from four-option items. Of the eight studies, six eliminated the least-popular option(Abad et al., 2001; Cizek &O'Day, 1994; Dehnad et al., 2014; Delgado & Prieto, 1998; Shizuka et al., 2006; Tarrant & Ware, 2010), one study removed the option with the lowest item discrimination (Trevisan et al., 1991),and one study randomly deleted an option (Baghei & Amrahi, 2011).

Some studies concluded that MC items with three options were statistically more difficult than MC items with four options, which is counterintuitive (Landrum et al., 1993; Rogers & Harley, 1999). Rodriguez (2005) conducted a meta-analysis and examined 48 empirical studies from 1925 to 1999 in order to uncover the effect of the number of options upon psychometric characteristics of MC items. Of these 48 studies related to achievement and aptitude tests, 27studies included pertinent results. The results supported that three-option items were slightly easier than four-option items.

In brief, many studies have shown that item difficulty did not significantly vary when the number of options in MC items in a form decreased. However, more research is essential to obtain generalized results regarding the impact of the number of options on MC item difficulty.

Item discrimination

Numerous empirical studies have investigated item discrimination for MC items with four options and with three options. Similar to the findings on item difficulty, these researchers found mixed results for item discrimination. Item discrimination between MC items with four options and items with three options was not statistically different in seven studies (Cizek & O'Day, 1994; Crehan et al., 1993; Dehnad et al., 2014; Delgado & Prieto, 1998; Rogers & Harley, 1999; Shizuka et al.,2006; Tarrant & Ware, 2010). Yet, four studies provided statistically significant evidence that item discrimination for MC items with three options was higher than item discrimination for MC items with four options (Baghei &Amrahi, 2011; Landrum et al., 1993; Rodriguez, 2005; Trevisan et al., 1991). Consequently, the literature reveals that three-option items do not harm item discrimination. However, more empirical studies are necessary to confirm the impact of the number of options in an MC item on item discrimination.

Test reliability

Six studies conducted over the past 25 years have analyzed the impact of three options versus four options on test reliability, still the researchers found mixed results. Three of them indicated that the number of options did not have a statistically significant impact on test reliability (Baghei & Amrahi, 2011; Delgado & Prieto, 1998; Rogers & Harley, 1999). Two studies found that test reliability increased when the forms with three options were employed (Rodriquez, 2005; Tarrant & Ware, 2010). Furthermore, one study investigated test reliability for low-, average-, and high-ability students (Trevisan et al., 1991). This study found that reliability coefficients decreased when the number of options decreased from four to three for low-ability students, while there was no statistical difference for average-ability students and high-ability students.

In summary, the literature indicates that constructing MC items with three answer options generally does not harm the psychometric characteristics of the items orthe test.

Significance of the Study and Research Questions

Previous studies examined the impact of number of options and NOTA options on item and test psychometric properties for different disciplines, such as Medicine/Nursing (Cizek & O'Day, 1994; Dehnad et al., 2014; Kolstad & Kolstad, 1991; Tarrant & Ware, 2010), Psychology (Crehan & Haladyna, 1991; Crehan et al., 1993; Landrum et al., 1993), Verbal/Vocabulary (Abad et al., 2001; Baghei & Amrahi, 2011; Delgado & Prieto, 1998; Lee & Winke, 2013; Shizuka et al., 2006; Trevisan et al., 1991), general knowledge (DiBattista et al., 2014) and statistics with qualitative problems (Tollefson, 1987). Different from previous studies, this study investigates how the number of options and NOTA options have an influence on mathematics item and test psychometric properties as constructing plausible distractors for MC mathematics items related to students' common errors are challenging and time-consuming for item writers and mathematics teachers. Thus, constructing MC mathematics with NOTA options and fewer number of options can take less time for item writers and mathematics teachers.

In addition, this study contributes to the literature by simultaneously applying NOTA options and number of options. Although the study by Crehan, Haladyna, and Brewer (1993) examined the impact of NOTA options and number of options on item characteristics (item difficulty and item discrimination), we extend this line of research through examining the impact of the two guidelines on both item characteristics and test reliability by addressing the following research questions:

1. Do item characteristics (item difficulty and item discrimination) vary across different types of mathematics items (four-option MC items, three-option MC items, and MC items with an NOTA option)?
2. Does test reliability vary across different types of mathematics items (four-option MC items, three-option MC items, and MC items with an NOTA option)?

METHODOLOGY

This section covers participants, instrument development, final instrument, data collection, and data analysis.

Participants

The pilot study was carried out with seventh and eighth grade students from the United States during the preparation process of creating the final instrument. The convenience sampling method was applied for both implementations. The pilot implementation was held during the spring semester of 2012 and 100 seventh grade students participated from one school in the United States. The final test was applied to 585 seventh and eighth grade students participated from five schools in the United States during the spring semester of 2013.

Instrument Development

One of the significant aspects of test development is to determine the appropriate content for the construct. This research has chosen four seventh grade math standards from the content area of the mathematics, expressions and equations, based on the Common Core State Standards Initiative (2010), and 29 item pairs (identical specification for the two items within each pair) were written based on content area. The item pairs were divided into two forms (Form A and Form B), each with 29 items. Thus, the participants would be able to answer all items in one form in a single class period.

Students' responses to each item in the pilot forms were used to calculate item difficulty and discrimination parameters. Of the 58 pilot items, 27 four-option MC items with high discrimination and medium difficulty were selected for use in the final instrument. Then, we generated three-choice MC versions of 9 of the original 27 items and created NOTA versions of another 9 of those items. The last version consisted of a 27 item test with 9 four-choice, 9 NOTA items, and 9 three-choice, with the content within each item type set parallel to the others. It means that parallel items were constructed based on the same specific learning mathematics standard from Common Core State Standards (CCSS) and so they measured the same educational objectives students should possess at critical point of mathematics. The items in Table 2 illustrates parallel items used in the final instrument, measuring the standard of CCSS, "*7.EE.1.*Apply properties of operations as strategies to add, subtract, factor, and expand linear expressions with rational coefficients." They were constructed with the same content and rationale of distractors, but the numbers are different.

Table 2.
Examples of parallel items.

MC item with four options	MC item with three options	MC item with the NOTA option
Q1. Which expression could be used to find 3 times 5 more than *x*?	Q15. Which expression could be used to find 8 times 10 more than *x*?	Q24. Which expression could be used to find 12 times 16 more than *x*?
A. 3*x* + 5	A. 8*x* + 10	A. 12*x* + 16
B. 5*x* + 3	B. 10*x* + 8	B. 16*x* + 12
C. 3(*x* + 5)*	C. 8(*x* + 10)*	C. 12(*x* + 16)*
D. (5 + 3)*x*	D. None of the above	

*Key for all three versions

Option replacement and elimination

The research used mixed methods, which means that the most, the second, or the least selected distractor of each item was randomly chosen and deleted to construct three option and NOTA mathematics items when the same method was used for each triplet of parallel items (four option, three options, NOTA items) to eliminate option as shown in Table 3. For example, while the least-selected distractor was selected and deleted for the parallel items in set #1, the distractor of parallel items in set #3 was selected and deleted via the most second selected method.

Table 3.
Elimination and Replacement Methods used for each set of parallel items

Sets	MC item with four options	MC item with three options	MC item with the NOTA option
#1	-	The least-selected	The least-selected
#2	-	The least-selected	The least-selected
#3	-	The second most selected	The second most selected
#4	-	The most selected	The most selected
#5	-	The least-selected	The least-selected
#6	-	The second most selected	The second most selected
#7	-	The most selected	The most selected
#8	-	The most selected	The most selected
#9	-	The least-selected	The least-selected

Final Instrument

Taking into account how many multiple-choice mathematics items students can complete during class period, approximately 40–50 minutes in length, 27 multiple-choice items were selected to use in the final administration. Parallel MC mathematics items were written in three different formats: items with four options, items with the NOTA option, and items with three options, respectively. Parallel items were selected across the three formats. The final test was gathered with four-option items in positions 1–9, NOTA items in positions 10–18, and three-option items in positions 19–27. The final instrument was then administered to students. Within each section the parallel items were administered in a randomly different position. For instance, one triplet of parallel items were located in positions #1, #15, and #24 (see Table 2).

Data Collection

The forms in the pilot study were only applied to seventh grade students because the content area of the forms, "Expressions & Equations" is a seventh grade domainin CCSS (CCSSI, 2010). However, seventh grade math teachers in most of the schools did not completely cover the content of the form during the pilot administration. Therefore, a paper-pencil form was applied to seventh and eighth grade students at the beginning of the 2013 spring semester in the United States. The research has used convenience sampling method in order to select the schools in the United States. This research holds a total of 585 students from five different schools located in three different states. Students individually responded to each item by selecting an answer option on the paper form during one class period, approximately 40–50 minutes in length.

Data Analysis

A repeated measures ANOVA was used to determine whether item characteristics vary across different types of items since each student answered all items in terms of the first research question.. As for the second research question, coefficient alpha was calculated for each set of nine items. Standard errors of coefficient alpha were calculated through a method developed by Duhachek and Iacobucci (2004), and .95 confidence intervals have been taken into account to determine whether the differences in test reliability were statistically significant (four-option MC items, three-option MC items, and MC items with an NOTA option).

RESULTS

This section consisted of two parts, one of which provides the psychometric characteristics of each item, while the second part presents the results of the analyses of variance testing overall mean differences in item difficulty and discrimination across the three types of multiple-choice items. Two software packages, STATA (StataCorp, 2013) and SPSS (IBM, 2012), were used so as to perform the analyses.

Item Statistics

Table 4 displays classical item statistics (item difficulty and item discrimination) of 27 items from three groups. Each row shows the characteristics of the three items with parallel content in the different item formats.

Table 4.
Item difficulty (p) and item discrimination (r) by parallel item groups

	Item Difficulty (*p*)			Item Discrimination (*r*)		
Item Group	Four Options	NOTA Options	Three Option	Four Options	NOTA Options	Three option
1	0.64	0.58	0.65	0.37	0.54	0.50
2	0.30	0.29*	0.49	0.55	0.47*	0.43
3	0.55	0.51	0.55	0.56	0.56	0.52
4	0.67	0.43*	0.60	0.35	0.50*	0.44
5	0.41	0.40	0.45	0.50	0.54	0.39
6	0.30	0.24	0.44	0.43	0.16	0.34
7	0.59	0.59	0.60	0.51	0.51	0.52
8	0.40	0.39*	0.45	0.38	0.08*	0.41
9	0.32	0.30	0.43	0.40	0.38	0.35

* NOTA as key

The first column for each characteristics in Table 4 was based on 9 conventional four-option items. Item difficulty was calculated as the proportion of examinees answering the item correctly while item discrimination was calculated as item-total correlation index in this study, which is one of most widely used method (Downing, 2005). The item difficulty index (*p*) of the four-option items ranged from .30 (item groups2 and 4) to .67 (item group 4). The item discrimination index (*r*) of the items with four options ranged between .35 (item group 4) and .56 (item group 3).

The second column for each characteristic holds 9 NOTA items in which NOTA replaced the randomly chosen option of the conventional item (this process was reported in Instrument Development). The item difficulty index ranged from .24 (item group 6) to .59 (item group 7). The item discrimination index of the items with NOTA options was determined to be between .08 (item group 8) and 0.56 (item group 3).

The last column for each characteristic displays the indexes of the three-option items, in which the weakest distractor of the conventional items was eliminated (this option elimination process was described in Instrument Development). The item difficulty index of three-option items ranged from .43 (item group 9) to .65 (item group 1). The item discrimination index of each item with three options was acceptable ranging from .34 (item group 6) to .51 (item groups 3 and 7).

The repeated measures ANOVA was carried out for the other 9 groups with 27 items to determine whether item characteristics vary across different types of items. Figure 1 presents the distribution of students' total score in four MC items, three MC items, and NOTA MC items and their Q-Q plots. The plots have identified that students' scores in all type of items are normally distributed.

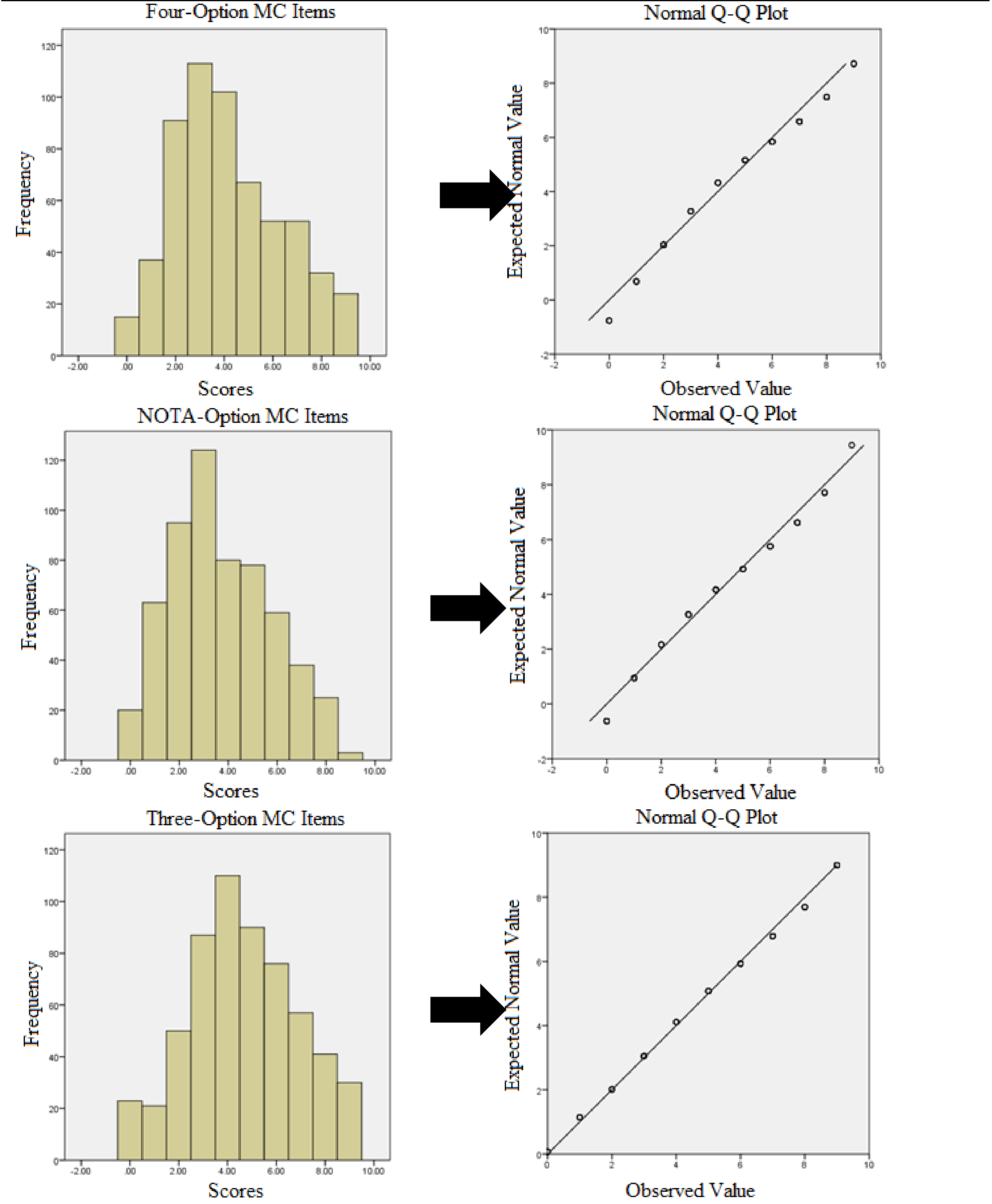

Figure 1. *Distribution of students' scores in different type of items*

Mean Differences in Item Difficulty and Discrimination

Item difficulty

A repeated measures ANOVA was conducted to analyze as to whether item characteristics vary across different types of items. A significant item type effect has been determined in terms of item difficulty, F(2,7) = 7.97, p< .05. The research has also conducted post-hoc comparison with the use of Bonferroni test with a view to evaluating pairwise comparison among the means of item difficulty for types of items. The means of item difficulty of four-option MC items (M_{four}= 0.46, SD_{four}=0.15) were not statistically different from those of NOTA-option (M_{NOTA}=0.41, SD_{NOTA}=0.13) and three-option MC items (M_{three}= 0.52, SD_{three}= 0.8). However, the means of item difficulty of NOTA-option MC items were statistically lower than those of three-option.

A repeated measures ANOVA has been also conducted to test how item difficulty change when NOTA as the distractors are used (6 sets of items out of 9 sets: 6 four-option MC items, 6 three-option MC items, and 6 items with NOTA as distractor). No significant item type effect has been identified in terms of item difficulty, F(2,5) = 4.53, p=.09. Besides, post-hoc comparison with the use of Bonferroni test was conducted in order to evaluate pairwise comparison among the means of item difficulty for types of items. The means of item difficulty of four-option MC items (M_{four}= 0.47, SD_{four}=0.14) were not found to be statistically different from those of NOTA-option as distractor (M_{NOTA_d}=0.44, SD_{NOTA_d}=0.15) and three-option MC items (M_{three}= 0.52 SD_{three}= 0.9). Similarly, the means of item difficulty of NOTA-option MC items were not statistically significant from those of three-option.

Item discrimination

ANOVA results for item discrimination have also indicated that item type effect is free from any significance, F(2,7) = .22, p=.80. The mean of item discrimination of four-option MC items (M_{four}= 0.45, SD_{four}=0.08) was not found to be statistically different from those of NOTA-option (M_{NOTA}=0.42, SD_{NOTA}=0.18), and three-option MC items (M_{three}= 0.43, SD_{three}= 0.07). A similar finding has been found related to the means of item discrimination of NOTA-option MC items compared to those of three-option.

A repeated measures ANOVA was also conducted to test how item discrimination change when NOTA as the distractors are used. The research findings have revealed that item discrimination indicated a not significant item type effect, F(2,5) = .24, p= .80. The mean difference in item discrimination of the three test forms with four-option MC items (M_{four}= 0.46, SD_{four}=0.07), MC items with NOTA-option (M_{NOTA}=0.45, SD_{NOTA}=0.16), and three-option MC items (M_{three}= 0.44, SD_{three}= 0.09) was not statistically different.

Differences in Reliability

The Cronbach's Alpha reliability of 27 multiple-choice mathematics used in the final implementation was found to be 0.84. For the second research question, .95 confidence intervals have been determined for each group of items. Cronbach's Alpha values are 0.65, 0.58, and 0.63 for MC items with four options, items with an NOTA option, and items with three options, respectively. Their *SE* values were ±0.02, ±0.03, and ±0.02, respectively. Thus, differences in the observed reliability coefficients of the three test forms (one form of MC items with four options, one form of MC items with three options, and one form of MC items with NOTA options) were not statistically different as their.95 confidence intervals overlapped:.61 and .69 for MC items with four options,.52 and .64 for MC items with NOTA options, and .59 and .67 for MC items with three options.

Test reliability was also calculated for only items with NOTA as the distractors in the current study. It means that The Cronbach's Alpha was calculated for18 items with 6 four-option MC items, 6 three-option MC items and 6 MC items with NOTA as distractors (see Table 4). Cronbach's alpha reliability coefficient was found to be 0.80 for the overall test. Cronbach Alpha values for four-option MC items, three-option MC items, and MC items with NOTA option are .62, .55, and.52, while *SE* values of three groups were ±0.03, ±0.03, and ±0.03, respectively. This shows the differences in the observed reliability coefficients of the three test forms were not statistically different as their .95 confidence intervals overlapped: .57 and .67 for MC items with four options, .49 and .60 for MC items with NOTA options, and .46 and .59 for MC items with three options.

DISCUSSION and CONCLUSION

MC items are effective for high-stakes tests and classroom assessment due to the small amount of time needed for the implementation, scoring, and analysis. However, writing plausible distractors in order to construct quality MC items is a challenging part of the item-writing process. Fortunately, there are some alternative ways to decrease effort in constructing items. Haladyna et al. (2002) addressed two writing guidelines related to this issue: "Write as many plausible distractors as you can" and "Use carefully *None of the above*" (p. 341). This study has empirically evaluated the impact of such guidelines (specifically, the use of NOTA option and three answer options) on the psychometric properties of MC items and tests.

The research findings have revealed that item discrimination and test reliability did not statistically differ in terms of MC items with four options, three options, and NOTA options; furthermore, the means of item difficulty of four-option MC items were not statistically different from those of NOTA-option and three-option MC items. However, the means of item difficulty of NOTA-option MC items were statistically lower compared to those of three-option.

The findings of the current study are consistent with previous studies in terms of item discrimination and test reliability. Accordingly, there is no statistically significant difference between four-option and three-option MC items for item discrimination and test reliability. (Another finding has indicated that there is no statistically significant difference between four-option and three-option MC items related to item difficulty even if using three-option items increases the value of item difficulty by .06 (making the items easier). In parallel to these studies, Delgado and Prieto (1998) reported that reducing 4 options to 3 increased the value of item difficulty between .03 and .07 after three different tests was applied to students, yet the change was not statistically significant. Rodriguez (2005) also found consistent and similar change in item difficulty (.04) between four-option MC items and three-option MC items. However, this small change was found to be statistically significant compared to the current study and that of Delgado and Prieto (1998). This may derive from the use of a different method, which is a meta-analysis of 26 studies.Rodriguez (2005) used meta-analysis of 26 studies for data analysis, which is different method from the ones that are used in the current study and in Delgado and Prieto's study (1998).

For the items with NOTA as distractor, no significant difference has been identified between conventional MC items and NOTA items when using NOTA as a distractor for item difficulty. Similar results emerged in studies conducted by DiBattista et al. (2014) and Tollefson (1987). The research results have also approved the consistency with the existing literature in terms of item discrimination (Crehan & Haladyna, 1991; Crehan, Haladyna, & Brewer, 1993; DiBattista et al., 2014; Frary, 1991; Knowles & Welch, 1992; Tollefson, 1987) and test reliability (Kolstad & Kolstad, 1991; Rich & Johanson, 1990).

Based on the findings of the current study, it is likely that using three-option MC items did not affect item and test psychometrics properties, hence it is much more practical compared to four-option MC items (Rodriguez, 2005). The findings related to NOTA option provides empirical support for the literature in that the use of NOTA as a distractor does not change item and test psychometric properties while the use of NOTA as key is problematic (DiBattista et al., 2014; Tollefson, 1987). We did not directly test how the use of NOTA as key influences the psychometric properties as the number of items with NOTA as key is few. DiBattista et al. (2014) claimed that examinees receive more scores than they deserve for the items with NOTA as key. This is supported by the item set #8 in Table 4. Item difficult of four-option MC item and the item with NOTA as key are 0.40 and 0.39, respectively. In other words, the use of NOTA as key decreases item difficulty by 0.01, which is a very small change. For the same item set, item discrimination of four-option MC item and the item with NOTA as key are 0.38 and 0.08, respectively, which is very dramatical change.. In educational perspective, the proportion of students who choose the correct answer for four option MC items and the item with NOTA as key are almost the same. However, the item with NOTA as key in the item set #8 does not make any differentiation among students with high ability and those with low as some of students with low ability choose correct answer for this item although they do not deserve.

Similar result was also observed in item set #6 in Table 4 although the item with NOTA as distractor was used in this set. While item difficult value decreases from 0.30 to 0.24, item discrimination value reduces from 0.43 to 0.16, which is not acceptable for a reliable and valid test. Besides, it is recommended that placement of key (correct answer) be balanced in constructing a reliable and valid test with MC items even if NOTA items are used. Considering position of NOTA option as a last option of a MC items, the use of NOTA as key is inevitable. Consequently, the use of NOTA is not recommended (DiBattista et al., 2014).

The findings of the current study should be considered regarding other factors in terms of item characteristics. First, because mathematics items are susceptible to back-solving (Author, 2013) and items in most other subjects typically are not, the impact of the number of distractors may affect other content areas differently. Therefore, use of three answer options and NOTA options may not influence item and test psychometric characteristics for mathematics items but items in other content areas.

The findings are also significant for mathematics teachers as well as designers of large-scale assessments. This research demonstrates that item writers and mathematics teachers may construct reliable and valid MC mathematics items with three answer options and NOTA options more quickly and easily. These items are less challenging and less time consuming for item writers as they require fewer plausible distractors than do conventional MC items with four options. This reduces the difficulty of constructing MC items.

Although this study showed no difference in the reliability of tests based on three-option MC items versus four-choice MC items when the number of items is the same, constructing a test with three-option MC items can increase test reliability since students require less time to respond, allowing more items to be applied in a given amount of time.

This study has several limitations. First, the test included only mathematics items constructed based upon one particular content area within the seventh grade in the United States. Future research needs to examine how psychometric characteristics of each MC item type vary when administering mathematics tests in different content areas or tests measuring different subject areas to students from different grades. With more research, we would be able to make generalizations from our current findings and apply them to a larger population.Second, the current study included 27 items: 9 MC items with four options, 9 MC items with NOTA options, and 9 MC items with three option. Longer tests may provide more opportunity for differences to manifest themselves.Third, the test forms A, B, and C in the current study included MC items with four options, MC items with NOTA options, and MC items with three options, respectively. The test was applied in this fixed order to all of the students. This causes a systematic error due to order effect. In the future, the tests should be designed to counterbalance the order of MC items in order to decrease error. Fourth, when converting conventional MC items with four options to MC items with NOTA options and three options, we eliminated an option of any item via mixed methods, such as the most, the second, and the least selected. Using only one of the elimination methods might lead to different results.

The validity of the inferences one can make from test scores rests on the quality of the items making up the test. Improving this quality will require more empirical study of the test development process. Unfortunately such studies remain relatively under-represented in the research base of the testing enterprise. This study is one small step to improve this situation. We encourage other researchers to add to this literature.

Acknowledgement

This paper is based on a chapter of Erkan Hasan Atalmis's Ph.D. dissertation prepared under the supervision of Neal Martin Kingston at the School of Education, University of Kansas.

REFERENCES

Abad, F. J., Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number of alternatives from the Item Response Theory. *Psicothema*, *13*(1), 152-158.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*(2), 192-211.

Burton, S.J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. *Brigham Young University Department of Instructional Science*. Retrieved from http://testing.byu.edu/info/handbooks/betteritems.pdf

Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement*, *54*(4), 861–872. doi:10.1177/0013164494054004002

Common Core State Standards Initiative (CCSSI). (2010). Common Core state standards for mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Crehan, K. D., & Haladyna, T. M. (1991).The validity of two item-writing rules. *The Journal of Experimental Education*, *59*(2), 183–192. doi:10.1080/00220973.1991.10806560

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, *53*(1), 241-247. doi:10.1177/0013164493053001027

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). A comparison between three-and four-option multiple choice questions. *Procedia-Social and Behavioral Sciences*, *98*, 398–403. doi:10.1016/j.sbspro.2014.03.432

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment, 14*(3), 197–201. doi:10.1027/1015-5759.14.3.197

DiBattista, D., Sinnige-Egger, J., & Fortuna, G. (2014). The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education, 82*(2), 168-183. doi:10.1080/00220973.2013.795127

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education, 10*(2), 133-143.doi:10.1007/s10459-004-4019-5

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*(5), 792–808. doi:10.1037/0021-9010.89.5.792

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education, 4*(2), 115–124. doi:10.1207/s15324818ame0402_2

Frey, B.B., Petersen, S., Edwards, L. M., Pedrotti, J.T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*(4), 357–364.doi:10.1016/j.tate.2005.01.008

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

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and psychological measurement, 53*(4), 999–1010.doi:10.1177/0013164493053004013

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choiceitem-writing guidelines for classroom assessment. *Applied Measurement in Education,15*(3), 309–334. doi:10.1207/S15324818AME1503_5

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

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business 73*(2), 94–97. doi:10.1080/08832329709601623

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Kingston, N.M., & Kramer, L.B. (2013). Highstakes test construction and test use. In T.D. Little (Ed.),*The Oxford handbook of quantitative methods: Vol. 1 foundations* (pp. 189–205). New York, NY: Oxford University Press.

Kingston, N.M., Scheuring, S.T., & Kramer, L.B. (2013). Test Development Strategies. In Kurt Geisinger (Ed.) *APA Handbook of Testing and Assessment in Psychology*. Washington, DC: APA Books.

Knowles, S. L.,&Welch, C. A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using "none-of-the-above." *Educational and Psychological Measurement, 52*(3), 571–577. doi:10.1177/0013164492052003006

Kolstad, R. K., & Kolstad, R. A. (1991). The option "none of these" improves multiple-choice test items. *Journal of Dental Education, 55*(2), 161–163.

Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement, 53*(3), 771–778. doi:10.1177/0013164493053003021

Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, *30*(1), 99-123.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher, 26*(8), 709–712. doi:10.1080/01421590400013495

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, *15*(8), 873–885. doi:10.1080/09658210701746621

Rich, C. E., & Johanson, G. A. (1990, April).*An item-level analysis of "none of the above."*Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. Retrieved from http://files.eric.ed.gov/fulltext/ED400299.pdf

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysisof 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13. doi:0.1111/j.1745-3992.2005.00006.x

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. doi:10.1177/00131649921969820

StataCorp (2013). Stata: release 13 - statistical software. College Station, TX: StataCorp LP

Shizuka, T., Takeuchi, O., Yashima, T. & Yoshizawa, K. (2006). A comparison of three-and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, *23*(1), 35–57. doi:10.1191/0265532206lt319oa

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis.*BMC Medical Education*, *9*(1), 40. doi:10.1186/1472-6920-9-40

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three-and four-option multiple-choice questions in nursing assessments. *Nurse Education Today*, *30*(6), 539–543. doi:10.1016/j.nedt.2009.11.002

Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using the"none of the above" and one correct response options. *Educational and Psychological Measurement, 47*(2), 377–383.doi:10.1177/0013164487472010

Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement*, *51*(4),829–837. doi:10.1177/001316449105100404

TÜRKÇE GENİŞLETİLMİŞ ÖZET

Çoktan seçmeli sorular diğer soru türlerine göre doğru ve objektif puanlama yapabilme, uygulama ve puanlama kolaylığı sağlama avantajlarından dolayı farklı alanlarda yapılan değerlendirme süreçlerinde yaygın olarak kullanılmaktadır (Haladyna, Downing, & Rodriguez 2002; McCoubrie, 2004). Doğru cevap seçeneği ve çeldiricilerden oluşan çoktan seçmeli soruları yazarken dikkat edilmesi gereken önemli noktalardan birisi de uygun ve mantıklı çeldiriciler oluşturmaktır (Haladyna ve diğerleri, 2002). Çünkü çeldiriciler soruların kalitesini gösteren madde ve testin psikometrik özelliklerini sağlamlaştırmada önemli rol oynamaktadır. Ancak uygun ve mantıklı çeldiricileri oluşturmak madde yazma sürecinin zor bir aşamasıdır (Haladyna & Downing, 1989; Hansen& Dexter, 1997; Rich & Johanson, 1990). Bu nedenle, bu nitelikteki çeldiricileri oluşturmak için soru yazan kişilerin önemli bir bilgi alt yapısına ve soru hazırlama süresine ihtiyaçları vardır. Bu durum gerek zaman gerekse zaman maliyetine paralel olarak para maliyetini artırmaktadır (Haladyna & Downing, 1993; Haladyna & Rodriguez, 2013).

Etkili çeldiricileri daha maaliyetsiz bir şekilde yazmak için bazı alternatif yaklaşımların olduğu bu konu ile ilgili olarak yapılan önceki çalışmalarda görülmektedir. Örneğin, daha az seçenek içeren ya da son seçenek olarak "Hiçbiri" seçeneğinin kullanımı, çoktan seçmeli soruları daha kolay bir şekilde oluşturmak için kullanılabilecek yaklaşımlardandır. Haladyna, Rodriguez ve Haladyna (2002, s.341) tarafından yapılan çalışmada, geçerli soru yazma ilkeleri açıklanmıştır. Bu ilkelerden 18.si "*Olabildiği kadar uygun ve mantıklı çeldiricileri yazınız*" ve 25.si "*Hiçbiri seçeneğini dikkatli kullanınız"* dır. Yapılan az sayıdaki ampirik çalışmalar bu iki ilkenin soru kalitesini nasıl etkilediğini araştırmış ve bu amaç doğrultusunda madde ve testin psikometrik özelliklerini hesaplayarak karşılaştırmışlardır. Çalışmalarda madde istatistikleri için madde zorluk ve madde ayırt edicilik indeksi hesaplanırken, test istatistiklerinden test güvenirlik katsayısı hesaplanmıştır. Madde zorluk indeksi, doğru cevabı seçen öğrencilerin oranı olarak tanımlanırken, madde ayırt edicilik indeksi ise soruların bilen öğrenci ile bilmeyen öğrenciyi birbirinden ayırmanın derecesini göstermektedir. Test güvenirliği ise bir ölçme aracının ölçme hatalarından arınık ve tutarlı ölçümler yapması olarak tanımlanmaktadır.

Yapılan önceki çalışmaların birçoğunda, "Hiçbiri" seçeneğinin kullanımının madde zorluğu üzerindeki etkisi araştırılmıştır. Bu çalışmaların bir kısmında "Hiçbiri" seçeneğinin madde zorluğunu istatistiksel olarak anlamlı bir şekilde değiştirmediğini bulurken (DiBattista, Sinnige-Egger, & Fortuna, 2014 ; Knowles ve Welch, 1992), çalışmaların çoğunda ise "Hiçbiri" seçeneğini kullanmanın soruyu istatistiksel olarak zorlaştırdığı ortaya çıkmıştır (Crehan, Haladyna, & Brewer, 1993; Frary, 1991; Kolstad & Kolstad, 1991; Rich & Johanson, 1990; Tollefson, 1987). Madde ayırt edicilik indeksine bakıldığında çalışmaların çoğunda, "Hiçbiri" seçeneğinin madde ayırt edicilik indeksini istatistiksel olarak anlamlı bir şekilde değiştirmediği bulunurken (Crehan ve Haladyna, 1991; DiBattista ve diğerleri, 2014; Frary, 1991; Knowles & Welch, 1992; Tollefson, 1987), Rich ve Johanson (1990) tarafından yapılan araştırma sonucunda ise "Hiçbiri" seçeneğinin madde ayırt edicilik indeksini artırdığı ortaya konulmuştur. Son olarak test güvenilirliği ile yapılan çalışmalarda ise "Hiçbiri" seçeneğinin testlerin güvenirliğini istatistiksel olarak anlamlı bir şekilde değiştirmediği bulunmuştur (Kolstad & Kolstad, 1991; Rich & Johanson, 1990).

Dört seçenekli ve üç seçenekli çoktan seçmeli sorulara yönelik yapılan sınırlı sayıdaki ampirik çalışmalara bakıldığında madde zorluğuna yönelik bazı çalışmalarda üç seçenekli soruların, maddeleri daha zorlaştırdığı görülürken (Landrum, Cashin, & Theis, 1993; Rogers & Harley, 1999), Rodriguez (2005) tarafından yapılan ve 1925 ile 1999 yılları arasındaki 48 ampirik çalışmanın incelenmesini içeren meta-çalışmada ise üç seçenekli soruların maddeleri daha da kolaylaştırdığı bulunmuştur. Madde ayırt edicilik indeksine yönelik çalışmalara bakıldığında ise çalışmaların çoğunda madde ayırt edicilik indeksinin üç seçenekli sorular ve dört seçenekli sorular arasında istatistiksel olarak farklılık görülmediği bulunurken (Cizek & O'Day, 1994; Crehan et al., 1993; Dehnad, Nasser, & Hosseini, 2014; Delgado & Prieto, 1998; Rogers & Harley, 1999; Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006; Tarrant & Ware, 2010), bazı çalışmalarda üç seçenekli soruların madde ayırt edicilik indeksini artırdığı ortaya konulmuştur (Baghei &Amrahi, 2011; Landrum et al., 1993; Rodriguez, 2005; Trevisan et al., 1991). Test güvenirliğine yönelik yapılan çalışmalarda ise farklı sonuçlar bulunmuştur. Bazı çalışmalarda üç seçenekli ve dört seçenekli sorular arasında test güvenirliğinin istatistiksel olarak anlamlı bir şekilde farklılık göstermediği ortaya konulurken (Baghei & Amrahi, 2011; Delgado & Prieto, 1998; Rogers & Harley, 1999), diğer çalışmalar ise üç seçenekli soruların test güvenirliğini artırdığı ortaya konulmuştur (Rodriquez, 2005; Tarrant & Ware, 2010). Sonuç olarak gerek üç seçenekli soruların gerekse "Hiçbiri" seçeneğinin kullanımının, soruların madde ayırt edicilik ve test gvenirliğine zarar vermediği bulunmuştur.

Yapılan bu araştırmalara bakıldığında seçenek sayısının ve "Hiçbiri" seçeneğinin kullanımının madde ve test istatistiklerini nasıl etkilediği farklı alanlar için incelenmiştir. Bu alanlar tıp, hemşirelik, psikoloji, sözcük bilgisi, genel kültür ve nitel problemlerin yer aldığı istatistik problemleridir.

Önceki çalışmalardan farklı olarak bu çalışmada ise seçenek sayısının ve "Hiçbiri" seçeneğinin kullanımının matematik sorularının madde ve test psikometrik özelliklerine etkisi incelenmiştir. Çoktan seçmeli matematik sorularının yazımında öğrencilerin muhtemel hatalarına uygun çeldiricileri bulmak matematik öğretmenleri ve soru yazan kişiler için oldukça uğraştıcı ve zaman alıcıdır. Bu sebepten dolayı daha az seçenekli ve "Hiçbiri" seçeneğini içeren çoktan seçmeli soruların yazılması soru yazımı için harcanan emeği ve zamanı azaltmayı sağlayacaktır. Yine bu çalışmada önceki çalışmalardan farklı olarak üç seçenekli, dört seçenekli ve "Hiçbiri" seçeneğinin madde ve test istatistiklerine etkisi eş zamanlı olarak bakılmış, üç seçenekli soruların "Hiçbiri" seçeneği içeren sorularla karşılaştırılmasına olanak tanınmıştır. Bu bağlamda aşağıda yer alan araştırma sorularına yanıt aranmıştır:

1. Madde psikometrik özellikleri (madde zorluğu ve madde ayırt ediciliği), farklı formattaki çoktan seçmeli matematik soruları (üç seçenekli, dört seçenekli ve "Hiçbiri" seçeneğini içeren dört seçenekli çoktan seçmeli sorular) arasında değişim göstermekte midir?
2. Test güvenirliği, farklı formattaki çoktan seçmeli matematik soruları arasında değişim göstermekte midir?

Araştırma sorularını incelemek için öncelikle bu çalışmada test geliştirilmiştir. Test geliştirilirken öncelikle Amerika Birleşik Develetlerin'de birçok eyalet tarafından kullanılan ortak olarak kullanılan "Common Core State Standards" programına uygun olarak dört seçenekli 58 çoktan seçmeli matematik sorusu yazılmış ardından sorular Amerika'da seçilen bir devlet okulundaki 7. ve 8. sınıf 100 öğrenciye pilot grup olarak uygulanmıştır. Ardından 58 soru içerisinden psikometrik olarak uygun değerlere sahip 27 çoktan seçmeli soru seçilerek, test gerçek grupta uygulanmak üzere yeniden düzenlenmiştir. Bu sorular düzenlenirken sorulardan 9'u dört seçenekli soru olarak değiştirilmeden bırakılmış, 9 soru üç seçenekli çoktan seçmeli soru ve 9 soru ise son seçeneği "Hiçbiri" olan soru formatına dönüştürülmüştür. Herbir gruptaki sorular kazanım ve çeldiri yazımı olarak birbirine oldukça paralel sorulardan oluşmaktadır. Dört seçenekli sorular üç seçenekli sorulara ve "Hiçbiri" seçeneği içeren sorulara dönüştürülürken rasgele olarak bir çeldiri silinmiştir. Son olarak yeni formattaki 27 soru bir test altında birleştirilerek Amerika'daki 5 okulda 585 7. ve 8.sınıf öğrenciye uygulanmıştır.

Uygulama sonucunda dört seçenekli, üç seçenekli ve "Hiçbiri" seçenekli soruların madde zorluk indeksleri, madde ayırt edicilik indeksleri ve bulundukları testlerin test güvenirliği hesaplanmıştır. Bu araştırmanın 1. sorusunun cevabını bulmak için dört seçenekli, üç seçenekli ve "Hiçbiri" seçeneğini içeren testler arasındaki karşılaştırmayı yapmak için tekrarlı ölçümler için tek faktörlü varyans analizi (ANOVA) uygulanmıştır. Araştırmanın 2. sorusu olan gruplar arasındaki test güvenirlik farkını hesaplamak için de herbir testin Croanbach's Alpha güvenirlik katsayısı ve bu katsayısıların standart hatası hesaplanmıştır. Analiz sonucunda, seçenek düzenlemesinin madde zorluk derecesine istatistiksel olarak anlamlı bir etkisi olduğu görülmüştür *(F*(2,7) = 7.97, *p*< .05.) Uygulanan Bonferroni testi sonucunda, madde zorluk indeksleri karşılaştırıldığında dört seçenekli soruların madde zorluk indeksleri ortalamasının ($M_{dört}$= 0.46, $SD_{dört}$=0.15) "Hiçbiri" seçeneğini içeren çoktan seçmeli soruların ($M_{hiçbiri}$=0.41, $SD_{hiçbiri}$=0.13) ve üç seçenekli soruların ($M_{üç}$= 0.52, $SD_{üç}$= 0.8) madde zorluk indekslerinin ortalamalarından istatistiksel olarak farklı olmadığı bulunmuştur. Fakat, "Hiçbiri" seçeneğini içeren soruların madde zorluk indekslerinin ortalamalarının üç seçenekli soruların madde zorluk indeklerininin ortalamalarında istatistiksel olarak düşük olduğu görülmüştür. Madde ayırt edicilik indeksine bakıldığında üç soru formatının da madde ayırt edicilik indeksini istatistiksel olarak anlamlı bir şekilde değiştirmediği ortaya çıkmıştır (*F*(2,7) = .22, *p*=.80). Diğer bir ifade ile dört seçenekli soruların madde ayırt edicilik indeksleri ortalamasının ($M_{dört}$= 0.45, $SD_{dört}$=0.07) "Hiçbiri" seçeneğine sahip çoktan seçmeli soruların ($M_{hiçbiri}$=0.45, $SD_{hiçbiri}$=0.16) ve üç seçenekli soruların ($M_{üç}$= 0.44, $SD_{üç}$= 0.09) madde ayırt edicilik indeks ortalamalarından istatistiksel olarak farklı olmadığı görülmüştür. Test güvenirliği hesaplandığında, 27 soru için testin güvenirliği 0.84 olarak bulunmuştur. Soru formatlarına göre incelendiğinde, dört seçenekli, üç seçenekli ve "Hiçbiri" seçeneğinden oluşan soru gruplarının Croanbach Alpha katsayı değerleri sırasıyla 0.65, 0.63 ve 0.58 olarak bulunurken, standart hata değerleri ise ±0.02, ±0.03 ve ±0.02 olarak hesaplanmıştır. Bu üç grubun 0.95 düzeyinde güven aralığı hesaplandığında sırasıyla (0.61, 0.69), (0.59, 0.67) ve (0.52, 0.64) olarak bulunduğu ve bu aralıklar birbirleriyle kesiştiğinden dolayı dört seçenekli, üç seçenekli ve "Hiçbiri" seçeneğinden oluşan soruların test güvenirliklerinin birbirinden farklı olmadığı ortaya çıkmıştır.

Sonuç olarak dört seçenekli, üç seçenekli ve "Hiçbiri" seçeneğinden oluşan çoktan seçmeli matematik soruları karşılaştırıldığında test güvenilirliği ve madde ayırıcılık indeksinin etkilenmediği, "Hiçbiri" seçeneğini içeren soruların üç seçenekli sorulara göre daha zor olduğu bulunmuştur.

Bu sonuçlardan hareketle çoktan seçmeli matematik sorularının seçenekleri yazılırken dört seçenek yerine üç seçenek ve "Hiçbiri" seçeneğinin kullanımının sorunun psikometrik özelliklerini değiştirmediği ifade edilebilir. Bu durumun özellikle soru yazan uzmanlarınve matematik öğretmenlerinin dört seçenekli sorular yerine 3 seçenekli ve "Hiçbiri" seçenekli sorular kullanarak soru yazmak için daha az zaman harcamasına olanak sağlayacağı görüşünü desteklemektedir.