Psychometric Properties of 3-, 4-, and 5-Option Item Tests: Do Test Takers' Personality Traits Make a Difference?

Fatemeh Khaleghi, M.A, Imam Khomeini International University, Qazvin, Iran, Email: fkhaleqi@yahoo.com

Rajab Esfandiari, Corresponding author, Assistant Professor, Imam Khomeini International University, Qazvin, Iran, Email: esfandiari@hum.ikiu.ac.ir

Abstract

Prior research has yielded mixed results regarding what contributes psychometrically sound multiple-choice (MC) items. The purpose of the present study was, therefore, twofold: (a) to compare 3-, 4-, and 5-option multiple-choice (MC) tests in terms of psychometric characteristics, and (b) to investigate the relationships between three MC tests and five personality traits. To that end, 150 students were asked to answer three stem equivalent MC item tests. A Big Five Inventory was used to find students' personality traits. Moreover, an attitude questionnaire was utilized to seek students' opinions of these three MC tests. The results of one-way repeated measures ANOVA revealed statistically significant differences for item difficulty, while no statistically significant differences were found for item discrimination and reliability across three MC tests. The results of the Pearson correlation showed no correlation between personality traits and three different versions of MC tests. The results of the attitude questionnaire indicated mixed views towards MC tests. The findings of this study suggest that test developers consider statistical, affective, and contextual factors in order to develop different formats of MC tests.

1. Introduction

Multiple-choice (MC) tests, as one of the most important assessment tools, are widely used in educational settings to measure students' language abilities. Such multiple-choice formats are commonly used in "elementary and high-school achievement tests, college and graduate-school admission tests, personnel tests, and certification, licensing, competency, and proficiency examination in education and other professions" (Haladyna & Downing, 1989, p.38). In fact, vital decisions are made based on MC test scores. The main reason for the widespread use of such tests is that they are more practical (Ng & Chan, 2009). Furthermore, they cover a large number of items, are scored rapidly and objectively, save time and cost, and can be fair and standardized (Al-Rukaban, 2006; Lee & Winke, 2012; Rogers & Harley, 1999). Therefore, MC tests are employed to achieve different purposes across the globe.

MC tests have a variety of formats, of which the traditional MC test is the most commonly used test format (Ng & Chan, 2009). This type of test consists of a stem, including three, four, or five options. One of these options is referred to as the key or correct answer, and the others are distractors (Farhady, Ja'farpur, & Birjandi, 1994; Fulcher, 2010; Haladyna, 2004; Osterlind, 2002). The optimal number of options in MC tests has been a challenging issue in language testing, with no agreement among researchers (Nwadinigwe & Naibi, 2013; Rodriguez, 2005). This debate of the optimal number of options in MC tests has continued for more than 80 years (Rodriguez, 2005). Some researchers advocate four or five options (Hodson, 1984; Ramos & Stern, 1973); others have found that presenting many options can be as plausible as 4-, or 5-option items; there are also some recommendations for 3-option MC tests (Haladyna & Downing, 1993).

An MC test may be affected by some factors such as distractors, number of options, and difficulty, and item discrimination, which may lead to the

improvement of the quality of items. In addition to these factors, the personality traits of language learners may also affect the psychometric properties of an MC test. However, reviewing the existing literature, to date, few studies, if any, have focused on the possible relationships between personality traits and the properties of MC tests. Accordingly, the present study seeks to compare 3-, 4-, and 5-option MC items in terms of item facility, discrimination, and reliability. It is also intended to explore the relationships between personality traits and students' performance in MC tests. Additionally, this study aims to find students' preferences in terms of 3-, 4-, and 5-option MC tests.

2. Literature Review

2.2. Theoretical Background

2.2.1. MC Tests

MC tests are frequently used in education in general and in applied linguistics in particular. Regardless of the purpose of the test, there are various types of a test, but MC tests are one of the extensively used formats in the world. An MC test is the best manifestation of objective tests (Heaton, 1988; Razavipour, 2014). Otis (1910, as cited in Landrum, Cashin & Theis, 1993) developed early MC tests.

As claimed by Osterlind (2002), the reason for the widespread use of an MC test its diverse advantages that lead to appropriate decision making; one of the advantages is that MC tests can be used for various contents, abilities, and skills. Indeed, MC tests are applicable to a diversity of academic goals and contents in a short amount of time (Heaton, 1990). Another advantage is the objectivity of an MC test, which can be scored by anyone (Downing & Haladyna, 2006; Nejati & Moradi, 2015). The other strength of MC tests is that test takers cannot conceal their restricted knowledge of content (Osterlind, 2002).

Although MC tests are still very commonly used in high-stakes tests such as Test of English for International Communication (TOEIC) for admission purposes, the construction of MC items brings challenges. Over the years, researchers have presented guidelines about how to write the most effective MC items (Haladyna, & Downing, 1993; Haladyna, Downing, & Rodriguez, 2002; Hansen & Dexter 1997), but such guidelines do not necessarily guarantee a successfully written item. One of the most serious factors which have been shown to affect the quality of an MC item refers to the number of options an MC may include. In the following section, more information on this topic is given.

2.2.2. Number of Options in MC Tests

 The optimal number of options has long been debated; there is no agreement about the desirable number of options. Some researchers have recommended the use of 3-option MC tests (Asmus, 2016; Currie & Chiramanee, 2010; Nwadinigwe & Naibi, 2013; Rodriguez, 2005; Schneid et al., 2014; Shizuka et al., 2006). As explained by Asmus (2016), 3-option MC tests are more efficient in eliciting examinee's knowledge in comparison to five-option MC tests; they also save time due to lack of constructing and reading two extra distractors per item. Additionally, the construction of the three options is less laborious than 4-option items (Crehan, Haladyna, & Brewer, 1993). In spite of supporting evidence from findings of numerous studies, five- and four-option multiple-choice tests are still used for high-stakes assessments and measurements due to being less susceptible to guessing and high reliability (Mehrens & Lehman, 1991; Thorndike & Thorndike-Christ, 2010). Indeed, correct guessing happens increasingly in 3-option MC items (Schneid et al., 2014).

Some other researchers advocate longer than 3-option items (Budescu & Nevo, 1985; Farhady & Shakery, 2000; Green, Sax, & Michael, 1982; Woodford & Bancraft, 2004). In standard measurement theory, it is believed that more

options in an MC test contribute to having a highly reliable test (Hopkins, 1998; Mehrens & Lehman, 1991; Thorndike & Thorndike-Christ, 2010). As mentioned by Haladyna and Downing (1993), more options enhance reliability when the distractors are functional; more options also reduce the chance of correct guessing compared with two- and three-option MC tests. Despite these advantages, Owen and Froman (1987) argued that an MC test with more options provides unwanted cues, which lead to test-wiseness Providing more options requires more time, effort, and cost. (Delgado & Prieto, 1998; Haladyna, 2004; Rodrigeuz, 2005; Shizuka et al., 2006).

"Determining the optimal number of options among three, four, or five is important because, all other things about a test being equal, the number of response alternatives will affect the reliability of a test" (Osterlind, 2002, p. 146). In addition to reliability, research has shown that the number of options also affects the difficulty and discrimination of an MC item. More information about the psychometric characteristics of items is given in the following section.

2.2.3. Psychometric Characteristics of Items

An effective language item should discriminate between low-ability and high-ability language learners and is neither too easy not too difficult. Item difficulty determines the efficacy of items (Bachman, 2004). It "reflects how difficult or easy an item is, and numerically expressed as the proportion of students who answer an item correctly" (Nwadinigwe & Niabi, 2013, p. 189). Item difficulty indices range from 0.00 to 1.00. The 0.00 value shows the test is very difficult, whereas 1.00 value shows that the test is very easy (Bachman, 2004; Brown, 2005). The ideal item difficulty index is 0.50. Still, the values of 0.37 to 0.63 are acceptable, so an item with the IF index below and beyond the determined values should be eliminated in norm-referend language tests (Farhady, Ja'farpur, & Birjandi, 1994).

Item discrimination (ID) is the degree to which an item discriminates between high- and low-ability test takers (Brown, 2005). As noted by Onunkwo (2002), item discrimination relies on item facility and plausibility of distractors. As a result, ID is calculated by subtracting the IF of the low-ability group from high-ability groups, so ID index ranges from +1.00 and -1.00; the former shows that all high-ability test takers answer an item correctly, but all low-ability students get it wrong while the latter shows all of the low-ability test takers answer an item correctly, but all high-ability students get it wrong. Brown (2005) added that the point biserial correlation can be used to estimate item discrimination. In comparison to IF, the ideal index for item discrimination is +1.00, and the values beyond 0.40 are acceptable.

Reliability is a fundamental quality of a test or any type of measurement or assessment, which can be affected by the number of item options. Richards and Schmidt (2002) defined reliability as "a numerical index of test reliability that can be obtained by correlating two sets of scores on parallel tests, repeated test administrations, or the two halves of a test. Theoretically, its values range between 0.0 and + 1.0" (p. 453). Reliability has been conceptualized differently according to theories, including classical test score theory, generalizability theory, and item response theory (Bachman, 1990). Although different methods have been proposed to estimate reliability, Cronbach's alpha is the most commonly reported method in the literature to estimate reliability for different data types.

Research has shown that the number of options the items include can influence the reliability of an MC test. Cizek, Robinson, and O' Day (1998) stated that reducing the non-functioning distractors led to enhanced score reliability. Similarly, Straton and Catts (1980) noted that a 3-alternative item test was superior to other MC tests in terms of reliability. However, Rogausch,

Hofer, and Krebs (2010) found that reliability decreased significantly in 3-option MC tests.

2.2.4. Personality Traits

Personality traits are an inescapable aspect of teaching and testing used as a performance predictor. Personality traits as individual differences received increasing attention in the Piagetian constructionist era, whereas little attention was given to it in the behaviorist and sociocultural era (Jensen, 2015). Various definitions of personality have been expressed; for example, Funder (2004, p. 5) defined personality as "an individual's characteristic patterns of thought, emotion, and behavior, together with the psychological mechanisms – hidden or not – behind those patterns." Pervin, Cervone, and John (2005) conceptualized personality in another way that focused on personal characteristics, which led to "consistent patterns of feelings, thinking, and behaving" (p. 6).

Various models and classifications of personality traits have emerged after extensive research (John, Hampson, & Goldberg, 1991; McAdams, 1995). Additionally, a review by John, Naumann, and Soto (2008) revealed that until 2006, the number of studies using the big five model (BFM) exceeded other models. Although Goldberg (1981) coined the term 'Big Five,' McCrae and Costa (1992) proposed that BFM comprised five broad personality traits, each of them, including six sub-traits or facets. In other words, each trait is composed of six separate facets (Di Blas & Carrao, 2011; McCrae & Costa, 1992, 2006; Mervielde, Buyst, & De Fruyt, 1995). The five personality traits formed the OCEAN acronym derived from Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Dawson, 2015; Dornyie & Ryan, 2015). The components of Big Five personality traits are described in Table 1.

Table 1. *Personality Traits and Facets*

Personality traits	Facets
Extroversion (E) Introversion	Talkative, a joiner, physically active, affectionate, passionate, fun-loving, Reserved, seeking solitude (a loner), physically passive, quiet, sober, unfeeling
Openness / Open to new experiences (O)	Imagination, creativity, originality, prefer variety, curiosity, liberal
Traditionalist	Down to Earth, uncreative, conventional, uncurious, prefer routines, conservative
Conscientiousness(C) Careless	Conscientious, hard-working, ambitious, well organized, persevering, punctual, Quitting, negligent, lazy, disorganized, aimless, late
Agreeableness (A) Self-centered	Softhearted, trusting, generous, acquiescent, lenient, good-natured Self-centered Suspicious, ruthless, stingy, antagonistic, critical, irritable
Neuroticism (N) Emotionally stable	Worrying, temperamental, self-pitying, self-conscious, emotional, vulnerable Calm, even-tempered, self-satisfied, comfortable, unemotional, hardy

Note. Adapted from "Personality Traits, Learning and Academic Achievements" by M. Jensen (2015). *Journal of Education and Learning, 4*(4), p. 92

Each of the above components of personality traits is elaborated in the following paragraphs. Eysenck (1959, as cited in Dornyie & Ryan, 2015) was the first researcher to identify characteristics of extroversion. Extroversion describes the tendency to interact with others and assert themselves. It is also described as being a positive, friendly, energetic, and adaptable person (Cherry, 2017).

Openness is the other personality trait of the BFI. Open individuals like novelty, innovation, adventure, new experience (Fraser-Thill, 2017). Openness is described as the tendency to be original, open-minded, and curious about the

universe (Grahek, 2007). As claimed by McCrae and Costa (1997, p. 826), openness is perceived "in both structural and motivational terms. Openness is seen in the breadth, depth, and permeability of consciousness, and in the recurrent need to enlarge and examine experience".

Conscientiousness refers to the tendency to control impulses, planning, organizing, and doing activities (Grahek, 2007). To put it simply, individuals with this trait behave, think, and feel in a constant manner over time in specific circumstances (Roberts et al., 2009). Moreover, conscientiousness as an individual difference has to do with obeying rules and being purposeful, firm, and goal-directed; conscientious people also intend to plan, think, and be careful in doing an action (John & Srivastava, 1999).

Agreeable people prefer to have social interaction and collaboration (Fraser-Thill, 2017). Grahek (2007) described agreeableness as a tendency to be helpful, sympathetic, patient, and indulgent. This dimension was used with different names such as "tender-mindedness, friendly compliance versus hostile noncompliance, likeability, communion, and even love versus hate" (Graziano & Tobin, 2002, p. 47). Agreeable people tend to be cooperative, altruistic, emotional, sympathetic, well-mannered, and good-natured (Barrick & Mount, 1991; Digman, 1990).

Among all the five dimensions of personality trait, only neuroticism is considered a negative trait (Dawson, 2015). The term Neurosis was coined by Cullen (1769, as cited in Widiger, 2009), who was a Scottish doctor and related it to general affection disorders of the nervous system. Costa and McCrae (1992) defined neuroticism as "the general tendency to experience negative effects such as fear, sadness, embarrassment, anger, guilt and disgust is the core of the neuroticism domain" (p. 14).

Although these personality traits have been studied in applied linguistics, no information is known about the possible relationships between such traits and the number of options in MC tests.

2.3. Empirical Background

In the following paragraphs, the findings of the studies, which have examined the psychometric properties of MC tests, are summarised. It has been attempted to select and report the findings of only very few studies due to space limitations.

Shizuka et al. (2006) compared 3-, and 4-option reading tests of university entrance examination in Japan by using the Rasch model. Two groups of 1,000 and 192 students were given 3-, and 4-option tests, respectively. 3-option items were created by removing the least frequently endorsed alternative per each item. The statistical analyses showed that the removal of the least popular distractor did not significantly change the mean item facility and the mean item discrimination. The researchers concluded that item difficulty, item discrimination, and test reliability were not necessarily influenced by the number of options. However, they noted that other options should be considered before firm conclusions are reached.

Baghaei and Amrahi (2011) conducted a study with 180 participants who were graduate and undergraduate English students. The goals of this study were to find the optimal number of options among 3-, 4-, and 5-option of English vocabulary tests by using the Rasch model regarding item statistics, person statistics, the performance of distractors, and reliability. Thirty-five MC items were used as the original item bank from which 3-, and 4-option MC tests were constructed by deleting distractors randomly. The data analyses by using Rasch model revealed that item difficulty, fitness of items, and reliability did not produce statistically significant changes in 3-, 4-, and 5- option MC tests,

although reducing the number of options increased the discrimination power, which prompted them to recommend 3-option items for educational uses.

Rodriguez (2005) conducted a meta-analysis of 80 years of research on the optimal number of options in MC tests. The findings showed that although, in most cases, reducing the number of options led to reducing item discrimination and item difficulty, reducing the number of options from four to three resulted in an insignificant increase in item discrimination. The lowest item discrimination was observed in two-option items. In most cases, option reductions caused a decrease in score reliability, but reducing four to three options increased score reliability slightly. Using the findings of the meta-analysis, Rodriguez concluded that 3-option items could be safely used to measure students' abilities.

Lee and Winke (2012) attempted to differentiate between 3-, 4, and 5-option items in the context of Korean high school English language learners. Three versions of 3-, 4-, and 4-option items were created from College Scholastic Ability Tests (CSAT). Seventy-three Korean speakers were asked to delete the least plausible options in two rounds. Two hundred and sixty-four Korean high school students were divided into three groups to take all three versions of three-, four-, and five-option tests with the one-week interval between each examination. A survey was used to find test takers' opinions about the number of options in the multiple-choice tests. The quantitative results revealed that the 3-option version was easier than 4-, and 5-option versions, while no significant differences were observed in mean item discrimination across the three versions. The qualitative data analysis revealed that most of the students preferred a 3-option test. It was recommended that test developers and researchers take into account the "statistical, affective, and contextual factors in determining the optimal number of options" (Lee & Winke, 2012, p. 119).

Sadeghi and Akhavan Masoumi (2017) aimed to investigate the effects of the differing number of options on item difficulty and discrimination of 3-, 4-, 5-, and 6-option MC vocabulary tests. A 6-option MC test was given to 194 pre-university students. Then, 3-, 4-, and 5-option MC tests were generated based on omitting the rarely selected distractors of 6-option MC test administration. The findings of this study showed that a significant difference emerged in item difficulty, and the authors explained that increasing the number of distractors led to a difficult test. In other words, a 6-option MC test was more difficult than 5-, 4-, and 3-option MC tests, respectively. On the other hand, 3-option MC test was easier than 4-, 5-, 6-option MC tests, respectively.

Most of these studies have examined the comparison of 3-, 4-, and 5-option MC tests in terms of validity, reliability, item discrimination, and difficulty, and mixed results have been reported. However, none of the studies have investigated the possible relationships between the personality traits of language learners and MC tests, so this research attempted to address this gap in the present study. The present study assesses the significance of the relationship of five personality traits, namely, openness, conscientiousness, extraversion, agreeableness, and neurotic and the performance of examinees in 3-, 4-, and 5-option MC tests. Likewise, this study is designed to investigate the effect of different number of options on psychometric properties including item difficulty, discrimination, and reliability in order to identify the optimal number of options. It is also intended to seek students' opinions about the optimal number of options. To achieve the objectives of the present research, we have used the following research questions to focus our study on:

1. What is the effect of the number of options on item difficulty, item discrimination, and reliability of an MC test?

2. What is the relationship between students' personality traits and the number of options on an MC test?

3. What are test takers' opinions regarding 3-, 4-, and 5-option MC items?

3. Method

3.1. Participants

Initially, one hundred and fifty freshman students aged between 18 and 22 were conveniently selected to take part in this study. Since this research was conducted on the same students through three sessions, all of the students were asked to fill in all of the given tests. However, some students were absent and did not answer some tests. Therefore, we excluded those students. After excluding outliers, 120, including 71 males and 49 females, were used for data analysis. They were from different parts of the country accepted at Imam Khomeini International University in Qazvin, Iran. The participants were in intact classes studying a three-credit general English course, majoring in different fields of study.

3.2. Instruments

To collect the data, the present mixed-methods study made use of three main instruments, including three multiple-choice tests, an inventory, and an attitude questionnaire as explained in the following paragraphs.

3.2.1. Multiple-choice Tests

The multiple-choice tests consisted of grammar, vocabulary, cloze test, and reading comprehension items. Three MC format tests, including 3-, 4-, and 5-option items were used. According to previous studies, there are two methods of test construction with varying number of options. The first is deletion method, which uses different types of deletion of distractors such as random deletion of distractors, elimination of the least discriminating distractors, deletion of the least frequently endorsed distractors, and the deletion of least plausible

distractors (Crehan et al., 1993; Delgado & Prieto, 1998; Owen & Froman, 1987; Shizuka et al., 2006; Sidick et al., 1994; Trevisan et al., 1991; Williams & Ebel, 1957). The second procedure is addition method, which uses adding distractors instead of removing them (Trevisan, Sax, & Michael, 1994).

After selecting the items for administration, 3-, and 5-option MC tests from our original 4-option MC test were created. The reason for these types of constructions was that the original form of our test had four options. For the construction of 3-option MC test, the distractors that were least frequently endorsed by students in the piloting stage, as explained in the following procedure section, were eliminated (Delgado & Prieto, 1998; Sidick et al., 1994), while 5-option MC test were formed by adding a distractor based on the judgment of a testing specialist (Trevisan, Sax, & Michael, 1994).

3.2.2. Big Five Inventory (BFI)

The Big Five Inventory consists of 44 items with a 5-point Likert scale from 1= strongly disagree to 5=strongly agree measuring five trait dimensions of personality: extroversion (8 items), agreeableness (9 items), conscientiousness (9 items), neuroticism (8 items) and openness (10 items). This inventory was drawn from John, Nuamann, and Soto (2008); the reliability of the BFI was 0.83. This inventory was translated into Persian to avoid confusion. The reliability of the translated scale turned out to be .72 in the present study. The items of this scale were also short and easy to understand.

3.2.3. Attitude Questionnaire

An attitude questionnaire was used to seek the opinions of students regarding 3-, 4-, and 5-option MC tests. This questionnaire asked all participants about their preferences in the 3-, 4-, and 5-option MC tests. This questionnaire was

adapted from Lee and Winke (2012). It was originally in English, but the Persian version was distributed to the participants. As for the validity of the Persian version, two testing experts in the filed were asked to examine the wording, clarity, and the appropriateness of the translated items to ensure the content validity of the Persian version. Therefore, vague words, unclear Persian equivalents in the items, and wordy phrases were removed from the items before the translated version was administered to students. The attitude questionnaire comprised two open-ended and two multiple-choice questions.

3.3. Procedure

Initially, a pilot test, including 46 items was administered to 60 undergraduate students. The participants were asked to select the best possible answer for each item. Out of the 46 items, 35 items had acceptable item discrimination, item difficulty, and reliability for the study. Then, the multiple-choice test with 35 items was distributed to 150 students. The outliers were deleted, and 120 students remained.

In administration, our procedures were explained to students, and students were asked to record their names in order to compare their performance on 3-, 4-, and 5-option MC tests. It was also attempted to find the relationship between their personality traits regarding their performance in different number of options. Pearson correlation was used to establish the relationship between personality traits and students' performance on the tests. There was a one-week interval between administrations. In the first week, the 5-option multiple-choice test was distributed to all of the participants. In the second week, all the participants were asked to complete the Big Five Inventory and the multiple-choice test containing 4-option items. In the third week, all the participants were asked to respond to the 3-option multiple-choice test and an attitude

questionnaire to find the viewpoints of participants about the optimal number of options in MC tests.

3.4. Data Analysis

After all the tests were administered and the data were collected, IBM SPSS Statistics (version 21) was used to analyze the data quantitatively. Regarding the first research question, descriptive statistics and one-way repeated measures ANOVA were used. With regard to the second research question, the Pearson product-moment correlation was run. To answer the third research question, themes were located based on key terms of students' responses. Percent values for each theme based on its frequency in their responses were computed. Finally, themes with high percent values were considered for analysis. In fact, these themes were their reasons of the preference. Such procedures are consistent with the guidelines given in Dornyei (2005) for qualitative analysis.

4. Results

4.1. Quantitative Results

4.1.1. RQ1: What is the Effect of the Number of Options on Item Difficulty, Item Discrimination, and Reliability of an MC Test?

4.1.1.1. Item Difficulty

According to the results of the descriptive statistics, 3-option MC test was the easiest, five-option MC test was the most difficult, and 4-option MC test was in-between (Table 2)

Table 2.*Descriptive Statistics for Option Formats Based on Item Difficulty*

	N	Mean	Std. Deviation
Three Options	35	23.44	6.29
Four Options	35	22.43	6.94
Five Options	35	19.93	6.47

Note. N=Number of items; Std=standard

A one-way repeated measures ANOVA was conducted to compare students' scores on 3-option MC test, 4-option MC test, and 5-option MC test. The means and standard deviations are presented in Table 2. There was a significant effect for scores, Wilks' Lambda=.54, $F_{(2,\ 119)}$=49.13, p<.0005, multivariate partial eta squared=.45.

Table 3. *One-Way Repeated Measures ANOVA Results for Option Formats in Terms of Item Difficulty*

	Value	F	Sig.	Partial Eta Squared
Pillai's Trace	.54	49.13	.000	.45
Wilks' Lambda	.54	49.13	.000	.45
Hotelling's Trace	.83	49.13	.000	.45
Roy's Largest Root	.83	49.13	.000	.45

We used post-hoc test so as to find the exact location of mean differences among three tests. The results of LSD test indicated that item difficulties were statistically significant among three types of MC tests. As Table 4 shows, all of the tests had significant differences in item difficulty. In fact, significant differences were observed in the item difficulty of 3-, 4-, and 5-option MC tests.

Table 4. *Post hoc Results for Option Formats According to Item Difficulty*

Tests		Mean Difference (tests)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
3 options	4	1.00*	.36	.006	.29	1.72
	5	3.50*	.37	.000	2.76	4.25
4 options	3	-1.00*	.36	.006	-1.72	-.29
	5	2.50*	.33	.000	1.84	3.16
5 options	3	-3.50*	.37	.000	-4.25	-2.76
	4	-2.50*	.33	.000	-3.16	-1.84

Note: *=The mean difference is significant at the .05 level.

4.1.1.2. Item Discrimination

Following past research (Lee & Winke, 2012; Swanson et al., 2006; Tarrant, Ware, & Mohammed, 2009), point biserial correlation was used for the calculation of item discrimination. According to descriptive statistics, 4-option MC tests had the highest mean item discrimination and 3-option MC test had the lowest mean item discrimination (see Table 5).

Table 5. *Descriptive Statistics for Option Formats based on Item Discrimination*

	N	Mean	Std. Deviation
Three Options	35	.34	.19
Four Options	35	.37	.17
Five Options	35	.35	.17

Note. N=Number of items

As shown in Table 6, the results of one-way repeated measures ANOVA revealed that statistically no significant differences were observed in item discrimination (Wilks' Lambda=$F_{(2, 119)}$=.97 , p=.66, η_{p2}=.024).

Table 6. *One-Way Repeated Measures ANOVA Results for Option Formats in Terms of Item Discrimination*

	Value	F	Sig.	Partial Eta Squared
Pillai's Trace	.02	.40	.66	.02
Wilks' Lambda	.97	.40	.66	.02
Hotelling's Trace	.02	.40	.66	.02
Roy's Largest Root	.02	.40	.66	.02

4.1.1.3. Reliability Coefficients

To examine the reliability of three MC tests, Cronbach's alpha was run. Reliability coefficient in 5-option MC test is the highest, while 3-option MC test

shows the lowest reliability. 4-option and 5-option MC tests are relatively similar (Table 7)

Table 7. *Reliability of Three MC Tests*

	Three options	Four options	Five options
Reliability	.82	.85	.86

4.2. RQ2: What is the Relationship between Students' Personality Traits and the Number of Options on an MC Test?

The second research question of this study was concerned with the relationship between the number of options and personality traits. In order to answer this research question, the Pearson product-moment correlation was run. As Table 8 shows, only extroversion trait is correlated with 3-option MC tests ($p < .05$). In fact, there is a low, negative correlation between Extroverted language learners and three-option MC test ($r=-.175$, $n=3$, $p=.05$, $R^2=.030$) with a very small effect size. According to Cohen (1988), correlation .10 to .29 or -.10 to -.29 is small. The rest of the personality traits are not correlated with three MC format tests (three-, four-, and five-option MC tests).

Table 8. *Pearson Product-Moment Correlation Coefficients between Types of Personality Traits and Three Forms of MC Tests*

Personality Traits		Three Options	Four Options	Five Options
Extraversion	Correlation	-.17	-.06	-.08
	Sig.	.04[*]	.45	.34
Agreeableness	Correlation	-.14	-.07	-.17
	Sig.	.10	.39	.06
Conscientiousness	Correlation	.02	.08	.02
	Sig.	.79	.38	.77
Neuroticism	Correlation	.00	-.03	-.08
	Sig.	.95	.72	.38
Openness	Correlation	-.02	.02	.00
	Sig.	.77	.76	.99

Note: [*] = The mean difference is significant at the .05 level

4.3. Qualitative Results

4.3. RQ3: What Are Test Takers' Opinions Regarding 3-, 4-, and 5-Option MC Items?
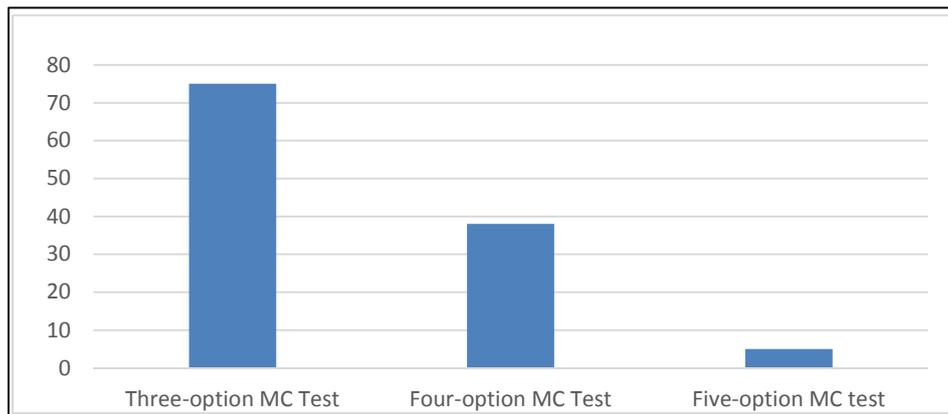

Figure 1. *Students' Preference of 3-, 4-, and 5-option MC Tests*

Regarding the first question of attitude questionnaire as which multiple-choice item format do you prefer? (3-, 4-, and 5-option MC tests), 62.5 % of test takers (n=75) preferred 3-option MC test, 31 % selected 4-option MC test, and 5.8 % preferred 5-option MC test (see Figure 1). Overall, more than half of the students preferred 3-option MC test.

The second question of attitude questionnaire is open-ended, aiming to reveal the reasons of students' responses to the first question of attitude questionnaire. After extracting themes out of the students' responses, these themes were classified and the percent values of each theme were computed. Table 9 shows all the extracted themes and the number of students and the percentage of students who provided the themes.

Table 9. *Extracted Themes out of Given Responses*

MC Tests	Theme	Number	Percent
Agree with Three Options	High probability of correct answer (.33 being correct for 3-option tests vs. .25 being correct for 4-option tests)	12	%10
	Easiness of test and selection of options	9	%7.5
	Saving time and increasing your speed in reading and answering tests	12	%10
	Low probability of the error	13	%10.8
	Lack of nonfunctional options	1	%0.8
	More concentration and less confusion, distraction, and stress	7	%5.8
	Reducing uncertainty	5	%4.16
	High chance of correct answer	3	%2.5
	Less boring	1	%0.8
	Less challenging	2	%1.6
	Total	**65**	**%54**
Agree with Four Options	4-option tests are neither difficult like 5-option test nor easy like 3-option test. They go in between.	15	%12.5
	Accustomed to four-option MC test and also being familiar with this format which leads to easiness	15	%12.5
	Students believed that they select the correct answer from two or three options because one or two options, which were nonfunctional options, were always deleted by test takers.	2	%1.16
	Less confusion	1	%0.8
	Being standard test format	6	%5
	Having less probability of the error	1	%0.8
	Total	**40**	**%33**
Disagree with Four Options	Difficulty in test and selection of options	1	%0.8
	Increasing the probability of error	2	%1.16
	Needing more time for reading and answering the options		
	Total	**3**	**%2.5**
Agree with Five Options	Having more options to choose	1	%0.8
	High ability to answer 5-option items	1	%0.8
	Low chance of correct answer	1	%0.8
	Difficulty of test	1	%0.8
	Better comparison of performance	1	%0.8
	Total	**5**	**%4.16**
Disagree with Five Options	Having nonfunctional options	1	%0.8
	Total	**1**	**% 0.83**

According to Table 9, most of the students (54 %) preferred 3-option MC test; reasons for such preferences include high probability of correct answer, low probability of the error, and saving time in reading and answering the test. In 4-option MC test, 40 out of 120 students (% 33) preferred this format because 4-option MC test is neither difficult like 5-option MC test nor easy like 3-option MC test. They also expressed that the majority of the tests, which they take, were 4-choice MC tests. In this way, they were familiar and comfortable with this test format. However, 3 out of 120 students disagreed, arguing that it is difficult and time-consuming. Very few students, 5 out of 120, favored 5-option MC test due to having more alternatives to select, the difficulty of the test, better comparison of test takers' performance, and requiring high capacity.

5. Discussion

The present study intended to compare 3-, 4-, and 5-option MC tests in terms of psychometric characteristics including item difficulty, item discrimination, and reliability so as to find optimal number of options. The study also aimed to investigate the relationships between five personality traits, namely, openness, conscientiousness, extroversion, agreeableness, and neuroticism and three MC tests. In this section, the findings of the present study are presented and discussed.

The first finding of this study showed that statistically significant differences were observed in item difficulty of 3-, 4-, and 5-option MC tests. The results revealed that increasing the number of options made the test more difficult. Therefore, 5-option MC test was the most difficult test, 3-option MC test was the easiest one, and 4-option MC test was in-between. The finding of this study is consistent with the findings of some other previous studies (Asmus, 2016; Kilgour & Tayyaba, 2016; Lee & Winke, 2012; Nwadinigwe & Niabi, 2013; Rogers & Harley, 1999; Sadeghi & Akhavan Masoumi, 2017). Lee and Winke

(2012), for example, obtained similar findings: Significant differences in all 3-, 4-, and 5-option MC tests in which 3-option MC test was easiest one, 5-option MC test was the hardest one, and 5-option MC test was in-between. Asmus (2016) and Nwadinigwe and Niabi (2013) found significant differences in item difficulty of 3- and 5-option MC tests; 3-option MC test was significantly easier than 3-option MC test. Kilgour and Tayyaba (2016) also found significant differences in 4-, and 5-option MC tests and 3- and 4-option MC tests. Recently, Sadeghi and Akhavan Masoumi (2017) compared item difficulty of 3-, 4-, 5-, and 6-option MC tests and found significant differences in item difficulty of 3-, 5-, and 6-option MC tests and 4-, 5-, and 6-option MC tests. However, no significant differences were found between 3-and 4-option MC tests.

The first finding of the study is not in keeping with those of some other studies. Baghaei and Amrahi (2011), for instance, compared item difficulty of 3-, 4-, and 5-option MC tests. They found no statistically significant differences among these three test forms in terms of item difficulty. Similarly, Farhady and Shakery (2000) compared 3-, 4-, and 5-option MC tests in terms of item facility.

The results echoed those of Baghaei and Amrahi (2011). In the same way, Currie and Chiramanee (2010) and Thanyapa and Currie (2014) and Schneid et al., (2014) made this comparison and found similar results. We speculate that one possible explanation for significant results in item difficulty is students of the present study may never have encountered 5-, and 3-option MC tests, so their performance was very low and high in these tests, respectively and big differences yielded in difficulty of tests by deleting distractors.

In comparison to the first finding, the second finding of this study revealed that item discrimination and reliability of 3-, 4-, and 5-option MC tests were not statistically significant. Four-option MC test had higher discrimination (.37), 3-, and 5-option MC tests had nearly the same discrimination (.34 and .35). In terms of reliability, 3-option MC test had the lowest reliability coefficient (.82); the

reliability of 4-, and 5-option MC tests were nearly equal (.851 and .855). The results of this study are in line with those of Lee and Winke (2012) and Schneid et al., (2014), who compared 3-, 4-, and 5-option MC tests in terms of item discrimination and reliability. Both studies obtained the same results, which were no significant differences in item discrimination and reliability. Likewise, Sadeghi and Akhavan Masoumi (2017) compared item discrimination of 3-, 4-, 5-, and 6-option MC tests and reported the same findings. Moreover, Shizuka et al., (2006) and Rogers and Harley (1999) explored the item discrimination and reliability of 3-, and 4-option MC tests. The findings indicated that reducing the number of options changed item discrimination and reliability, but these differences were not statistically significant. Similarly, Tarrant and Ware (2010) investigated the discrimination of 3- and 4-option MC tests and found insignificant differences.

In terms of reliability, the findings of the present study are in line with some of the previous studies. Farhady and Shakery (2000) compared the reliability of 3-, 4-, and 5-option MC tests. Their findings were the same as those in the present study in which reliability of 4- and 5-option MC tests was nearly the same and higher than that of 3-option MC test, but the 3-option test had the lowest reliability. Baghaei and Amrahi (2011) also found high reliability of 4-, 5-, and 3-option MC test, respectively. However, no significant differences in the reliability of 3-, 4-, and 5-option MC tests were observed. Furthermore, Thanyapa and Currie (2014) and Vegada et al. (2016) found no significant differences in reliability. Asmus (2016) and Nwadinigwe and Niabi (2013) also compared the reliability of 3- and 5-option MC tests and found insignificant differences.

The insignificant differences in item discrimination could be because all of the items had the same stem and correct answer. They differed only in the number of options in which more options led to more burden in deleting

nonfunctional distractors. As Lee and Winke (2012) argued, although having more options makes a test more difficult, change in number of options does not affect item discrimination and reliability.

However, these findings do not support those of Thanyapa and Currie (2014), and Hogben (2015). Thanyapa and Currie (2014) investigated item discrimination of 3-, 4-, and 5-option MC tests. They found that 5-option MC test had significantly larger discrimination than to 3- and 4-option MC tests. Hogben (2015) compared 3-, 4-, and 5-option MC tests in terms of item discrimination and reliability. He found significant differences in discrimination and reliability of 3- and 4-option MC tests and 4- and 5-option MC tests. Such mixed results reflect the very fact that researchers may have adopted different research methods, data types, and research designs, as a result of which significant differences in item discrimination and reliability have been reported in some studies, but not in other studies.

Another finding of this study was that a relatively low, negative correlation was found between extroversion and 3-option MC test. Therefore, the more extroverted the students are the poorer their performance on 3-option items. The rest of the personality traits are not correlated with different formats of MC tests. This implies that there are no correlations between different personality traits and different formats of MC tests. Further research on the possible relationship between personality traits and differing number of options needs to be done to come up with more robust findings in this area. These findings support those of Esfandiari and Radfar (2017), who found that extroversion did not correlate with students' test performance. Furthermore, in Ehrman and Oxford's (1995) study, no significant relationship was found between personality traits and language learners' performance on reading comprehension tests. The findings of the present study and those of previous research suggest that researchers have reported mixed results on the relationship between personality

traits and language learners' test performance. Some possible reasons for such inconsistencies stem from small sample sizes, poor research designs, unsophisticated methodological approaches, different research settings, and the absence of clearly operationalized variables.

The qualitative results of the present study revealed that more than half of the students (62.5 %) preferred 3-option MC test, 38 students (31 %) selected 4-option MC test, and seven students (5.8 %) preferred 5-option MC test. The students were asked to reveal how much they liked 3-option MC test (using a seven–point Likert-scale). 58 % of students liked 3-option MC test, while 19 % of them disliked it, and 22 % of them had no idea about it.

6. Conclusion

According to the findings of the present study, fewer numbers of options made an MC test easier because nonfunctional options were removed. When fewer options are used in an MC test, test takers have more time to concentrate, read, and get an item correct. Removing these nonfunctional options also reduced confusion, distraction, and boredom. As Lee and Winke (2012) also concluded, students get low scores on a 5-option MC test due to the anxiety it creates. Sadeghi and Akhavan Masoumi (2017) also added that when students were faced with an unfamiliar test format like 5-, and 6-option MC tests, they were more anxious, confused, and stressed out.

The second conclusion drawn from the findings of the present study is that the item discrimination and reliability of a test are not affected by the number of options. Therefore, "statistical, affective, and contextual factors" should be taken into account in order to find appropriate number of options (Lee & Winke, 2012, p. 119). In fact, these factors should also be used to find the optimal number of options in addition to mathematical factors including item discrimination and reliability.

The third conclusion of this study is that, generally, no relationship can be established between students' personality traits and MC test formats. No matter whether students are open, neurotic, conscientious, agreeable, or extroverted, their performance is not affected by their personality traits. Therefore, test developers should not worry about the personality traits of test takers when constructing MC test with differing number of options. However, this conclusion should be treated with caution because the number of students in the present study was very low.

According to the qualitative findings, although the majority of students preferred 3-option MC tests, a considerable number of students also endorsed 4-option MC tests. The practicality of the test and familiarity with the test format are two important factors for the selection of the appropriate number of options on MC tests. Overall, it can be concluded that we should consider not only statistical, affective, and contextual factors, but practicality and familiarity of the test should be considered in order to choose optimal number of options.

The findings of this study have some implications for language teachers and test developers. Both test developers and language teachers can benefit from the findings in order to develop optimal number of options in both formative and summative tests. The findings of this study also suggest that test developers can construct 3-, 4- or 5-option MC tests because differing number of options impact on item difficulty but not on item discrimination and reliability. Therefore, they can use different numbers of options for MC tests based on different situational factors. Additionally, the findings of the present study showed that students' performance with different personality traits did not correlate with different MC test formats. Therefore, it can be beneficial for language teachers to construct different MC test formats without considering students' personality traits.

References

Al-Rukban, M. O. (2006). Guidelines for the construction of multiple choice questions tests. *Journal of Family & Community Medicine, 13*(3), 125-133.

Asmus, E. P. (1981). The effect of altering the number of choices per item on test statistics: Is there better than five?. *Bulletin of the Council for Research in Music Education, 65*(1), 1-15.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford university press.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

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.

Brown, J. D. (2005*). Testing in language programs: A comprehensive guide to English language assessment.* McGraw-Hill College.

Budescu, D. V., & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement, 22*(3), 138-196.

Crehan, K., 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.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research.* California: Sage Publications, Inc.

Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-491.

Dawson, J. R. (2015). Comparative analysis on personality traits and motivation on the international student's academic performance in universities in Taiwan. *International Journal of Humanities and Management Sciences, 4*(3), 209-216.

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment, 14*(3), 197.

Dornyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* London: Lawrence Erlbaum Associates.

Dornyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited.* New York: Routledge.

Ehrman, M. E., & Oxford, R. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal, 79*(1), 67-89.

Esfandiari, R., & Radfar, S. (2017). An investigation into the relationship between personality traits and Iranian EFL learners' performance on C-test. *Teaching English Language, 11*(1), 167-188.

Farhady, H., & Shakery, S. (2000). Number of options and economy of multiple-choice tests. *Roshd Foreign Language Teaching Journal, 14*(1), 57.

Farhady, H., Ja'farpur, A. & Birjandi, P. (1994). *Testing language skills from theory to practice.* Tehran: SAMT.

Fulcher, G. (210). *Practical language testing.* London: Routlege.

Green, K., Sax, G., & Michael, W. (1982). Validity and reliability of tests having differing numbers of options for students of differing level of ability. *Educational and Psychological Measurement, 42*(1), 239-245.

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

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.

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.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: item-writing guidelines and an analysis of auditing test banks. *Journal of Education for Business, 73*(2), 94-97.

Heaton, J. B. (1988). *Writing English language test: Longman handbooks for language teachers.* New York: Longman.

Hogben, D. (2015). The reliability, discrimination and difficulty of word-knowledge tests employing multiple-choice items containing three, four, or five alternatives. *The Australian Journal of Education, 17*(1), 63-68.

Jensen, M. (2015). Personality traits, learning and academic achievements. *Journal of Education and Learning, 4*(4), 91-118. doi:10.5539/jel.v4n4p91

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality. *Theory and Research, 2*(1999), 102-138.

John, O. P., Robins, R. W., & Pervin. L. A. (Eds.). (2008). *Handbook of personality: Theory and research.* New York: Guilford Press.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education, 21*(3), 571-585.

Lee, H., & Winke, P. (2012). The differences among three-, four-, five-option-item format in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-123. doi: 10.1177/0265532212451235

Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology.* Holt, Rinehart and Winston. Inc.: Orlando, FL.

Nejati, R., & Moradi, M. (2015). Utility of complex alternatives in multiple-choice items: The case of all of the above. *Research in Applied Linguistics, 6*(1), 87-97.

Ng, A., & Chan, A. (2009). Different methods of multiple-choice test, implications and design for further research. *Proceedings of the International MultiConference of Engineers and Computer Scientists, 2*(1), 1-6.

Nwadinibgwe, P. I., & Naibi, L. (2013). The number of options in a multiple-choice test item and the psychometric characteristics. *Journal of Education and Practice, 28*(4), 189-196.

Nwadinibgwe, P. I., & Naibi, L. (2013). The number of options in a multiple-choice test item and the psychometric characteristics. *Journal of Education and Practice, 28*(4), 189-196.

Osterlind, S. J. (2002). *Constructing test items multiple-choice, constructed-response, performance, and other formats.* Dordrecht: Springer.

Osterlind, S. J. (2002). *Constructing test items multiple-choice, constructed-response, performance, and other formats.* Dordrecht: Springer.

Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple-choice items?. *Educational and Psychological Measurement, 47*(2), 513-522.

Ramos, R. A., & Stern, J. (1973). Item behavior associated with changes in the number of alternatives in multiple-choice items. *Journal of Educational Measurement, 10*(4), 305-310.

Razavipour, K. (2013). Assessing assessment literacy: Insights from a high-stakes test. *Research in Applied Linguistics, 4*(1), 111-131.

Richards, J. C., & Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics.* London: Pearson Education Limited.

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.

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.

Sadeghi, K., & Akhavan Masoumi, G. (2017). Does number of options in multiple choice tests affect item facility and discrimination? An examination of test-taker preferences. *Journal of English Language Teaching and Learning, 9*(19), 123-143.

Schneid, S. D., Armour, C., Park, S. Y., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: Response time, psychometrics and standard setting. *Medical Education, 48*(10), 1020-1027.

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.

Foreign Language Education and Research, 23(1), 35-57. doi: 10.1191/0265532206lt319oa

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

Swanson, D. B., Holtzman, K. Z., Allbee, K., & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine, 81*(10), S52-S55.

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.

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), 1-8.

Thanyapa, I., & Currie, M. (2014). The number of options in multiple-choice items in language tests: does it make any difference? *Evidence from Thailand. Language Testing in Asia, 8*(4), 1-21.

Trevisan, M. S., Sax, G., & Michael, W. B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement, 54*(1), 86-91.

Vegada, B., Shukla, A., Khilnani, A., Charan, J., & Desai, C. (2016). Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian Journal of Pharmacology, 48*(5), 571-575.