iojes
www.iojes.net

International Online Journal of Educational Sciences

ISSN: 1309-2707

Examining Gender Bias in Multiple Choice Item Formats Violating Item-Writing Guidelines

Research Article

Mehmet KAPLAN[1], Erkan Hasan ATALMIS[2]

[1] *Artvin Coruh University, Turkey,* ORCID: 0000-0002-4175-3899
[2] *Kahramanmaras Sutcu Imam University, Turkey,* ORCID: 0000-0001-9610-491X

To cite this article: Kaplan, M., Atalmis, E. H. (2019). Examining Gender Bias in Multiple Choice Item Formats Violating Item-Writing Guidelines, *International Online Journal of Educational Sciences*, 11(1), ??.

ARTICLE INFO

Article History:
Received: 30.11.2018

Available online:
09.01.2018

ABSTRACT

Multiple choice items (MCIs) are commonly used in high-stake testing and classroom assessment because of their reliable assessment results. However, the recent literature has revealed that item-writing guidelines have been repeatedly violated in creating MCIs, which could also threaten reliability and validity. Another threat to the validity occurs when items favor certain groups even though the magnitude of underlying ability of the different groups is the same, and this is called differential item functioning (DIF). This empirical study aims to compare item parameters for MCIs with negative wording stem and complex MCIs, which are commonly used MCI formats that violate item-writing guidelines for MCIs, and to investigate the impact of DIF on gender differences considering the use of these item formats. The results of this study showed that DIF detection methods flagged two complex MCIs favoring male students because of the item format and tendency of male students' taking more risk on solving MCIs.

Keywords:
Multiple choice items, guideline violation, complex multiple choice item, negative stem, DIF

Introduction and Background

Classroom assessment plays an important role in determining and assessing the level of knowledge and cognitive skills of students, and their learning progressions. Although different assessment methods are facilitated in the educational context, tests comprising short answer questions and multiple choice items (MCIs) are commonly used in classroom assessments due to high reliability and objectivity of their results. Critical decisions, such as pass-or-fail decision and proficiency level determination, can be made about individuals based on the results of such kind of tests.

[1] Corresponding author's address: Artvin Coruh Univeristy
Telephone: +90 466 215 1043
Fax: +90 466 215 1042
e-mail: mehmet.kaplan2@gmail.com
DOI: https://doi.org/10.15345/iojes.2019.01.015

Tests are generally composed of selected-response (e.g., multiple choice, true or false, or matching) and constructed-response (e.g., short answer or essay) items separately, or possibly the combination of them. Although both item formats are widely used in classroom assessments, Thorndike (2005) expressed that the choice of item formats primarily depends on the content demand, test administration time, and cognitive and scoring process. The main advantages of using selected-response items include objective scoring, easier and faster administration, and a broad range of learning outcomes in a short time (Collins, 2006; McCoubrie, 2004). In other words, this type of items provides practicability, and high degree of reliability and validity in assessments. Therefore, the usage frequency of the selected-responses items has recently increased in classroom testing (Caldwell, 2008).

Recent developments in educational assessment have also put great emphasis on differential item functioning (DIF). DIF occurs when examinees from different groups show different probabilities of success on an item after the underlying ability is matched (Zumbo, 1999). Moreover, items with DIF effect produces bias results which can threaten test validity and score reliability. The available literature shows that DIF has been mostly examined with respect to demographic characteristics such as gender, education, social class, ethnicity, and age. It is noteworthy that a very limited number of studies have focused on the potential sources of DIF resulted from MCIs with negative wording stem and complex MCIs, which are commonly used MCI formats that violate item-writing guidelines for MCIs. In this regard, this study aims to examine gender bias in MCI formats violating item-writing guidelines.

Item Formats in the Educational Context

The selected-response item formats are separated into different categories such as conventional MCIs, sentence-completion items, true-false items, and matching items in the literature (Haladyna & Rodriguez, 2013), and different formats can be used in the test design based on the kinds of skills and the content to be measured (Cohen & Wollack, 2004). For example, MCIs, the most commonly used selected-response items, are employed to assess a wide range of higher-order thinking skills, whereas short answer questions under constructed-response items can be used for disclosing deep learning steps in a cognitive process. MCIs have been increasingly used at different levels of education because they are easy and quick to score, which increase the objectivity and reliability of the test scores. Moreover, with the rapid development of educational technology, computers lead to use more practical and innovative MCI formats with instant feedback to the test takers (Douglas, Wilson, & Ennis, 2012; Harter & Harter, 2004; Nicol, 2007). Nonetheless, constructing MCIs with high-quality can be challenging and time consuming for the item writers. For example, producing plausible distractors to reveal students' possible errors for any item can be difficult (Haladyna & Downing, 1993; Haladyna, Downing, & Rodrigez, 2002), and previous studies showed that most of the MCIs used in state-wide examinations had poor functioning distractors (Delgado & Prieto, 1998; Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006; Tarrant, Ware, & Mohammed, 2009; Terzi & Yakar, 2018). Because constructing distractors that function adequately requires considerable time and skills for the item writers, they tend to use different MCI formats in the classroom assessment such as complex MCIs (i.e., K-type) and MCIs with negative stem rather than conventional MCIs to save time, effort, and cost. The appropriateness of using these MCI formats has been discussed with respect to item-writing guidelines for classroom assessment and high-stake testing to construct reliable scores and valid tests.

The origin of the item-writing guidelines dates backs to 1980s, and more than 40 item-writing guidelines were proposed after reviewing textbooks on measurement and evaluation in testing (Haladyna & Downing, 1989a, 1989b). Later, Haladyna et al. (2002) revised the proposed guidelines, and classified them into five categories based on content, format, style, and forming the stem, and the choices. Frey, Petersen, Edwards, Pedrotti, and Peyton (2005) identified the 40 most commonly used item-writing guidelines after reviewing and evaluating approximately 20 classroom assessment textbooks. Recently, Moreno, Martínez, and Muñiz

(2015) re-designed the existing item-writing guidelines, and decreased the number of the guidelines to 9 by evaluating them in terms of validity. Even though there has been a sufficient number of studies on item-writing guidelines in the literature, the previous studies have shown that the item-writing guidelines have been violated across different disciplines (e.g., auditing, medicine, and nursing), especially in high-stake testing (Tarrant & Ware, 2008), creating item banks (Hansen & Dexter, 1997; Masters et al., 2001), and classroom assessment (Pate & Caldwell, 2014; Tarrant, Knierim, Hayes, & Ware, 2006).

To give an example, Table 1 shows three different MCI formats measuring the same content. Specifically, the first column of Table 1 represents a conventional MCI, the second represents a complex MCI, and the last column represents a MCI with negative stem. Considering the complex MCIs and MCIs with negative stem in Table 1, two item-writing rules have been explicitly violated based on the guidelines in Haladyna et al. (2002): (i) "AVOID the complex MC format" and (ii) "Word the stem positively, avoid negatives such as NOT or EXCEPT" in creating items. These item formats are not recommended in the test design because they lower the test reliability and validity (Haladyna & Downing, 1989b; Parker & Somers, 1983). In addition, previous empirical studies also showed that the two guidelines have been commonly violated in testing. Furthermore, previous empirical studies have also showed that the two guidelines are commonly violated in testing. Unfortunately, a limited number of empirical studies have addressed the influence of the use of MCIs with negative stem and complex MCIs based on the psychometric properties of the items (e.g., item difficulty and discrimination), which are directly related to the test reliability and validity. The studies in concern showed that the use of complex MCIs increases item difficulty compared to the same item written in the conventional MCI format in anatomy and pharmacology (Nnodim, 1992; Tripp & Tollefson, 1985).

Table 1. Item types using different MCI formats

Conventional MCIs	**Complex MCIs**	**MCIs with negative stem**
Which is a city in Europe?	I. Berlin	Which city is NOT in Europe?
A. Berlin*	II. Chicago	**A.** Athens
B. Chicago	III. Paris	**B.** Berlin
C. Mexico City	Which city (or cities) is/ are in Europe?	**C.** Paris
D. New Delphi	**A.** only I **B.** only III	**D.** Chicago*
	C. I and II **D.** I and III*	

* Key of the items

The existing literature yields contradictory results related to the use of MCIs with negative stem. Namely, MCIs with negative stem were reported easier than conventional MCIs (Harasym, Doran, Brant, & Lorscheider, 1993) while Tamir (1993) found that the use of MCIs with negative wording stem produced more difficult items for high-thinking levels. Atalmis (2016), on the other hand, concluded that the use of MCIs with negative stem did not statistically change item difficulty parameters; however, complex MCIs yielded more difficult items in comparison to conventional MCIs. Moreover, the literature review indicates that an insufficient number of studies compared MCIs with negative stem and conventional MCIs in terms of item discrimination. Nnodim (1992) found that complex MCIs were items more discriminating and difficult than conventional MCIs, whereas Tripp and Tollefson (1985) found item discrimination level of the complex MCIs was not statistically different from that of the conventional MCIs while complex MCIs were more difficult than conventional MCIs. Consequently, the literature states that the complex items generally reveal items with larger item difficulty parameters (more difficult) compared to the conventional MCIs; however, no consensus has been reached on the comparison of the complex MCIs and conventional MCIs in terms of item discrimination, and the comparison of MCIs with negative stem and conventional MCIs in terms of item difficulty.

Previous studies examining gender differences have also showed that MCIs favor male students and constructed-response items favor female students (Bolger & Kellaghan, 1990; DeMars, 2000; Klein et al., 1997). In another study, Ben-Shakhar & Sinai (1991) found that male students were more likely to use test taking strategies and guess the answer on MCIs when they were uncertain about the item.

Differential Item Functioning

Several DIF detection procedures have been proposed in the literature. These procedures are generally categorized based on either item response theory (IRT) approaches or non-IRT approaches. For example, Lord's chi-squared DIF method (Lord, 1980) is an IRT-based approach, whereas the Mantel-Haenszel (Holland & Thayer, 1988) and the logistic regression methods (Swaminathan & Rogers, 1990) are non-IRT based approaches used for DIF detection. In addition, two types of DIF effects, namely, uniform and non-uniform, could be examined using the logistic regression and Lord's chi-squared methods. In general, the main effects of group differences indicate uniform DIF and the interaction effects of group by ability indicate non-uniform DIF (Zumbo, 2007).

As previously mentioned, DIF has generally been investigated in terms of demographic factors such as gender and ethnicity. However, only few of these studies have focused on DIF effects by using different item formats (Hamilton, 1999; Hamilton & Snow, 1998; Liu & Wilson, 2009; Mazzeo, Schmitt, & Bleistein, 1993; Zenisky, Hambleton, & Robin, 2004). The results of these studies showed that MCIs favor male examinees while open-ended items favor mostly female examinees. Specifically, Liu and Wilson (2009) investigated the impact of DIF on the probability of success with respect to gender through examining both item domain (i.e., space and shape, and quantity) and item type (i.e., conventional MCIs, complex MCIs, and open constructed-response items). The authors analyzed the PISA 2000 and 2003 mathematics assessment results, and found the largest gender difference in unconventional item type in favor of male students. Hence, in this study, it is also aimed to analyze the impact of DIF on gender difference considering the use of MCIs with negative stem and complex MCIs.

In this regards, we have examined two research questions as follows:

- Does test difficulty change across MCIs with negative stem and complex MCIs?
- Does the use of MCIs with negative stem and complex MCIs cause DIF effect on gender difference?

Method

Participants

651 senior undergraduate students enrolled in a 28-week teaching certification program participated in this study. Participants had different academic backgrounds (e.g., history, geography, and business); however, they all completed the program successfully. The responses of 32 participants were removed from the data set because they failed to specify their gender which is a critical variable in DIF detection for the study. Therefore, the statistical analyses were performed with the sample size of 619. It is significant to note that approximately 58% of the participants were female (N = 361), and slightly over 42% of them were male (N = 258).

Instrument

A total of 32 MCIs with four options were developed for the final test to be administered at the end of a course in educational science. Three types of item formats were used in constructing the test including 10 conventional MCIs, 11 MCIs with negative stem, and 11 complex MCIs. The conventional MCIs were chosen from an item pool composed of items with high reliability and validity where the items were administered to 190 students in the previous year, and their difficulty and discrimination parameters were calculated based on the Classical Test Theory approach because of the small sample size. The results are shown in Table 2. In

this study, item difficulty was computed as the proportion of examinees who responded the item correctly while item discrimination was calculated as correlating individuals' item and total scores, which is one of the commonly used method (Downing, 2005). The item difficulty parameters ranged between .42 (item #1) and .95 (item #8) while the item discrimination parameters ranged from .28 (item #4) to .48 (item #7). Because item discrimination parameters of the items were greater than .25, all items were acceptable to use in subsequent tests (Thorndike, 2005). However, the MCIs with negative stem and complex MCIs were constructed for the purpose of this study by the two content experts in Measurement and Evaluation field, pursuing the following steps. First, the MCIs with negative stem and complex MCIs were strictly parallel items intended to measure the same content and to have the same rationale of distractors. To test how strictly parallel items, these experts created codes showing similarity and consistency of the items by using the formula proposed by Miles and Huberman (1994; Reliability = Agreement/(Aggrement and Disaggrement)). The findings indicated that the reliability of the codings was calculated as 0.99, which provides the reliability of coding.

Second, the stem of complex MCIs was constructed by randomly selecting three out of four options from the MCIs with negative stem. For example, four options of the MCI with negative stem in Table 1 were a) Athens, b) Berlin, c) Paris, and d) Chicago. The options b, c, and d were randomly selected to construct the stem of the complex MCI. Then, two forms consisting of 21 items on each, were created (i.e., Form A and Form B). In particular, 10 conventional MCIs were common items in both forms (i.e., anchor), and 11 MCIs with negative stem and 11 complex MCIs were included in Form A and Form B, respectively. Moreover, the order of the MCIs was designed as the conventional MCIs were followed by the MCIs with negative stem in Form A, and followed by the complex MCIs in Form B.

Table 2. Item difficulty and item discrimination indexes of the conventional MCIs

Items	Item difficulty	Item discrimination
#1	.42	.40
#2	.66	.35
#3	.90	.32
#4	.93	.28
#5	.89	.31
#6	.53	.46
#7	.44	.48
#8	.95	.28
#9	.87	.34
#10	.70	.42

Data Analysis

Form A and Form B were randomly assigned to the participants regardless of their academic background and gender. To investigate whether both forms were assigned to the groups comprising of examinees who did not significantly differ in terms of academic achievement, an independent t-test was conducted to compare the mean proportion of the correct responses of the anchor items in both forms. In addition, the psychometric properties of different item types (i.e., conventional MCIs, MCIs with negative stem, and complex MCIs) in different forms were also analyzed using IRT.

<table>
<tr><th>IDs</th><th colspan="3">Items</th></tr>
<tr><td>Participants given Form A</td><td rowspan="2">Anchor Items</td><td>MCIs with Negative Stem</td><td>-</td></tr>
<tr><td>Participants given Form B</td><td>-</td><td>Complex MCIs</td></tr>
</table>

Figure 1. The combined data sets for the IRT analyses

The data set was fitted to an appropriate IRT model, and item parameters were analyzed. To eliminate the data fit issues in the IRT analysis, the two data sets were combined to increase the sample size in item parameter calibration. As in shown in Figure 1, the data sets from the two forms were combined as all the participants responded to the anchors items (i.e., the conventional MCIs) while some of them responded the MCIs with negative stem rather than the complex MCIs, and vice versa. Therefore, missing data occurred on the reverse diagonal of the data set. In addition, how item difficulty parameters differed across two different item formats was examined (i.e., MCIs with negative stem and complex MCIs). In the second part of the study, the DIF detection methods were employed with respect to gender considering different item formats used in this study using through both non-IRT-based and IRT-based approaches. Specifically, a logistic regression method (non-IRT-based) and the Lord's chi-squared method (IRT-based) were used in DIF detection. The Wald test (Thissen, Steinberg, & Wainer, 1988) was also used as the test statistic for the logistic regression. Item was flagged as DIF if one of the methods resulted in a significant test statistics.

Results

Test Forms Equality

The random assignment of two forms to the participants might yield different group performances, and might have an impact on the results. To eliminate this issue, groups were formed with students who do not significantly differ in terms of academic achievement. Figure 2 shows the distribution of the total scores obtained from Form A and Form B. It can be inferred from the histograms that the scores are approximately normally distributed. An independent t-test was conducted to compare the mean proportion of correct responses provided to the anchor items (i.e., first 10 items) in both forms. The means and standard deviations of the total scores for the anchor items obtained from Form A (N = 310) and Form B (N = 309) were calculated $\bar{X}_{A\_anchor} = 7.46$ and $S_{A\_anchor} = 1.38$, and $\bar{X}_{B\_anchor} = 7.43$ and $S_{B\_anchor} = 1.45$, respectively. The t-test results showed that there was no statistically significant differences between the two forms ($t_{(617)} = 0.27$, $p = 0.79$) based on the anchor items. However, when the t-test was performed using the entire test for Form A and Form B, it was found that there was a significant difference between the forms ($t_{(617)} = 9.53$, $p = 0.00$) in terms of difficulty. Form B was relatively more difficult ($\bar{X}_B = 12.66$) compared to Form A ($\bar{X}_A = 14.50$) because it included the complex MCIs.

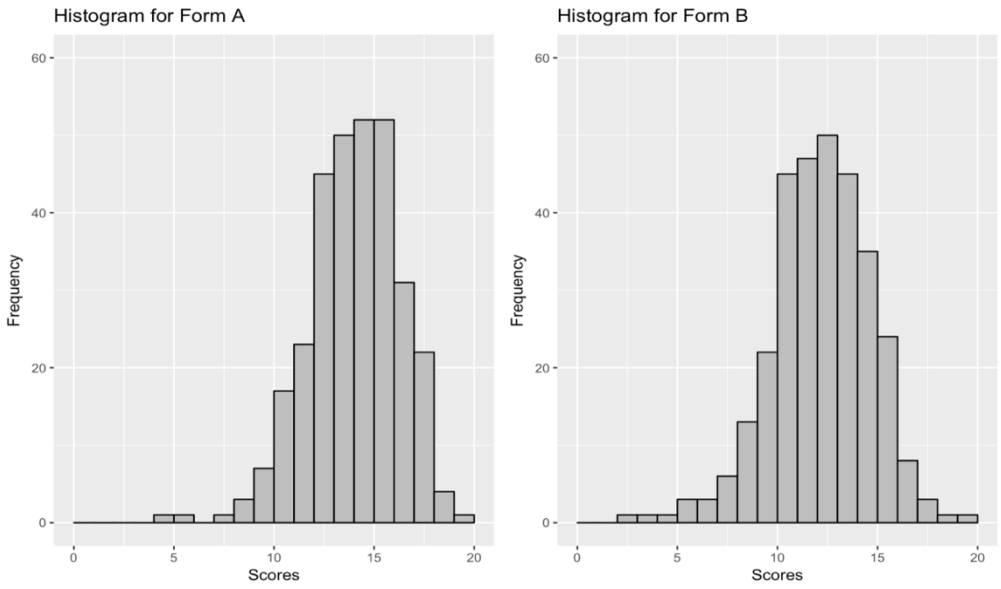

Figure 2. Distribution of the total scores obtained from Form A and Form B

Psychometric Properties of the Forms

Because of the small sample size, item parameters were estimated using the Rasch model, and analyses were conducted using "ltm" package in R (Rizopoulos, 2006). The unidimensionality assumption for the Rasch model was checked empirically (Drasgow & Lissak, 1983), and the result showed that the assumption was not violated. Table 2 shows the item difficulty parameters based on the Rasch model. Specifically, the first row shows the parameters for the entire test, the second row shows the parameters for the anchor items, the third row shows the parameters for the last 11 items in Form A (i.e., the MCIs with negative stem), and the last row shows the parameters for the last 11 items in Form B (i.e., the complex MCIs).

The mean of the item parameters across the total test was -0.77, which ranged from -4.36 to 1.47. This implies that the test was relatively easy because the mean of the parameters was below zero. Moreover, the means of the item difficulty for the MCIs with negative stem and complex MCIs were -0.93 and 0.04, respectively. Based on the results, the complex MCIs were more difficult and this was consistent with the findings in the literature (Atalmis, 2016; Nnodim, 1992; Tripp and Tollefson, 1985). Last, the mean of the anchor item parameters was -1.48, which ranged from -3.15 to 0.49.

Table 3. Descriptive statistics of item difficulty parameters based on the Rasch Model

	Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
The Entire Test	-4.36	-1.97	-0.71	-0.77	0.38	1.47
Anchor Items	-3.15	-2.37	-1.49	-1.48	-0.72	0.49
MCIs with Negative Stem (Only in Form A)	-4.36	-1.53	-0.74	-0.93	0.18	0.64
Complex MCIs (Only in Form B)	-2.91	-0.65	0.48	0.04	0.92	1.47

DIF Detection

The DIF analyses were conducted using "difR" package in R (Magis, Beland, Tuerlinckx, & De Boeck, 2010). A logistic regression method and Lord's chi-squared DIF method using the Rasch model were performed for the DIF detection in this study. Specifically, the former models the test score, group membership, and interaction between these two as covariate, and tests for the significant main and interaction

effects. The latter tests the hypothesis that the item difficulty parameters obtained from the model for one group are equal to those in a second group. Moreover, it should be noted that the anchor items were excluded from the DIF analyses in this study.

Based on the logistic regression model and using the Wald test as the test statistic, Item 30 showed uniform DIF effect favoring the focal group (i.e., male participants). In other words, it was assumed that there was no difference between male and female participants with respect to ability levels in this content; nonetheless, the male participants took advantage of correctly responding Item 30. This occurred because Item 30 was a complex MCI and male students had a higher tendency to take risk and guess it correctly (Ben-Shakhar & Sinai, 1991). Therefore, the violation of item-writing guidelines could have caused DIF effect in this particular test. However, all the effect sizes based on the measure proposed by Zumbo and Thomas (1997), and Jodoin and Gierl (2001) were negligible using this method. In addition, using the Lord's chi-squared DIF method, Items 30 and 34, complex MCIs, showed uniform DIF effect and again favored male students. Moreover, Items 13, 26, 33, and 34 showed moderate, and Item 30 showed large effect sizes based on the measure proposed by Holland and Thayer (1985).

In summary, the first 10 items were conventional MCIs, and the item-writing guidelines were obeyed in this type of items. The same items were written in two different formats (i.e., the MCIs with negative stem and complex MCIs) and caused possible DIF effect, especially with complex MCIs. Item 30 was flagged as DIF by two methods, which should be removed from the test evaluation.

Discussion

The complex MCIs and MCIs with negative stem are the most commonly used unconventional MCI formats because creating distractors for them is relatively easy but not plausible (Boland, Lester, & Williams, 2010). However, these two item formats can cause issues in reliability and validity because of the item-writing guideline violations for MCIs. In addition, DIF could lead to fairness issues in educational assessment. The purpose of this study was to investigate the impact of DIF on gender differences considering the two different item formats, the complex MCIs and MCIs with negative stem.

The previous studies emphasized the negative impact of using MCIs with negative stem and complex MCIs on the psychometric properties of the item. For example, the complex MCIs produced larger IRT item difficulty parameters compared to the conventional MCIs. However, the findings of the studies were inconsistent extent to which the use of MCIs with negative stem influences item difficulty. Even though several studies have compared MCIs with negative stem and complex MCIs with the conventional MCIs, only a few of them investigated the comparison of these two item formats considering DIF. Accordingly, this particular study aimed to compare them in terms of psychometric properties, and to examine whether their use could cause DIF. A total of 32 MCIs was developed using the conventional MCIs, MCIs with negative stem, and complex MCIs. In the first part, the psychometric properties of these two item formats were analyzed based on the Rasch model. In the second part of the study, two DIF detection methods were employed based on gender differences using different item formats. The results of the study showed that the difficulty levels of the complex MCI format were found relatively higher compared to the entire test. In DIF detection, the logistic regression model flagged one and the Lord's chi-squared DIF method flagged two different items as DIF, favoring the male students. Specifically, the former method flagged Item 30, and the latter method flagged Items 30 and 34 as uniform DIF. Moreover, four items showed moderate and one item showed large effect sizes based on gender. Because of the design of the study, the conventional MCIs did not show any DIF; however, when the items were created using complex MCI format, a DIF effect was detected because of the tendency of male students. The use of MCIs with negative stem did not yield any DIF effect based on gender difference in this study.

The findings of this study could exceptionally contribute to the growing body of research focusing on violation of item-writing guidelines. They supported the findings from previous studies, and showed that the violation of item-writing guidelines can be a source of DIF based on gender differences. However, more research needs to be done to fully understand the potential source of DIF related to the item-writing guideline violations. First, this study was conducted using a final exam administered at the end of the semester, and results cannot be generalizable due to the sample size and sampling. A large study should be designed to compare the impact of using different item formats for high-stake testing to obtain more generalizable results. Second, only gender differences were taken into account for the DIF detection related to the item formats in this study. However, considering other demographic characteristics such as ethnicity or social class would be important in DIF detection in this concern. Third, although the design of this study consisted of creating parallel test forms with negative stem and complex MCIs, and comparing these two formats, it would be interesting to investigate DIF when these two item formats are created from conventional MCIs, and how the psychometric properties of the items could change. Fourth, this study was limited to the investigation of two item-writing guideline violations. Hence, the other item-writing guideline violations in DIF can be further examined. Last, it is suggested in the literature that differential distractor functioning (DDF) analyses can shed light on designing fair tests for different groups at the same ability levels because DDF can cause DIF in the correct responses (Terzi & Suh, 2015; Terzi & Yakar, 2018). In a future study, it could be interesting to investigate MCI and DIF along with DDF analyses.

GENİŞLETİLMİŞ ÖZET

Soru Yazma İlkelerinin İhlal Edildiği Çoktan Seçmeli Sorularda Cinsiyet Yanlılığının İncelenmesi

Sınıf içi değerlendirme, öğrencilerin bilgi düzeyinin, bilişsel becerilerinin ve öğrenme ilerlemelerinin belirlenmesinde önemli bir rol oynamaktadır. Yüksek güvenilirliği ve objektifliği nedeniyle genellikle kısa cevaplı ya da çoktan seçmeli madde türlerinden oluşan testler, bu gibi değerlendirmelerde yaygın olarak kullanılmaktadır. Ayrıca, bu tür testlerin sonuçlarına dayanılarak bireyler hakkında geçme-kalmaya yönelik karar alınabilmekte ya da bu bireylerin belirli bir konu alanında yeterlik düzeyleri belirlenebilmektedir.

Testler genellikle, seçmeli (çoktan seçmeli, doğru-yanlış veya eşleştirme) ve yazmalı maddeler (kısa cevap veya kompozisyon) veya bunların kombinasyonundan oluşmaktadır. Her iki madde formatı da test katılımcıları için pratik olmasına rağmen, Thorndike (2005) madde formatı seçiminin öncelikli olarak içerik talebine, test yönetim süresine ve puanlama sürecine bağlı olduğunu ifade etmiştir. Seçmeli maddelerinin kullanılmasının başlıca avantajları arasında nesnel puanlama, kolay ve hızlı yönetim ve kısa sürede çok sayıda öğrenme çıktıları bulunmaktadır (Collins, 2006; McCoubrie, 2004). Başka bir deyişle, bu tür maddeler değerlendirmelerde uygulanabilirlik ve yüksek derecede güvenilirlik ve geçerlilik sağlamaktadır. Bu nedenle, seçmeli maddelerin kullanım sıklığı son zamanlarda sınıf içi değerlendirme artma eğilimi göstermiştir (Caldwell, 2008).

Eğitimsel değerlendirmedeki son gelişmeler, değişen madde fonksiyonuna (DIF) büyük önem vermiştir. DIF, farklı gruplar için bu gruplardaki kişilerin yetenek seviyeleri eşleştirildikten sonra belirli konu alanında farklı başarı olasılıkları gösterdiğinde ortaya çıkar (Zumbo, 1999). Ayrıca, DIF etkisine sahip maddeler, test geçerliğini ve skor güvenilirğini tehdit edebilecek yanlılık sonuçları üretebilir. Bu konudaki ilgili literatür incelendiğinde, DIF'in çoğunlukla cinsiyet, eğitim, sosyal sınıf, etnik köken ve yaş gibi demografik özellikler açısından incelendiğini göstermektedir.

Çalışmanın Amacı

Alanyazında çok sınırlı sayıda çalışma, çoktan seçmeli sorular için madde-yazım kurallarını ihlal eden formatları olan negative köklü ve birleşik yanıt gerektiren madde yapılarına dikkat çekmiş ve bu durumun DIF'e sebep olacağı vurgusu yapmıştır. Ancak bu durum önceki çalışmalarda genellikle ampirik (deneysel) olarak test edilmemiştir. Bu bağlamda, bu çalışmanın amacı aynı içeriği ölçen bu tip çoktan seçmeli maddeleri oluşturmak ve cinsiyet farklılıklarına bağlı olarak DIF'in varlığını araştırmaktır. Bu bağlamda, aşağıdaki iki araştırma sorusu incelenecektir:

• Test zorluk dereceleri, negatif köklü ve birleşik yanıt gerektiren çoktan seçmeli maddeler için farklılık gösteriyor mu?

• Negatif köklü ve birleşik yanıt gerektiren çoktan seçmeli maddelerin kullanılması DIF'in cinsiyet farkı üzerine etkisine sebep oluyor mu?

Yöntem

Bu çalışmanın örneklemini 28 haftalık bir öğretim sertifikasyon programına kayıtlı farklı akademik geçmişe sahip 619 lisans son sınıf öğrencileri oluşturmaktadır. Katılımcıların yaklaşık %58'i kadın (N = 361) ve %42'si erkektir (N = 258). Çalışmada kullanılacak test için eğitim bilimlerinde bulunan bir dersin sonunda yapılacak final sınavı için dört seçenekli toplam 32 geleneksel çoktan seçmeli madde geliştirilmiştir. Ardından bu maddelerden 10'u aynı formatta korunmuş, 11'i negatif köklü ve kalan 11'i ise birleşik yanıt gerektiren çoktan seçmeli madde formatına dönüştürülmüştür. Negative köklü ve birleşik yanıt gerektiren çoktan seçmeli maddeler oluşturulurken aynı içeriği ölçecek ve aynı mantıksal gerekçelere sahip olacak şekilde

tasarlanmıştır. Diğer bir ifade ile bu iki formattaki maddeler birbirine parallel olacak şekilde tasarlanmıştır. Ardından her biri 21 maddeden oluşan iki test formu oluşturulmuştur (Form A ve Form B). Bu test formlarından her ikisinde de 10 geleneksel çoktan seçmeli sorular ortak olarak yer alırken, negatif köklü çoktan seçmeli sorular Form A'da birleşik yanıt gerektiren çoktan seçmeli sorular ise Form B'de bulunmaktadır.

Form A ve Form B test fomları oluşturulduktan sonra öğrencilerin akademik cinsiyet farklılıkları dikkate alınmaksızın öğrencilere rasgele dağıtılmıştır. Bu test formlarının eşitliğini doğrulamak için sadece ortak maddeler üzerinden bağımsız örneklem t-testi yapılırken, farklı fomattaki maddelerin (negative köklü ve birleşik yanıt gerektiren çoktan seçmeli maddeler) psikometrik özelliklerini karşılaştırmak için madde tepki kuramı uygulanmıştır. DIF tespiti için bir lojistik regresyon yöntemi (madde tepki kuram tabanlı olmayan) ve Lord'un ki-kare yöntemi (madde tepki kuram tabanlı) kullanılmıştır.

Bulgular

Öncelikle Form A ve Form B eşitliğini göstermek için sadece ortak 10 madde üzerinden yapılan bağımsız örneklem t-test sonucunda iki form arasında istatistiksel olarak anlamlı bir fark olmadığı bulunmuştur ($t(617) = 0.27$, $p = 0.79$). Ancak t-testi formlardaki tüm maddeler için uygulandığında, zorluklar açısından formlar arasında istatistiksel olarak anlamlı bir fark olduğu bulunmuştur ($t(617) = 9.53$, $p = 0.00$). Birleşik yanıt gerektiren çoktan seçmeli maddelerin içerdiği formun negatif köke sahip maddeleri içeren forma göre daha zor olduğu görülmüştür. Yine madde parametre tahminleri için uygulanan Rasch modeli bu sonucu da desteklemiştir. Diğer bir ifade ile negatif köklü maddelerin madde zorluk ortalaması -.93 bulunurken, birleşik yanıt gerektiren maddelerin ortalaması ise .04 bulunmuştur. Bu sonuçlar önceki yapılan çalışmaların bulgularıyla da tutarlılık göstermektedir (Atalmis, 2016; Nnodim, 1992; Tripp ve Tollefson, 1985).

Son olarak maddelerde cinsiyete göre DIF olup olmadığını araştırmak için öncelikle lojistik regresyon modeline dayanan Wald testi ve madde tepki kuramına dayalı Lord'un ki-kare DIF yöntemi olmak üzere iki farklı yöntem kullanılmıştır. İlk yönteme göre birleşik yanıt gerektiren bir madde olan madde 30'da erkek öğrenciler lehine bir başarı artmasının olduğu görülmüştür. İkinci yöntemde ise birleşik yanıt gerektiren maddelerden madde 30 ve 34'de yine erkek öğrencilerin lehine DIF ortaya çıkmıştır.

Sonuç ve Öneriler

Negatif köklü ve birleşik yanıt gerektiren çoktan seçmeli sorular en yaygın olarak kullanılan geleneksel olmayan madde formatlarıdır. Bunun nedeni ise bu maddeler için çeldiri oluşturmak geleneksel çoktan seçmeli maddelere göre nispeten kolaydır (Boland, Lester ve Williams, 2010). Ancak, bu iki madde formatı, çoktan seçmeli maddeler için madde-yazma ilkelerini ihlalleri nedeniyle güvenilirlik ve geçerlik sorunlarına neden olabilmekte bu durum bu formattaki maddelerin DIF'e sebep olabilme potansyelini artırmaktadır. Bu çalışmada özellikle DIF'in iki farklı madde formatı için cinsiyet farklılıklarına etkisi araştırılmıştır. Elde edilen sonuçlara göre negatif kök kullanımı madde zorluk derecesini etkilemezken, birleşik yanıt gerektiren çoktan seçmeli soruların daha zor olduğu görülmüştür. Ayrıca, birleşik yanıt gerektiren sadece bir çoktan seçmeli madde için erkek öğrencilerin eğilimi nedeniyle DIF etkisi tespit edildiği bulunurken, negatif köklü çoktan seçmeli maddelerde ise cinsiyete göre farklılık göstermediği bulunmuştur.

Bu çalışmanın bulguları, madde-yazma ilkelerinin ihlal edilmesine odaklanan ve giderek büyüyen araştırmalara katkıda bulunabilmektedir. Özellikle bu çalışmanın bulguları önceki çalışmalardan elde edilen bulguları ve madde-yazım kurallarının ihlal edilmesinin cinsiyet farklılıklarına dayalı bir DIF kaynağı olabileceğini ampirik olarak desteklenmektedir. Bununla birlikte, madde-yazım ilkeleri ihlalleri ile ilgili potansiyel DIF kaynağını tam olarak anlamak için daha fazla araştırma yapılması gerekmektedir. Özellikle daha fazla genellenebilir sonuçlar elde etmek için yüksek katılımlı testler için farklı madde formatlarının

kullanılmasının etkisini karşılaştırmak için büyük bir çalışma tasarlanmalı ve sadece DIF tespitinde cinsiyet farklılıkları değil, etnik köken veya sosyal sınıf gibi diğer demografik özellikler dikkate alınmalıdır.

REFERENCES

Atalmis, E. H. (2016). Do the guideline violations influence test difficulty of high-stake test? An investigation on university entrance examination in Turkey. *Journal of Education and Training Studies, 4*(10), 1-7. http://dx.doi.org/10.11114/jets.v4i10.1738

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*(1), 23-35. https://doi.org/10.1111/j.1745-3984.1991.tb00341.x

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry, 34*(4), 310-316.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*(2), 165-174. http://www.jstor.org/stable/1434975

Caldwell, J. S. (2008). *Comprehension assessment: A classroom guide.* New York, NY: The Guildford Pub.

Cohen, A. S., & Wollack, J. A. (2004). *Handbook on test development: Helpful tips for creating reliable and valid classroom tests.* University of Wisconsin-Madison, WI, USA, 2004. https://www.researchgate.net/profile/Allan_Cohen2/publication/248808614_Handbook_on_Test_Development_Helpful_Tips_for_Creating_Reliable_and_Valid_Classroom_Tests/links/5632378d08aefa44c367cea8.pdf

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics, 26*(2), 543-551. https://doi.org/10.1148/rg.262055145

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. https://doi.org/10.1027/1015-5759.14.3.197

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77. https://doi.org/10.1207/s15324818ame1301_3

Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International, 49*(2), 111-121. https://doi.org/10.1080/14703297.2012.677596

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. https://doi.org/10.1007/s10459-004-4019-5

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363-373. http://dx.doi.org/10.1037/0021-9010.68.3.363

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. https://doi.org/10.1016/j.tate.2005.01.008

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

Haladyna, T. M., & Downing, S. M. (l989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51-78. https://doi.org/10.1207/s15324818ame0201_4

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. https://doi.org/10.1177/0013164493053004013

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

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

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science

test. *Applied Measurement in Education, 12*(3), 211-235. https://doi.org/10.1207/S15324818AME1203_1

Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests (CSE Tech. Rep. No. 483).* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing test banks. [Super about MCQ]. *The Journal of Education for Business, 73*(2), 94-97. https://doi.org/10.1080/08832329709601623

Harasym, P. H., Doran, M. L., Brant, R., & Lorscheider, F. L. (1993). Negation in stems of single-response multiple-choice items: An overestimation of student ability. *Evaluation & the Health Professions, 16*(3), 342-357. https://doi.org/10.1177/016327879301600307

Harter, C. L., & Harter, J. F. R. (2004). Teaching with technology: Does access to computer technology increase student achievement? *Eastern Economic Journal, 30*(4), 505-514. https://www.jstor.org/stable/40326144

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty. Research Report RR-85-4.* Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale NJ: Erlbaum.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349. https://doi.org/10.1207/S15324818AME1404_2

Klein, S. P., Jovanovic, J., Stecher, B. M., McCaffrey, D., Shavelson, R. J., Haertel, E., Solano-Flores, G., &

Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation & Policy Analysis, 19*(2), 83-97. https://doi.org/10.3102/01623737019002083

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*(2), 164-184. https://doi.org/10.1080/08957340902754635

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862. https://doi.org/10.3758/BRM.42.3.847

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40*(1), 25-32. https://doi.org/10.3928/0148-4834-20010101-07

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement Examinations. College Board Report No.* 92-7. New York, NY: College Entrance Examination Board. https://doi.org/10.1002/j.2333-8504.1993.tb01516.x

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

Miles, M. B. ve Huberman, A. M. (1994). *Qualitative data analysis (*2nd ed.). Thousand Oaks, CA: Sage.

Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*(4), 388-394. https://doi.org/10.7334/psicothema2015.110

Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education, 31*(1), 53-64. https://doi.org/10.1080/03098770601167922

Nnodim, J. O. (1992). Multiple-choice testing in anatomy. *Medical Education, 26*(4), 301-309. https://doi.org/10.1111/j.1365-2923.1992.tb00173.x

Parker, C., & Somers, J. (1983, December). *A comparison of the difficulty and reliability of type K and best response test items.* Paper presented at the Iowa Evaluation and Research Association Conference, Des Moines, IA.

Pate, A., & Caldwell, D. J. (2014). Effects of multiple-choice item-writing guideline utilization item and student performance. *Currents in Pharmacy Teaching and Learning, 6*(1), 130-134. https://doi.org/10.1016/j.cptl.2013.09.003

Rizopoulos, D. (2006). Ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software, 17*(5), 1-25. https://core.ac.uk/download/pdf/6305163.pdf

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. https://doi.org/10.1191/0265532206lt319oa

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tamir, P. (1993). Positive and negative multiple choice items: How difficult are they? *Studies in Educational Evaluation,* 19(3), 311-332. https://eric.ed.gov/?id=EJ471898

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*(8), 354-363. https://doi.org/10.1016/j.nedt.2006.07.006

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*(2), 198-206. https://doi.org/10.1111/j.1365-2923.2007.02957.x

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), 40-47. https://doi.org/10.1186/1472-6920-9-40

Terzi, R., & Suh, Y. (2015). An odds ratio approach for detecting DDF under the nested logit modeling framework. *Journal of Educational Measurement*, *52*(4), 376–398. https://doi.org/10.1111/jedm.12091

Terzi, R., & Yakar, L. (2018). Differential item and differential distractor functioning analyses on Turkish high school entrance exam. *Journal of Measurement and Evaluation in Education and Psychology*, *9*(2), 136-149. https://doi.org/10.21031/epod.368081

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Tripp, A., & Tollefson, N. (1985). Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options? *Journal of Nursing Education, 24*(3), 92-98. https://doi.org/10.3928/0148-4834-19850301-04

Zenisky, A.L., Hambleton, R.K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1-2), 61-78. https://doi.org/10.1080/10627197.2004.9652959

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. https://s3.amazonaws.com/academia.edu.documents/33861736/handbook4__pdf_dif.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1535487960&Signature=lTCvN%2BycU1dLORbreXBM0vYO89w%3D&response-content-disposition=inline%3B%20filename%3DHandbook-4_pdf_dif.pdf

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF.* Prince

George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.