International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 314–324

DOI: 10.21449/ijate.421167

Published at http://www.ijate.net http://dergipark.gov.tr/ijate Research Article

# The Use of Three-Option Multiple Choice Items for Classroom Assessment

**Erkan Hasan Atalmış** [1*]

[1]Kahraman Sutcu Imam University, Faculty of Education, Department of Educational Measurement and Evaluation, Kahramanmaras, Turkey

**Abstract:** Although multiple-choice items (MCIs) are widely used for classroom assessment, designing MCIs with sufficient number of plausible distracters is very challenging for teachers. In this regard, previous empirical studies reveal that using three-option MCIs provides various advantages when compared to four-option MCIs due to less preparation and administration time. This study examines how different elimination methods; namely, the least selected and the random methods, influence item difficulty, item discrimination and test reliability on decreasing the number of options in MCIs from four to three. The research findings have revealed that the concerning methods did not affect item difficulty, item discrimination, and test reliability negatively. Results are discussed in relation to promoting quality classroom assessment.

**ARTICLE HISTORY**

*Received: 01 January 2018*
*Revised: 23 April 2018*
*Accepted: 30 April 2018*

**KEYWORDS**

Classroom Assessment,
Item-Writing Guidelines,
Number of Options,
Multiple-Choice Items,
Test Quality

## 1. INTRODUCTION

Classroom assessment is an indispensable period of education and training. To what extent the goals and behaviors that students need to gain during the semester has been determined and how much teachers teach what they think they are teaching has been presented through classroom assessment. Therefore, it is of high importance for teachers to carry out an effective in-class assessment, and teachers are required to spend a significant part of their professional work life in classroom assessment studies (Darling-Hammond & Youngs, 2002; Stiggins, 1991). Upon examining the related literature, the significance of in-class assessment was revealed and various recommendations were presented in this context. Among these recommendations are that paper-pencil tests which are the mostly used method of classroom assessment should be prepared by the teachers themselves (Frey & Schmitt, 2010). This allows the assessment tool be consistent and compatible with the class activities as the measurement tool.

Multiple-choice items (MCIs) are one of the most commonly used item type in classroom assessment (Haladyna & Rodriguez, 2013). When previous studies were analyzed, both theoretical and empirical studies regarding reliability and validity of these item types were

CONTACT: Erkan Hasan Atalmış ✉ erkanatalmis@gmail.com Kahraman Sutcu Imam University, Faculty of Education, Department of Educational Measurement and Evaluation, Kahramanmaras, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

conducted and these were determined to be more reliable and valid than particularly open-ended items (Collins, 2006; Tarrant, Knierim, Hayes, & Ware, 2006; Thorndike, 2005). However, the studies emphasized the challenges of preparing the appropriate number of rational choices for MCIs, so they developed alternative ways related to MCIs.

One of these alternative methods has been considered as a reduction of the number of options. Although various studies revealed that reducing the number of options from 4 to 3 does not have a negative effect upon test reliability and item discrimination (Atalmis & Kingston, 2017; Delgado & Prieto, 1998), no consensus has been reached so far on the comparison of three-option and four-option items in terms of item difficulty. That is, it could not be exclusively argued that one type is more difficult than the other in all circumstances. Even though Rodriguez (2005) suggests that the number of options in MCIs may result from different methods used to reduce the number of options from 4 to 3, this is not revealed empirically.

In this regard, whether different methods used in reducing the number of options from 4 to 3 has an impact upon test reliability, item discrimination and item difficulty will be empirically examined and thus the use of 3 option items in the classroom assessment is thought to provide a new path.

### 1.1. Classroom assessment activities (Assessment Criteria)

The quality of classroom activities was discussed by educators and researchers as classroom assessment activities play a significant role in improving the outputs of the training. In this sense, researchers emphasized that classroom assessment activities should aim at increasing the quality of learning in the classroom, rather than largely through the traditional sense of passing and failing the exams (Chappuis & Stiggings, 2002; Leahy, Lyon, Thompson, & Wiliam, 2005). Hence, classroom assessment must have the ability to answer questions such as how well learners are learning and how effectively teachers teach (Angelo & Cross, 2001). The most important way to achieve this is to use classroom assessment methods that provide accurate and descriptive feedback to students and teachers about learning and teaching activities in the classroom. This is only possible with reliable, valid and useful measuring tools.

Reliability is defined as the accuracy or precision of measurement procedure and so it is the degree to which measurement are free from error (AERA, APA, & NCME, 2014; Thorndike, 2005). Errors can arise either from the measurement tool, the measured characteristic, and the person who measure or from the environment. In this context, test reliability is negatively influenced by such factors as incorrectly responded questions whose answers are known to the students, involvement of guessing factor, subjective evaluation of teachers, testing environment, and cheating. Thus, the fact that tests used in the classroom are mostly composed of more questions, objectively scored and sensitive in selecting the test environment will increase the test reliability.

Validity is the test quality that indicates the degree to which a measuring instrument measures the desired property (AERA, APA, & NCME, 2014; Haladyna & Rodriguez, 2013). Hence, the validity of a measurement tool is measured through different features, such as content-related validity, construct validity and criterion-related validity. Content-related validity is about how much the test covers the features desired to measure (Thorndike, 2005). To illustrate, the extent to which a test prepared in a mathematics class covers the acquisitions of the unit that is to be measured relates to content-related validity. In this respect, more question-based testing also increases content-related validity just as test reliability. The construct validity refers to the fact that the construct to be measured is measured without any other mixing (Messick, 1989). For instance, if a test would only measure students' mathematical skills, this test would violate the validity of the test when it involves such skills that include mathematics questions including reading and attention skills. Criterion-related validity is the

relationship between a test and another test (Thorndike, 2005). To exemplify, if the paper-pencil test to measure students' math skills helps solve the mathematical problems experienced by the person in real life, then the paper-pencil test's criterion-referenced validity might be strong.

Along with reliability and validity, another feature of the measurement tool is practicality which is defined as the economical and easy development, application and scoring of a test (Thorndike, 2005). For example, in a 20-person class, when a test type is evaluated for each student's lesson time, paper-pencil tests seem to be more useful than performance tests in both applying and scoring process.

### 1.2. Multiple-Choice Items

Multiple-choice items (MCIs) are widely used paper-pencil test to construct objective tests, which are considered as quickly and unambiguously scored, and minimize test administration and scoring time (Haladyna, Downing, & Rodriguez, 2002; Haladyna & Rodriguez, 2013). However, constructing these items consisting of plausible distractors and measuring desired objectives is challenging for item writers (Collins, 2006; Haladyna et al., 2002). Item preparation is also called as "a creative art". (Rodriguez, 1997). In this regard, item-writing guidelines which are supposed to help design items systematically with the aim of increasing validity evidence for the test have been reported in previous studies. Validity evidence is a scientific notion used to describe how to develop tests accurately and how to predict, evaluate, and interpret the test scores (AERA, APA, & NCME, 2014).

To date, a limited number of studies focused on item-writing guidelines for item and test construction. One of the pioneering ones in this concern was conducted by Haladyna and Downing (1989) who proposed 43 item-writing guidelines after reviewing textbooks published in the field of measurement and evaluation. Approximately two decades later, Haladyna et al. (2002) 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. Several years later, Frey, Petersen, Edwards, Pedrotti, and Peyton (2005) evaluated twenty classroom assessment textbooks and identified 40 most commonly used item writing guidelines. They also classified them depending on validity concerns, (i.e., potentially confusing wording or ambiguous requirements, guessing, rules addressing test-taking efficiency, and rules designed to control for test wiseness). Moreno, Martínez, and Muñiz (2006) designed a condensed version of the existing item writing guidelines and classified them into three groups with a focus on foundations, the expression of the domain and context in each item and test, and on response options. Likewise, the same researchers decreased the number of the guidelines to 9 and evaluated them according to the definition of validity (Moreno, Martinez, & Muniz, 2014). One of the common characteristics of item writing guidelines proposed by previous studies was that each study emphasized the construction of a sufficient number of options with plausible distractors for each item since one of the challenging part of the item-writing process of MCIs is to construct a sufficient number of plausible distracters (Haladyna et al., 2002). Plausible distractors are developed based on students' particular errors at some point in analyzing and solving the problem (Thorndike, 2005). Thus, MCI writers should have deep pedagogical content knowledge and teaching experience. This results in decreasing the probability of using more options in MCIs, such as the use of three options rather than four options.

### 1.3. Comparing Four Option with Three Option MCIs

Extant empirical studies concluded that using three-option MCIs provides various advantages compared to four-option MCIs due to less preparation and administration time (Balta & Eryılmaz, 2017; Haladyna & Downing, 1989; Haladyna et al., 2002; Rich & Johanson, 1990). Empirical studies have examined how item (test) difficulty, item discrimination and test

reliability vary across 4-choice items and 3 choice items. Item difficulty is defined as the proportion of students who choose the correct answer while item discrimination is defined as how well the item differentiates students with high ability in the construct of interest from students with low ability. Test reliability, as mentioned in the introduction section, is defined as the consistency of test results.

The studies have found opposite results regarding item difficulty. Some studies found that item difficulty was not statistically different between four-option items and three-option items (Abad, Olea & Ponsoda, 2001; Atalmis & Kingston, 2017; Baghei & Amrahi, 2011; Shizuka, Takeuchi, Yashima & Yoshizawa, 2006), whereas others concluded that MCIs with three options were statistically more difficult than MCIs with four options, which is counterintuitive (Landrum, Cashin, & Theis, 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 MCIs. Of these 48 studies related to achievement and aptitude tests, 27 studies included pertinent results. The results supported that three-option items were slightly easier than four-option items.

Considering studies on item discrimination, item discrimination between MCIs with four options and items with three options was not statistically different in most studies (Atalmis & Kingston, 2017; Cizek & O'Day, 1994; Crehan, Haladyna, & Brewer, 1993; Dehnad, Nasser, & Hosseini, 2014; Delgado & Prieto, 1998; Rogers & Harley, 1999; Shizuka et al., 2006; Tarrant & Ware, 2010). Yet, some studies provided statistically significant evidence that item discrimination for MCIs with three options was higher than that of MCIs with four options (Baghei & Amrahi, 2011; Landrum et al., 1993; Rodriguez, 2005; Trevisan, Sax, & Michael, 1991). Consequently, the literature reveals that three-option items do not affect negatively item discrimination.

A limited number of studies on test reliability have indicated that the number of options did not have a statistically significant impact on test reliability (Atalmis & Kingston, 2017; Baghei & Amrahi, 2011; Delgado & Prieto, 1998; Rogers & Harley, 1999) while some found that test reliability increased when the forms with three options were employed (Rodriquez, 2005; Tarrant & Ware, 2010).

### 1.4. Significance of the Study and Research Questions

Although previous studies revealed that reducing number of options from 4 to 3 did not have a negative effect upon test reliability and item discrimination, no consensus has been reached so far on the comparison of three-option and four-option items in terms of item difficulty. Even though Rodriguez (2005) suggests that the number of options in MCIs may result from different methods used to reduce the number of options from 4 to 3, this has not been revealed empirically. Given previous studies, these used different traditional methods to eliminate one of four-option to construct three-option MCIs, such as eliminating the least selected option or a random option. However, they did not consistently investigate the impact of elimination method on item and test characteristics. Therefore, this research aims to examine how item and test psychometric characteristics vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs. More specifically, we have examined three research questions as follows:

- Does item difficulty for mathematics items vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs?
- Does test reliability for mathematics items vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs?

- Does item discrimination for mathematics items vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs?

## 2. METHOD

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

### 2.1. Data Collection

Data collection procedure includes several phases. First, test forms including parallel four-option MCIs were designed, and each form was administered to 7th grade students in state primary schools located in Turkey for the pilot study. After calculating item psychometrics characteristics of each item, 20 of them were selected to be used in the final version of the instrument. Subsequently, two forms (Form B1 and B2) were designed with parallel items on each, one of the options of MCIs in each form was eliminated by using the least selected option in Form B1 and random option in Form B2. After two forms were administered to 7th and 8th grade students in Turkey, data analysis was conducted.

### 2.2. Participants

This research was carried out with 7th and 8th grade students in the pilot study and the main study in Turkey. Convenience sampling method was applied for piloting and the ultimate phase. The pilot test was administered to 1130 students attending sixteen state primary schools in the province of Manisa, Turkey. The final test was administered to 847 students who enrolled in eight schools in the provinces of Manisa and Kahramanmaras, Turkey. In both phases, only students' responses to mathematics items were collected without their academic proficiency and demographic features.

### 2.3. Instrument Development

Instrument development procedure of this study includes several phases. First, we developed an item pool including a large number of items representing the full range of the objectives of the mathematics topic, "equation and expression". Hence, after 58 MC mathematics items were developed, two forms (Form A1 and Form A2) composing of 29 items on each were constructed so that the examinees could answer all of them during a class period. After Form A1 and Form A2 was respectively administered to 474 and 656 students attending sixteen public schools in the province of Manisa, Turkey, item difficulty and item discrimination were computed for each item in Form A1 and Form A2. Item difficulty is defined as the proportion of the students choosing correct answers while item discrimination shows how well the item discriminates between students with high ability and low ability (Thorndike, 2005). Item-total correlation index, one of most widely used method, was used to calculate item discrimination for each item (Downing, 2005). The findings showed that item difficulty indexes for the items range between .19 and .74, while item discrimination indexes range between .25 and .65.

Second, 20 out of 58 MCIs which were higher quality (higher discrimination index and middle item difficulty) were selected and designed to be used in the final version of the instrument. Namely, items with discrimination index of .30 or greater than .30 and three functioning distractors were selected (Field, 2009). Previous studies proposed that functioning distractors were plausible distractors chosen by at least 5% of examinees (Haladyna & Dowing, 1993; Rodriguez, Kettler, & Elliott, 2014).

Third, two forms (Form B1 and B2) were redesigned with 10 parallel items measuring the same specific learning standards and objectives on each form. Moreover, parallel items were

constructed with the same content and rationale of distractors, but the numbers were different. Due to a small number of items in each form (*i.e.*10 items in each group), nonparametric tests were preferred by using item difficulty and item discrimination values of items on each form to show the equality of Form B1 and Form B2. A Mann-Whithey U Test was conducted; accordingly, it was found that average rank of item difficulty did not statistically differ between Form B1 ($z$=-1.17, $p$=.24). To test two item discrimination index, item-total correlation was used and then Fisher-Z transformation method is carried out to make correlation values normalized, as follows:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad (1)$$

where r is item-total correlation for each item and z is the transformed value of r. The findings showed item discrimination did not statistically vary across Form B1 and B2 ($z$=-.19, $p$=.85).

Fourth, one of the options of each of the MCIs in each form was eliminated via the least selected option in Form B1 and random method in Form B2. Table 1 reveals distractor selection frequencies, proportion of students choosing options of each item, and eliminated distractors in Form B1 and Form B2.

**Table 1.** *Distractor Selection Frequencies in Form B1 and Form B2*

| Item ID | Form B1 | | | | Item ID | Form B2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | a | b | c | d | | a | b | c | d |
| 1 | .25 | .15 | .53* | .07** | 11 | .27 | .14** | .54* | .05 |
| 2 | .12 | .10** | .42 | .36* | 12 | .52* | .17 | .17 | .13** |
| 3 | .11** | .23 | .19 | .47* | 13 | .13** | .07 | .44 | .36* |
| 4 | .40* | .27 | .21 | .12** | 14 | .16 | .53* | .17 | .14** |
| 5 | .19 | .50* | .18 | .13** | 15 | .15 | .20** | .53* | .11 |
| 6 | .54* | .15 | .11** | .20 | 16 | .55* | .13 | .12 | .20** |
| 7 | .54* | .12** | .15 | .19 | 17 | .10 | .13 | .44** | .33* |
| 8 | .26 | .31* | .09** | .34 | 18 | .27** | .34* | .14 | .25 |
| 9 | .23 | .39* | .15** | .23 | 19 | .22** | .25 | .34* | .19 |
| 10 | .26 | .27 | .28* | .18** | 20 | .23 | .19** | .30* | .28 |

* key; **eliminated distractor

Finally, the test consisting of all items was designed to counterbalance the order of MCIs in order to avoid systematic errors, entailing that some students were supposed to take the test with Form B1 followed by Form B2, while others take the tests in concern in the reverse order.

### 2.4. Data Analysis

In order to examine whether different elimination methods have a significant impact on psychometrics characteristics, item difficulty and discrimination were calculated for each item, and test reliability of forms (Form B1 and Form B2) was measured individually and together by using item response theory (IRT). Regarding item difficulty and item discrimination, two parameter logistic (2PL) model and three-parameter logistic (3PL) model data fit statistics were calculated for 20 items by employing IRTPRO 4.2. After likelihood ratio test was conducted to compare the models, it was found that the values of the -2 log likelihoods (i.e., -2ln$L$) for the 2 PL and 3 PL models were 19577.35 and 19306.99, respectively. The difference between these values was found as 270.36 with 19 degree of freedom, which is statistically significant.

Therefore, using 3PL model represents a statistically significant improvement in fit over the 2PL model.

Item difficulty, *b* parameter, shows the position of the Item Curve Characteristics (ICC) regarding the ability scale and item difficulty index ranges between -3 and +3. Easy items are located somewhat below 0 while difficult items are located somewhat above 0 (De Ayala, 2013). Item discrimination, *a* parameter, reveals the proportion to the slope of the ICC at the *b* point on the ability scale (Hambleton, Swaminathan, & Rogers, 1991). Items with higher *a* values are demanded because these items discriminate well among examinees. In this study, after item difficulty and item discrimination values of items on each form were calculated, nonparametric tests were conducted to compare Form B1 and Form B2.

Test reliability is defined as internal consistency of the test and it can be calculated using coefficient alpha (α) estimation method in Classical Test Theory (CTT). Subsequent to calculating the standard error of estimate (SEE) for each reliability coefficient value to obtain the 95% confidence interval (Duhachek & Iacobucci, 2004; Van Zyl, Neudecker, & Nel, 2000), it is examined whether the reliability coefficient is statistically different from one sub-test to another. However, calculating test reliability depends on the particular set of items in CTT. This is a disadvantage of CTT over IRT since each item contribute test reliability individually and independently. Therefore, test reliability is also calculated using IRT as well as CTT in this study.

In summary, item difficulty, item discrimination, and test reliability are important criteria to calculate psychometrics characteristics. The following section shows the results of item difficulty and item discrimination indexes for each item, and test reliability indexes for each form.

## 3. FINDINGS

This section provides the results of item difficulty and item discrimination, and test reliability, respectively.

### 3.1. Item Difficulty and Item Discrimination

Table 2 reveals item difficulty and item discrimination indexes for each item on Form 1 and Form 2. Item difficulty index of items ranges from -.22 to 1.14 in Form B1 with the median value of .43 while item difficulty index of items varies from -.86 and .77 in Form B2 with the median of .21. Item discrimination index of items ranges between 1.43 and 7.73 in Form B1 while these ranges are observed with 1.20 and 7.99 values in Form B2. The median of Form B1 and Form B2 are 2.21 and 2.26, respectively.

**Table 2.** *Item Difficulty and Discrimination Indexes across Forms*

| Item ID | Form B1 (The Least Selected Method) | | Item ID | Form B2 (Random Method) | |
|---|---|---|---|---|---|
| | *Item dif. (b)* | *Item disc. (a)* | | *Item dif. (b)* | *Item disc. (a)* |
| 1 | 0.63 | 7.73 | 11 | -0.21 | 3.21 |
| 2 | -0.22 | 3.06 | 12 | 0.52 | 1.92 |
| 3 | 0.82 | 2.98 | 13 | -0.79 | 1.54 |
| 4 | 0.83 | 2.14 | 14 | -0.86 | 1.20 |
| 5 | 0.19 | 2.22 | 15 | -0.10 | 1.71 |
| 6 | 0.22 | 1.43 | 16 | 0.63 | 2.59 |
| 7 | -0.21 | 3.27 | 17 | 0.66 | 7.99 |
| 8 | 0.05 | 1.84 | 18 | 0.76 | 1.37 |
| 9 | 1.14 | 1.92 | 19 | 0.77 | 3.23 |
| 10 | 0.78 | 2.20 | 20 | -0.26 | 4.19 |

To show the equality of item difficulty and item discrimination of Form B1 and Form B2, a nonparametric test; Mann-Whithey U Test, was run and it was found that average rank of item difficulty and item discrimination did not statistically vary across Forms B1 and B2 (item difficulty: $z$ = -1.36, $p$ = .17; item discrimination: $z$ = -.34, $p$ = .73).

### 3.2. Test Reliability

The reliability coefficient for each form was calculated through coefficient alpha estimation method in CTT. All forms had good internal consistency values ($\alpha_{\text{whole test}}$ = .84; $\alpha_{\text{form B1}}$ = .72; $\alpha_{\text{form B2}}$ = .73), which were greater than .70 (Thorndike, 2005). The 95% confidence interval for each alpha yielded great similarity in Form B1 and Form B2, which do not statistically differ from each other since their intervals are overlapped ( Form $B1_{(.70, .74)}$; Form $B2_{(.71,.75)}$). When test marginal reliability is calculated using IRT 3PL model, the findings showed .70 for Form B1 and Form B2 and .84 for the whole test, which indicated similar results to that of CTT's. It allows us to infer that applying different elimination methods to reduce four-options to three-options does not statistically influence the reliability of the test.

In summary, the findings showed that different elimination methods; the least selected method and the random method, did not affect item difficulty, item discrimination, and test reliability negatively.

## 4. CONCLUSION AND DISCUSSION

This study aimed to examine how psychometric properties of items and test vary when different elimination methods were used to reduce the number of options of MCIs by applying the least selected method and random method. Research results showed that item difficulty, item difficulty, and test reliability did not statistically differ across the elimination methods administered to the items.

The results of this study could contribute to the growing body of research focusing on the impact of the number of options on psychometric characteristics. Overall, earlier research on item and test characteristics generally put great emphasis on comparing three-option MCIs with four-option MCIs rather than option elimination in these items. Namely, most of the empirical studies usually employed a particular traditional elimination method to reduce number of options of MCIs (i.e. least frequently chosen option, least discriminating option, and a random option). The findings of the present study are consistent with those in the some of the previous studies which showed that three-option MCIs perform equally well as four-option MCIs in terms of item discrimination and test reliability regardless of elimination methods (Atalmis & Kingston, 2017; Baghei & Amrahi, 2011; Delgado & Prieto, 1998; Rogers & Harley, 1994; Sidick, Barett, & Doverspike, 1994; Tarrant & Ware, 2010). Prior research reported contradictory results for item difficulty; for instance, three-option MCIs have been determined to be more difficult than four-option MCIs in Crehan et al. (1993) whereas Rodriguez (2005) also found a small change in item difficulty (.04) between four-option MCIs and three-option MCIs. However, this change was found to be statistically significant, meaning that four-option makes items more difficult. Atalmis & Kingston (2017), Baghei & Amrahi (2011), Delgado & Prieto (1998), Shizuka et al. (2006) and Tarrant & Ware (2010) found both types of MCIs were found to be equally difficult. On the other hand, none of the previous studies examined the impact of elimination method on psychometric characteristics of the mathematics items and/or tests. Keeping this in mind, the current study addressed how two elimination methods differentially affect psychometric characteristics in concern.

The results of this study could not only provide empirical support for test development studies but also make several recommendations to the test designers and classroom teachers. The use of different elimination methods does not significantly influence item difficulty, item

discrimination, and reliability. In other words, psychometric characteristics did not vary when the least selected option and random option were deleted. Therefore, reducing the number of options of four-option MCIs to three options increases the efficiency of item-writing and administering test regardless of elimination methods. For instance, more three-option MCIs could be constructed in relatively shorter period of time as opposed to four-option MCIs (Aamodt & McSahne, 1992) since construction of a rationale distractor is widely considered as time consuming and one of the most challenging part of item writing (Haladyna et al., 2002). Besides, administering a test including three-option MCIs is expected to increase test reliability in certain ways as opposed to that composed of four-option MCIs. First, administering a test composed of three-option MCIs takes less time than that including four-option MCIs, and shorter tests are likely to decrease students' fatigue and test anxiety, which increases the reliability of the test. Second, the same amount of time is allotted to the implementation of the test including three-option MCIs and to that including four-options MCIs, the former is expected to contain a relatively higher number of items than the latter, which increases the test reliability.

Despite reporting findings on the use of elimination methods which are administered to MCIs, this study has certain limitations. Data in this study were obtained from the seventh and eighth grade students. So, different findings could be driven when the test is administered to the students attending lower or higher grades. For instance, relatively different findings are expected to be obtained when it is applied to those attending higher grades since they are considered to be more test-wise. It is also confined to the content of this test; namely, the items were constructed only in the scope of mathematics. So, tests prepared in different disciplines are expected to yield different results in terms of psychometric characteristics of MCIs and the test. The other limitation of this study is that convenience sampling method was applied for piloting and the ultimate phase from only two cities in Turkey. Although applying this sampling method is plausible as it is fast, inexpensive and easy, it might be implausible to generalize the findings for entire population. In addition, the items used in this study were at the middle difficulty level due to the fact that their item difficulty indexes ranged generally from -1 to +1. This could limit us to discriminate students in the upper and lower levels. The final limitation is that there are only 10 items used per test form. Although the reliability coefficient for each test composed of 10 items was found good internal consistency values in existing study, a small sample for statistical tests limits the generalizability.

**Acknowledgments:**

An earlier version of this paper was presented at the National Council on Measurement in Education in Chicago, US in 2015. The researcher would like to thank Dr. Marianne Perie, the director at Center for Educational Testing and Evaluation at Kansas, and Dr. Neal Kingston, the director of the Achievement and Assessment Institute at University of Kansas in the US, for their invaluable encouragement that made it possible to conduct this study.

## ORCID

Erkan Hasan Atalmış https://orcid.org/0000-0001-9610-491X

## 5. REFERENCES

Aamodt, M. G., & McShane, T. D. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management, 21*(2), 151–160.

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

AERA, APA, & NCME (2014). *Standards for educational and psychological tests*. Washington DC: American Psychological Association, American Educational Research Association, National Council on Measurement in Education.

Angelo, T. A., & Cross, K. P. (1993). Classroom assessment techniques: A handbook for college teachers. *San Francisco: Jossey-Bass*.

Atalmis, E. H., & Kingston, N. M. (2017). Three, four, and none of the above options in multiple-choice items. *Turkish Journal of Education*, *6*(4), 143-157.

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.

Balta, N., &Eryılmaz, A. (2017). Counterintuitive dynamics test. *International Journal of Science and Mathematics Education*, *15*(3), 411-431.

Chappuis, S., & Stiggins, R. J. (2002). Classroom assessment for learning. *Educational leadership*, *60*(1), 40-44.

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.

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions 63 for continuing medical education activities and self-assessment modules. *RadioGraphics, 26*(2), 543-551.

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.

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us?. *Educational researcher*, *31*(9), 13-25.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

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.

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.

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.

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.

Field, A. (2009). *Discovering statistics using SPSS*. London, England: Sage.

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.

Frey, B. B., & Schmitt, V. L. (2010). Teachers' classroom assessment practices. *Middle Grades Research Journal*, *5*(3), 107-117.

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

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.

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

Hambleton, R.K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of item response theory.* Beverly Hills, CA: Sage.

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.

Leahy, S., Lyon, C, Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership, 63*(3), 18-24.

Messick, S. (1989).Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology*, *2*(2), 65-72.

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.

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.

Rodriguez, M. C. (1997). The art & science of item writing: A meta-analysis of multiple choice item format effects. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13.

Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *Sage Open*, *4*(4), 1-10.

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

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.

Sidick, J.T., Barrett, G.V., & Doverspike, D. (1994). Three-alternative multiple choice14 tests: An attractive option. *Personnel Psychology, 47*(4), 829-835.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*(1), 7–12.

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 in Practice*, *6*(6), 354-363.

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.

Thordike, R.M. (2005). *Measurement and Evaluation in Psychology and* Education (7th Ed.). Upper Saddle River, NJ: Pearson Education.

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.

Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*(1), 271–280.