Journal of Modern Research in English Language Studies
Print ISSN: 2676-5357
Online ISSN: 2676-5985

Attitudes of Language Teachers toward Multiple-choice Item Writing Guidelines: An Exploratory Factor Analysis

Mahdi Ganji[1], Rajab Esfandiari[*2]

M.A. Student, Department of English Language, Faculty of Humanities, Imam Khomeini International University, Qazvin Iran, *mgmahdiganji@yahoo.com*
Assistant Professor, Department of English Language, Faculty of Humanities, Imam Khomeini International University, Qazvin, Iran, *esfandiari@hum.ikiu.ac.ir*

Abstract

Previous research has shown that the construction of multiple-choice (MC) items is a very difficult task. As such, textbook writers have proposed some guidelines to help item writers to write more effective items. However, such guidelines reflect the intuition of their writers, and most of them are not necessarily supported by empirical research, and what is preached may not be practiced. The purpose of the present study was, therefore, to analyze the attitudes of language teachers in an EFL setting to better understand if they follow the guidelines when developing MC items. To that end, a 28-item, 5-point Likert type, researcher-made questionnaire was used to collect data from 661 Iranian language teachers. The data were analyzed using SPSS (version, 25). Results from frequency tallies and percent values showed the significance of the majority of the guidelines in the construction of MC items. However, mixed results were reported for one of the guidelines, and another guideline was considered unimportant. Findings from factor analysis yielded four major factors underlying the guidelines: Developing plausible distractors, editing and proofreading guidelines, formatting and refining items, and avoiding clues to the correct response. Drawing on the findings, we discuss the pedagogical implications for how to best develop and fine-tune MC guidelines.

Keywords: Attitudes, Factor Analysis, Multiple-choice Items

Received 09 May 2020 Accepted 01 June 2020
Available online 02 June 2020 DOI: 10.30479/jmrels.2020.13247.1638

1. Introduction

Multiple-choice (MC) tests are one of the most popular test formats. Regardless of the arguments about the effectiveness of using MC tests in classroom assessment, an MC test is the most common method of measuring students' performance all around the world (Kiss & Selei, 2017). It consists of a stem followed by several alternatives, and test takers need to select the correct response among other alternatives known as distractors. Although MC tests are widespread in most high-stakes tests and classroom assessment, developing a high-quality MC test requires thorough and detailed consideration. Cohen and Swerdlik (1999) noted that, "the creation of a good test is not a matter of chance; it is the product of the thoughtful and sound application of established principles of test construction" (p. 215). Constructing multiple-choice items requires a combination of art, skill, and experience. As Crehan, Haladyna, and Brewer (1993) asserted, "item writing is often viewed as more art than science" (p. 241). Therefore, some teachers may not have the ability to produce high-quality MC items (Haladyna, Rodriguez, & Stevens, (2019).

Studies have shown that many MC items in teacher-made tests are of poor quality and do not meet the MC item writing guidelines (e.g., Tarrant, Knierim, Hayes, & Ware, 2006). Richichi (1996) found that items, which did not conform to item writing guidelines, had psychometric properties, including low discrimination. The vast number of measurement textbooks and empirical research have provided excellent advice to aid examiners in developing effective MC tests. MC item writing guidelines provide several suggestions for constructing test items. According to Mehrens and Lehmann (1991), teacher-made tests suffer from major deficiencies because teachers do not necessarily follow the guidelines handed down to them through measurement textbooks. In order to produce more effective MC items and, consequently, obtain more reliable and valid scores, teachers should consider several guidelines in developing multiple-choice items.

Recent developments in test construction have heightened the need for some guidelines in designing MC items. Despite the plethora of studies focusing on how to construct MC items and providing some guidelines in item writing, so far, very little attention has been paid to teachers' attitudes toward MC item writing guidelines. In other words, it is unclear how much teachers care about different MC item writing guidelines and whether they are aware of such guidelines. Therefore, this study explored the attitudes of Iranian English language teachers about multiple-choice item writing guidelines for the following reasons. Seeking language teachers' opinions helps us to figure out how well-versed they may be in knowing and applying the guidelines in the construction of effective MC items. Additionally, careful analyses of their attitudes may contribute to raising their consciousness

regarding the significance attached to the guidelines. Therefore, we posed the following research question in our study: What is the factor structure of the attitudes of Iranian English language teachers toward MC item writing guidelines?

2. Literature Review

In the following three subsections, a selective review of some concepts relevant to the present study is presented. First, the construction of MC tests is briefly explained. Next, the advantages and disadvantages of MC items are explained. Finally, guidelines related to MC writing are outlined.

2.1. Development of MC items

Developing high-quality multiple-choice tests is difficult. It has been a serious concern for a long time even for professional item writers (Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006). At first glance, it may seem easy to construct MC items, but when it comes to real item writing, some difficulties arise. Developing multiple-choice items is more difficult than other formats since item writers need to develop effective options and stems (Moreno, Martı´nez, & Mun˜iz, 2006). In addition, "Good MC tests are generally more complex and demand a lot of time to create compared with other types of tests. It requires a certain amount of skills and knowledge" (Torres, Lopes, Babo, & Azevedo, 2011, p. 2).

Difficulties of developing multiple-choice items are not just related to mechanical checking of the items. Developing high-quality multiple-choice items requires technical skills (Haladyna, 2004). "Technical writing—such as preparing test items—is especially difficult because it demands an extraordinarily high degree of precision in language use" (Osterlind, 2002, p. 109). Osterlind (2002) pointed out that item writers must possess abilities to predict how examinees react to items. In other words, item writers must consider feelings and attitudes of examinees in responding to different items and reduce the number of potential factors which cause examinees to answer items without adequate knowledge. As Wainer, Wadkins, and Rogers (1983) commented, developing high-quality test items "involves the consideration of every possible interpretation of the item" (p. 3).

Teachers usually complain about the lack of adequate time for examining and constructing quality tests (Schrock & Mueller, 1982). However, there are various measurement textbooks and empirical research which help test designers to develop high-quality multiple-choice tests.

2.2. Advantages and Disadvantages of MC Items

Multiple-choice tests are widely used all around the world compared to other test formats (Haladyna, et. al. 2019). Among different test formats,

the multiple-choice item can assess a large variety of course objectives (Nitko, 1985). Also, it is more flexible and may cover a large amount of content (Haladyna, & Downing, 1989a). Williams (1984) stated that the most important feature of objective tests which distinguish them from other types of tests is that syllabus content can be covered broadly and extensively.

Another advantage is that MC tests provide an opportunity for precise interpretation of test and content validity (Haladyna, & Downing, 1989a). MC items pave the way for examiners to determine which items are too easy or too difficult for examinees and help to detect the strengths and weaknesses of students in particular course material (Kubiszyn & Borich, 2013). MC tests do not require examinees to write their answers like essay type ones. Therefore, examinees are not able to hide their lack of knowledge by writing something unclear or difficult to understand (Osterlind, 2002).

Objectivity in scoring paves the way for MC items to compare the performance of students from class to class (Torres, et al., 2011). Also, the scoring of an MC test is easier and quicker compared with a constructed response format. Roediger and Marsh (2005) pointed out that since administration and scoring in MC tests are more comfortable than constructed-response tests, it is appropriate for large-scale measurements. Also, MC tests, as objective tests, have high reliability (Wilson & Wang, 1995). Bailey (2018) remarked that scoring reliability and practicality are considered to be strengths of MC tests.

In addition to merits, MC tests have their own limitations. Some educators have criticized the MC test since it encourages learners to recall some facts and may lead to surface learning (Burton, Sudweeks, Merrill, & Wood, 1991; Tamir, 1990). As Linn, Baker, and Dunbar (1991) argued, there is an uncertainty about the effect of the MC test on students' learning. These researchers question whether students who are able to answer MC items correctly have gained the ability to understand different concepts. MC items are incapable of gathering students' explanations about answer of a certain item or providing justification for items (Liu, Lee, & Linn, 2011). Paxton (2000) suggested that in order to use MC tests as a learning tool, their use should be limited to formative assessment instead of summative assessment.

Developing MC items is a time-consuming process and requires training to develop high-quality items (Kubiszyn & Borich, 2013). The reason is that MC item writers need to think of plausible distractors and stems. If MC items are not designed carefully, they may have more than one absolute correct option (Kubiszyn & Borich, 2013). Furthermore, Lennox (2009) believed that the use of wrong content could be troublesome. In addition, test-wise students may benefit from item constructors' bias in the setting of special content or format. For instance, test constructors may prefer

to put the correct answer in a specific position among alternatives, which provides a hint for test-wise students.

MC items are often criticized for measuring low-level cognitive processes (Martinez, 1999). The results of several studies have shown that MC items are best suited for measuring low-level learning like recall of the exact definition rather than high-level thinking (Tamir, 1990). In other words, MC items cannot provide opportunities for communicative competence and problem-solving activities (Paxton, 2000; Tamir, 1990). Torres, et. al. (2011, p. 2) argued that "the ability to organize the information or the communication and the creativity skills" is hard to measure by MC items. Therefore, MC items should not test superficial knowledge such as memorization of facts and should measure higher level thinking such as comprehension, evaluation, analysis, synthesis, and application (Kubiszyn & Borich, 2013).

2.3. Guidelines for Writing Effective MC Tests

Most educational measurement textbooks have devoted a certain amount of space, or a chapter, to MC item writing, which shows the significance of the quality of MC tests in educational measurement (Haladyna & Rodriguez, 2013). Authors in measurement textbooks and journals have recommended numerous guidelines to assist in developing high-quality MC items. Nitko (1985) remarked that "elder item writers pass down to novices' lists of rules and suggestions which they and their item-writing forefathers have learned through the process of applied art, empirical study, and practical experience" (p. 201). Some of the guidelines listed in different research articles and measurement textbooks are straightforward, and most of the test constructors agree on their usefulness in developing MC items (e.g. "using correct punctuation, spelling, and grammar in items"). However, other guidelines such as "avoid using all of the above" or "word the stem positively" have received numerous empirical research and their effectiveness in constructing multiple-choice items are controversial.

Haladyna and Downing (1989b) examined 46 textbooks and other sources in educational measurement and developed a taxonomy of 43 MC item writing guidelines. Some of these guidelines were mentioned in different textbooks and authors paid great attention to these guidelines. However, some other guidelines did not meet the strong consensus among textbooks. In fact, the taxonomy is the first and the most comprehensive taxonomy of item writing guidelines.

Haladyna and Downing (1989a) examined the results of 96 theoretical and empirical studies on the validity of item writing guidelines which appeared in previous research. They intended to determine whether support existed for every single guideline in the published articles. The results

showed that the guidelines regarding the number of options ("use as many options as feasible") received the highest attention. Also, about half of the rules remained unsupported since there were no theoretical and empirical studies.

Haladyna, Downing, and Rodriguez (2002) reorganized and updated the guidelines found by Haladyna and Downing (1989a) and classified them into five different categories: content concerns, formatting concerns, style concerns, writing the stem, and writing the choices. They used two sources of evidence to examine the validity of 31 multiple-choice item writing guidelines. These two sources were the consensus of 27 measurement textbooks and the results of 27 empirical research studies. They examined the validity of guidelines through evaluation of each study regarding guidelines as cited and supported, cited and not supported, or not cited. They reported that some of the guidelines were common in most of the measurement textbooks and had a strong consensus. However, some other guidelines were mentioned in fewer studies. Therefore, the authors could not justify all the guidelines.

Other researchers have conducted various studies to provide MC item writing guidelines. The majority of researchers used guidelines presented in the study of Haladyna et al. (2002) as their source for analysis. As a result, there were nuanced differences in their taxonomy in comparison to original guidelines. They took different approaches in their analysis of measurement textbooks and empirical research and found that all the original guidelines had strong support. In fact, the guidelines presented by Haladyna et al. (2002) are comprehensive and almost include all the valid guidelines mentioned in various measurement textbooks and empirical research. In the following paragraphs, we briefly review the findings of some studies in presenting MC item writing guidelines.

Vacc, et. al. (2001) provided a few general guidelines for writing effective item stems, keyed responses, and distractors. No differences were found in guidelines compared to those presented by Haladyna et al. (2002). Moreno, et al. (2004) produced a set of 12 guidelines as did Haladyna et al. (2002) with different phrasing in different categories. In fact, they produced a shortened version of guidelines by removing irrelevant and repetitive guidelines.

Frey, et al. (2005) conducted another similar study to obtain valid item writing rules. The authors examined 20 classroom assessment textbooks to identify a list of valid rules for item construction. The study provided 40 item writing rules for four different item formats: MC, matching, true-false, and completion. The guidelines presented in MC format were the same as those presented in the study of Haladyna, et al. (2002). The only difference had to

do with presenting new guidelines as follows: “all parts of an item or exercise should appear on the same page”.

Moreno et al. (2006) took a different approach in their study on presented guidelines in their previous study: To identify and measure merits and demerits of guidelines. Guidelines were given to two groups of assessors: experts or professionals in measurement and teachers as the users of the guidelines. “Both groups of assessors stress the utility of the set, which results from its parsimony and synthesis of other proposals, and rate as adequate the avoidance of overlap and contradictions between the guidelines” (Moreno et al., 2006, p. 67). Furthermore, the two groups recommended revisions to some ambiguous guidelines.

Moreno, et al. (2015) changed the process of developing MC item-writing guidelines. They claimed that “many different guidelines have been presented for the construction of multiple choice items. Those guidelines have been based on the observation of errors when constructing items but not on any clear scientific criterion” (p. 388). Therefore, they produced nine general guidelines based on validity criteria. They “used the properties of adjustment, precision, and differentiation, applying them to three basic phases of instrument construction: the definition of the objective and its context; their expression in the instrument and item stem; and the elaboration of response options” (p. 388). Finally, the authors provided a checklist containing 24 questions regarding presented guidelines to make it clear and more tangible for item designers.

3. Method

3.1. Participants

Participants of the present study were 661 male and female Iranian English language teachers holding BA, MA, and Ph.D. degrees in English Language Teaching (ELT), English Translation, English Literature, and Linguistics. They were native speakers of Persian and varied greatly in terms of teaching experience, ranging from novice to fully experienced teachers. Table 1 presents the demographic information of the participants, including the gender, age, teaching experience, academic degree, and field of study.

This study employed convenience sampling to have access to language teachers. This type of sampling is usually used when the participants possess certain key characteristics relating to the investigation. Since including a vast number of English teachers is difficult through convenience sampling, snowball sampling was also used to gather more participants. “Snowball sampling involves a ‘chain reaction’ whereby the researcher identifies a few people who meet the criteria of the particular study and then asks these participants to identify further members of the population” (Dörnyei, 2010, p.

61). This technique is useful when a researcher needs a large group of respondents.

Table 1

The Demographic Information of the Language Teachers

Demographic categories	Frequency	Percent	Mean
Gender			
Male	262	39.6	
Female	399	60.4	
Total	661	100	
Age			40.04
Teaching experience			17.50
Academic degree			
BA holders	312	47.2	
MA holders	322	48.7	
PhD holders	27	4.1	
Total	661	100	
Field of study			
ELT	503	76.1	
Literature	55	8.3	
Translation	66	10.0	
Linguistics	37	5.6	
Total	661	100	

3.2. Materials and Instruments

To investigate the research question posed in the present study, the researchers used a researcher-questionnaire (see Appendix A) consisting of two major parts. The first part of the questionnaire provided the participants' profile in terms of gender, age, year of teaching experience, educational level, field of study, and current teaching situation. The second part of the questionnaire included 28 items developed based on the relevant literature and the model proposed by Haladyna, et al. (2002), who presented 31 guidelines for constructing MC items, but in this study, the researchers ignored guidelines which were related to content because they did not measure linguistic abilities. Furthermore, the wording of guidelines was changed to statements and simplified for participants. Vague guidelines were clarified by examples to ensure the intention of the researchers for participants.

To ensure sufficient variation among the item scores, teachers were asked to mark their responses on a 5-point Likert scale ranging from "*not important*" (1) to "*very important*" (5) in the order of significance of MC item guidelines. Scores of 1 or 2 indicated teachers' disapproval of using those guidelines in constructing multiple-choice items whereas scores of 4 or 5 indicated approval.

The internal consistency reliability of the questionnaire was estimated using Cronbach's alpha, which turned out to be 0.848, indicating very good level of reliability. The construct validity of the questionnaire was confirmed through factor analysis yielding four factors.

3.3. Procedure

To make sure about the comprehension of guidelines by the participants, the questionnaire was given to some English language teachers to comment on each item and provide suggestions for ambiguous items. After gathering information about the first draft, the items were revised so that by rewording the questions, the writer's intent was made clearer to respondents.

In order to obtain high precision of measurement, a pilot study was conducted to evaluate feasibility and identify design issues before the main research so that there would be a less chance of unreliable results. The main purpose of piloting was to evaluate the correctness of the instructions that respondents in the pilot sample followed the directions as indicated. It also indicated whether the type of survey was useful in achieving the aim of the study.

In the present study, the main approach of gathering information was through an online questionnaire. In the first section, participants were reminded that participation in this study was voluntary, and their answers were completely confidential. The questionnaire was constructed in Google Docs. A link was provided to make access to a wide range of participants. English language teachers all around the country participated in the study by just having access to the link. The link of the questionnaire was distributed through social networks such as TELEGRAM.

3.4. Data Analysis

The results were analyzed using SPSS (statistical package for social sciences, version 25). The questionnaire responses were recorded in an excel spreadsheet and then imported to SPSS for statistical analysis. Descriptive statistics including frequencies and percentage of responses for each Likert point questionnaire were calculated. The importance of different MC item guidelines from the teachers' points of view was ranked based on their responses. We used exploratory factor analysis to extract the factors underlying guidelines.

4. Results and Discussion

4.1. Results

The research question of this study aimed to explore the underlying factor structure of the attitudes of Iranian English language teachers toward MC item writing guidelines. As can be seen in Table 2, the responses to each

item are assigned into five different Likert-point types consisting of *"not important", "slightly important", "moderately important", "important", and "very important"*.

According to Table 2, EFL teachers indicated that nine guidelines were the most important guidelines in developing MC items. These guidelines included Item 16 (There must be one correct answer) (84.7%), Item 3 (Items should be edited before given to examinees) (77.0%), Item 11(If negative words are used, one of the following strategies, or a combination of them, should be used: capitalizing, italicizing, underling, and bold typing) (59.5%), Item 17 (Choices should be homogeneous in content and grammatical structure) (56.6%), Item 28 (All parts of an item should appear on the same page) (52.8%), Item 1(Grammar, punctuation, capitalization, and spelling should be correct) (51.9%), Item 15 (The position of the correct answer should be randomly assigned) (48.1%), Item 9 (Irrelevant and extra information should be avoided) (42.8%), and Item 12 (The stem should not start with a blank) (42.1%).

The statistics gathered from EFL teachers also showed that 12 MC item writing guidelines were considered important in developing MC items. These guidelines included Item 14 (The number of options depends on the number of functional distractors, but research suggests three options are adequate) (46.1%), Item 23 (It is better to use typical errors of students in developing distractors) (44.8%), Item 7 (The stem should be written in a way that, without referring to the options, examinees know immediately what the focus of the item is) (44.0%), Item 4 (Items should be worded as simply as possible) (43.3%), Item 2 (Content of each item should be independent of that of other items) (43.0%), Item 13 (The stem can be in the form of a statement or a question) (42.1%), Item 8 (Main idea should be in the stem instead of the choices) (41.1%), Item 5 (Items should be as brief as possible) (40.7%), Item 22 (All distractors (wrong options) should seem correct and plausible for examinees) (36.9%), Item 21 (Clues to the right answer, such as using specific determiners (always, never, …) in choices or grammatical inconsistencies should be avoided) (34.8%), Item 18 (Length of the choices should be equal) (34.0%), and Item 19 ("None of the above" as an option should be used carefully) (34.0%).

As it can be seen in Table 2, responses to five items indicated that the majority of teachers regarded these guidelines as either important or very important guidelines in developing MC items. These included Item 24 (Repeated words, words which are common in all options, should be included in the stem) (60.1%), Item 25 (There should be only one blank in each stem) (56%), Item 20 ("All of the above" as an option should be avoided) (55%), Item 27 ("both a and b" or "neither c nor d" in options should be avoided)

(52.1%), and Item 26 (The length of the blanks should be equal in all stems) (49%).

Table 2

EFL Teachers' Attitudes toward MC Item Writing Guidelines

	Not Important		Slightly Important		Moderately Important		Important		Very Important	
Items	F	%	F	%	F	%	F	%	F	%
1	5	0.8	26	3.9	82	12.4	205	31.0	343	51.9
2	37	5.6	47	7.1	118	17.9	284	43.0	175	26.5
3	4	0.6	9	1.4	17	2.6	122	18.5	509	77.0
4	14	2.1	25	3.8	120	18.2	286	43.3	216	32.7
5	18	2.7	46	7.0	161	24.4	269	40.7	167	25.3
6	190	28.7	105	15.9	175	26.5	138	20.9	53	8.0
7	31	4.7	60	9.1	156	23.6	291	44.0	123	18.6
8	28	4.2	31	4.7	104	15.7	272	41.1	226	34.2
9	21	3.2	45	6.8	92	13.9	220	33.3	283	42.8
10	121	18.3	112	16.9	152	23.0	169	25.6	107	16.2
11	16	2.4	19	2.9	48	7.3	185	28.0	393	59.5
12	60	9.1	53	8.0	94	14.2	176	26.6	278	42.1
13	51	7.7	43	6.5	155	23.4	278	42.1	134	20.3
14	26	3.9	57	8.6	191	28.9	305	46.1	82	12.4
15	26	3.9	25	3.8	75	11.3	217	32.8	318	48.1
16	5	0.8	9	1.4	19	2.9	68	10.3	560	84.7
17	10	1.5	15	2.3	44	6.7	218	33.0	374	56.6
18	78	11.8	52	7.9	121	18.3	225	34.0	185	28.0
19	56	8.5	49	7.4	110	16.6	225	34.0	221	33.4
20	93	14.1	78	11.8	126	19.1	184	27.8	180	27.2
21	55	8.3	61	9.2	139	21.0	230	34.8	176	26.6
22	37	5.6	48	7.3	128	19.4	244	36.9	204	30.9
23	32	4.8	45	6.8	131	19.8	296	44.8	157	23.8
24	59	8.9	66	10.0	139	21.0	185	28.0	212	32.1
25	66	10.0	63	9.5	162	24.5	191	28.9	179	27.1
26	134	20.3	76	11.5	127	19.2	170	25.7	154	23.3
27	72	10.9	74	11.2	171	25.9	165	25.0	179	27.1
28	26	3.9	25	3.8	69	10.4	192	29.0	349	52.8

Note. F = Frequency; % = Percent; N = 661

Item 10 in the questionnaire was concerned with negative words in the stem and options. In other words, the guideline says that "Stem and options should not contain negative words such as NOT or EXCEPT". According to Table 2, 41.8 % of respondents considered this guideline to be either "Important" or "Very important" in writing MC items. However, 35.2% of teachers regarded this guideline as either "Not important" or "Slightly important". The remaining 23% of participants reported this guideline as "Moderately important".

Item 6 (Choices should be arranged vertically instead of horizontally) was the only guideline which the majority of respondents reported to be either "Not important" or "Slightly important". As Table 2 indicates, 44.6% of English language teachers considered this guideline either "Not important" or "Slightly important". It is also worth noting that language teachers did not endorse moderately important option in any item.

The questionnaire was also subjected to factor analysis to identify underlying factors. The steps of doing it are described below. In order to

determine the suitability of data for factor analysis, two issues of "sample size" and "strength of the inter-correlations" should be considered. According to Pallant (2016), the results of the small samples cannot be well generalized compared to large samples. Therefore, the larger the sample, the better for data analysis. According to Tabachnick and Fidell (2001), "it is comforting to have at least 300 cases for factor analysis" (p. 613). Also, Tabachnick and Fidell admit that smaller sample sizes (e.g., 150 cases) should be sufficient if the solutions have several high loading marker variables (above .80). The total number of the participants in the current study was 661, which satisfies the above condition of the sample size.

The second issue, strength of the inter-correlations, was also satisfied by the presence of many coefficients of .3 and above (values higher than .3 indicate that each item fits well with the other items). The Kaiser-Meyer-Olkin value was also 0.871, exceeding the recommended value of 0.6. Bartlett's test of sphericity which should be 0.5 or smaller. It also reached statistical significance (p = .000), supporting the factorability of data (see Table 3).

Table 3

KMO and Bartlett's Test for Suitability of Data

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.871
Bartlett's Test of Sphericity	Approx. Chi-Square	3506.928
	df	378
	Sig.	.000

After making sure the data were suitable for factor analysis, we turned to extracting the minimum number of factors which represent the interrelations among the variables. As a rule of thumb, in the current study, Eigen values greater than 1 and the scree plot (see Appendix B) were used to determine the number of factors. Initially, eight factors were identified. To ensure the factors were correctly identified, we also used parallel analysis. Three pieces of information were necessary to do parallel analysis. They are (a) the number of variables (in our case, 28 items); (b) the number of participants (in our case, 661); and (c) the number of replications (the program default requires 100). Then, the eigenvalues obtained from SPSS for eight factors were compared with the corresponding values generated by parallel analysis. Finally, those values which were larger than the criterion values from parallel analysis were kept and the other four factors needed to be excluded. Results are summarized in Table 4. As can be seen, only four of eight factors were accepted.

Four-factor solutions were examined based on principal components analysis. The four-factor solution explained a total of 36.965% of the variance, with Factor 1 contributing 20.322%, Factor 2 contributing 6.338%,

Factor 3 contributing 5.634%, Factor 4 contributing 4.671% to the total variance (see Appendix C).

Table 4

Comparison of Eigenvalues from the PCA and Criterion Values from Parallel Analysis

Component number	Eigenvalues from PCA	Criterion values from parallel analysis	Decision
1	5.690	1.4007	Accepted
2	1.775	1.3465	Accepted
3	1.577	1.3061	Accepted
4	1.308	1.2659	Accepted
5	1.178	1.2308	Rejected
6	1.104	1.2016	Rejected
7	1.066	1.1702	Rejected
8	1.015	1.1438	Rejected

A four-factor solution was rotated for this study, then based on it, some new information was taken into account. In this step, the Component Correlation Matrix Table was checked in order to show the strength of relationship between the factors. Based on the statistics presented in Table 5, Factor 1 and Factor 2 (r = 0.186), Factor 1 and Factor 3 (r = 0.326), Factor 1 and Factor 4 (r = 0.173), Factor 2 and Factor 3 (r = 0.150), Factor 2 and Factor 4 (r = 0.167), and Factor 3 and Factor 4 (r = 0.127) are correlated.

Table 5

Component Correlation Matrix

Component	1	2	3	4
1	1.000	.186	.326	.173
2	.186	1.000	.150	.167
3	.326	.150	1.000	.127
4	.173	.167	.127	1.000

Extraction Method: Principal Component Analysis

As the correlation among the components is low, the solutions from both Varimax and Oblimin rotation are similar (Pallant, 2016). Using Oblimin rotation, we received three tables of Component Correlation Matrix, Pattern Matrix, and Structure Matrix. Table 5 shows information for Component Correlation Matrix. The Pattern Matrix (showing the factor loadings of each of the variables) and the Structure Matrix (providing information about the correlation among variables and factors) are shown below in Table 6 and Table 7.

Table 6

Pattern Matrix for the PCA with Oblimin Rotation of four-Factor Solution of MC Item-Writing Guidelines Questionnaire.

	Component			
	1	2	3	4
I27	.774	.038	-.133	-.010
I20	.679	-.053	-.078	.148
I21	.635	-.011	-.047	.239
I25	.589	.017	.170	-.210
I26	.548	-.074	.169	-.098
I18	.541	.192	.123	-.022
I12	.494	.100	.216	-.157
I10	.485	-.081	.359	-.177
I28	.397	.338	-.134	.047
I19	.396	-.009	-.004	.302
I24	.362	.148	.061	.242
I16	-.035	.682	-.071	-.134
I3	-.184	.582	.135	.210
I17	.228	.524	.080	-.024
I11	.129	.500	.179	-.053
I15	.199	.356	.085	.233
I5	-.068	.134	.656	-.078
I6	.125	-.389	.604	-.038
I4	-.101	.187	.596	.096
I7	-.110	.129	.505	.086
I13	.181	-.114	.456	-.024
I14	.183	-.071	.391	.182
I8	.033	.017	.301	.167
I9	.170	.223	.266	-.040
I2	.062	-.157	.108	.635
I1	-.179	.049	.063	.568
I22	.350	.092	.043	.429
I23	.249	.086	-.016	.320

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.
a. Rotation converged in 10 iterations.

Table 7 shows the correlation between Factors and Items. This table confirms the accuracy of the Item loadings under each Factor. As it can be seen, the correlation of each Item is shown with all the Factors. The higher correlation of each Item with one of the factors shows that the Item should be under that Factor. Table 6 shows the items loading on the four factors with 11 items (Item 27, Item 20, Item 21, Item 18, Item 25, Item 26, Item 10, Item 12, Item 24, Item 19, and Item 28) loading above .3 on Component 1; with five Items (Item 16, Item 3, Item 17, Item 11, and Item 15) loading on Component 2; with eight Items (Item 5, Item 4, Item 6, Item 7, Item 13, Item 14, Item 9, and Item 8) on Component 3; and finally, with four Items (Item 2, Item 1, Item 22, and Item 23) loading on Component 4.

Table 7

Structure Matrix for the PCA with Oblimin Rotation of Four-Factor Solution of MC Item-Writing Guidelines Questionnaire.

	Factors			
Items	1	2	3	4
I27	.736	.161	.124	.114
I20	.669	.087	.154	.247
I21	.659	.140	.188	.341
I18	.613	.308	.325	.119
I25	.611	.117	.338	-.084
I26	.573	.037	.325	.006
I10	.556	.034	.483	-.061
I12	.556	.198	.372	-.027
I24	.452	.265	.231	.337
I19	.445	.114	.162	.368
I28	.424	.399	.051	.155
I16	.046	.642	.002	-.035
I3	.005	.603	.189	.293
I17	.347	.574	.229	.113
I11	.271	.542	.289	.075
I15	.333	.445	.232	.337
I5	.157	.206	.644	.016
I4	.145	.273	.603	.186
I6	.242	-.282	.581	-.005
I7	.093	.198	.499	.152
I13	.304	-.016	.495	.046
I14	.328	.052	.463	.252
I9	.291	.287	.349	.061
I8	.163	.096	.336	.214
I2	.178	-.023	.186	.634
I1	-.051	.120	.084	.553
I22	.455	.235	.226	.511
I23	.315	.184	.118	.375

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Highest loadings on each factor were used to label factors. Item 27 ("both a and b" or "neither c nor d" in options should be avoided)(77 %), Item 20 ("All of the above" as an option should be avoided) (67 %), and Item 21(Clues to the right answer, such as using specific determiners (always, never, …) in choices or grammatical inconsistencies should be avoided) (63%) had the highest loadings on Factor 1, so it was labelled "Developing plausible distractors".

Item 16 (There must be one correct answer) (68%), Item 3 (Items should be edited before given to examinees) (58 %), Item 17 (Choices should be homogeneous in content and grammatical structure) (52 %), and Item 11 (If negative words are used, one of the following strategies, or a combination

of them, should be used: capitalizing, italicizing, underling, and bold typing) (50%) had the highest loadings on Factor 2, so it was reasonable to label it "Editing and proofreading items".

Item 5 (Items should be as brief as possible) (65 %), Item 6 (Choices should be arranged vertically instead of horizontally) (60 %), Item 4 (Items should be worded as simply as possible) (59 %) and Item 7 (The stem should be written in a way that, without referring to the options, examinees know immediately what the focus of the item is) (50 %) were loaded highly on Factor 3, so it was labeled "Formatting and style of MC items".

Finally, Factor 4 was labeled "avoiding clues to the correct response" based on the following highest Item loadings: Item 2 (Content of each item should be independent of that of other items) (63 %) and Item 1(Grammar, punctuation, capitalization, and spelling should be correct) (56 %).

4.2. Discussion

The first finding of the study showed that nine guidelines were very important. These guidelines are mostly related to the style of MC items: Writing only one correct option, edition of items, specifying the negative words, homogenous choices, writing all parts of an item on the same page, using correct grammar and punctuation, assigning the correct option randomly among choices, removing irrelevant information in items, and avoidance of using blanks at the beginning of the stem. This finding is consistent with previous research regarding the high priority of these guidelines (Frey et al., 2005; Haladyna, et al, 2002; Haladyna, 2004; Kubiszyn & Borich, 2013).

MC items should be edited before given to examinees to reduce the possibility of any errors in constructing MC items. It is very important to follow this guideline because items containing errors may cause examinees to fail to get the item correct (Haladyna, 2004). Also, as Osterlind (2002) noted, an MC test should be edited because if an item is not constructed carefully, it may act as a hint for examinees. Furthermore, since item writers have to develop several plausible distractors in MC items, the possibility of making mistakes and giving clues in distractors increases. Therefore, items should be double checked to prevent giving any clues to the right answer.

Negative words in the stem should be specified to minimize the possibility of mistakes in responding to the items. Haladyna (2004) noted that since negative words require examinees to do the opposite, they should be specified to keep examinees alert. Similarly, Brame (2014) asserted that examinees are generally supposed to choose the correct option while in negative items examinees must do the reverse and find the wrong option. Therefore, negative words should be highlighted to alert the examinees (Haladyna & Rodriguez, 2013). If negative words are not highlighted,

examinees may make mistakes and do not answer items correctly despite having sufficient knowledge (Osterlind, 2002).

Including all parts of an item on the same page should be given sufficient attention. It is very important to follow this guideline because students may get confused if there is a gap between the stem and option. If stem and options are not constructed on the same page, examinees have to look back and forth, wasting time and being frustrated (Oermann, 2013)

Options should be homogeneous in content and grammatical structure. Using heterogeneous options may act as a hint or a clue for examinees. Examinees with limited knowledge may choose or remove some options if they are not homogeneous in content or grammatical structure. Using similar options in content and grammatical structure increases item difficulty and item discrimination of items (Ascalon, Meyers, Davis, & Smits, 2007; Kubiszyn & Borich, 2013). Grammar, punctuation, capitalization, and spelling should be correct. The incorrect grammar or punctuation in items may remain as a wrong template in examinees' minds and distract them.

Irrelevant and extra information should be avoided. Stem and options should be brief enough, and they should stick to the point without extra or irrelevant information so that examinees can comprehend the question and look for the solution in options (Osterlind, 2002). This guideline helps examinees to get the exact intention of the item and decreases the reading time for examinees and item development time for test constructors (Haladyna & Downing, 1989a).

The last guideline which was considered very important in the present study has to do with the position of blanks in the stem. The guideline emphasizes that the stem should not start with a blank. Haladyna (2004) provided some reasons for the significance of this guideline. Haladyna explained that putting blanks at the beginning of the stem increases the time of item development. Also, this format makes it difficult to read the stem and reduce the time for answering other items. Furthermore, by putting the blank at the beginning of the stem, examinees move from unknown information to known information which is not compatible with the theory of cognitive code learning (Farhady, Jafarpoor, & Birjandi, 1994).

The second finding of the present study was that 12 MC item-writing guidelines were considered important in developing MC items. As Haladyna et al. reported, most of these guidelines were supported in previous studies. For example, 100 percent of authors have cited that the main idea should be in the stem. Making distractors plausible (96 %), avoiding clues (96 %), equal length of choices (85 %), and clear direction in the stem (82 %) are other guidelines cited in textbooks and articles. However, a controversy also existed among these guidelines. For example, in the present study, 34% of

respondents regarded Item 19 ("None of the above" as an option should be used carefully) important and about 33% of respondents reported this guideline as very important in constructing MC items. However, in the analysis of Haladyna et al. about 44% of various sources regarding MC item-writing guidelines cited and supported this guideline, 7% did not cite this guideline, and 48% were against this guideline. The discrepancies in this guideline refer to different testing practices in different countries. For example, various studies (Crehan & Haladyna, 1991; Crehan et al., 1993; Frary, 1991) have reported no difference in item discrimination. However, Rich and Johanson (1990) pointed out that item discrimination and item difficulty increase with the use of "none of the above" option. However, Gross (1994) stated that "any stem or option format that by design diminishes an item's ability to distinguish between candidates with full versus misinformation, should not be used" (p. 125).

The third finding of this study showed that five guidelines were either important or very important in teachers' point of view. Several previous studies have reported the reasons for the significance of these guidelines, which the findings of this study confirm. Repeated words, words which are common in all options, should be included in the stem so that examinees have less trouble in answering MC items with short options (Hansen & Dexter, 1997).

Some studies provided several reasons for avoidance of 'all of the above" (AOTA) option. Some researchers believe that test-wise students can choose the AOTA option if they know at least two of the options are correct (Hansen & Dexter, 1997; Osterlind, 2002). Also, examinees can remove this option if they know that one of the options is incorrect. In fact, this option provides clues to the examinees. The main reason for avoiding AOTA option is also true for pair options. The use of "both A and B" or "neither A nor B" may run the risk of guessing. Examinees can use the logical relationship between options to remove some options and get to the right answer (Kubiszyn & Borich, 2013). If an item contains pair options, examinees with partial knowledge of materials can compare the options and find the correct option without having enough knowledge. Finally, the last guideline for the third finding referred to the length of the blanks. Schrock and Mueller (1982) stated that it is important to follow this guideline because ignorance of this guideline can give a clue to the right answer. Also, unequal length of the blanks may cause examinees to compare options and conclude that shorter blanks require shorter responses and longer blanks need longer responses (Holt & Kysilka, 2006).

The fourth finding of the study was that about 58.2 % of teachers do not regard the avoidance of negative words in the stem as an important guideline in constructing MC items. This finding supports the results of some

of those of previous studies (Ellsworth et al., 1990; Hansen & Dexter, 1997; Tarrant et al., 2006; Tarrant & Ware, 2008). According to these studies, one of the most frequent violations in MC item-writing guidelines has to do with using negative words in the stem. Hansen and Dexter (1997) reported that the most common problem in MC items was related to negative wording in the stem. Although some studies (Brame, 2014; Williams, 1984) have reported that the use of negative words in the stem makes it difficult for examinees to respond to the item, several other studies concluded that there is no difference in item difficulty and item discrimination of negatively worded items compared to positively worded ones (Rachor & Gray, 1996). Furthermore, Harasym et al. (1992) explained that the negative stem does not affect reliability

The sixth finding indicated that the majority of English language teachers regard the vertical formatting of options as the least important guideline. Some of the previous studies seem to have reported a similar finding. For example, Haladyna et al. (2002) reported that about 52% of various textbooks and articles did not cite this guideline and 11% of authors were against this guideline, as shown in Table 8. This finding showed that both textbook authors and EFL teachers do not regard this guideline important in constructing MC items. Writing options vertically occupies more space and has no effect on students' scores (Haladyna, 2004). Haladyna and Rodriguez (2013) pointed out that despite some advantages in vertical format, it is not cost-efficent for long MC tests.

Finally, fidnings from factor analysis regarding the four-factor solution of the guidelines partially confirm Haladyna et al.'s (2002) revised taxonomy of guidelines into five categories, notably, formatting concerns, style concerns, writing the stem, and writing the choices. Content concerns are not language-related guidelines, which is why we did not include items relating to content. The four-factor solution of our findings suggests that guidelines for the construction of effective MC items should focus on issues that address strategies to avoid providing hints in items, developing the most plausible distracters to function well, and using correct language in items to prevent examinees from fossilizing wrong structures.

5. Conclusion and Implications

In the present study, we set out to explore the significance of item writing guidelines, using a researcher-made questionnaire. Results from descriptive statistics showed that Iranian English language teachers generally considered the guidelines to be important for effective MC item writing. Findings of the factor analysis revealed four major underlying factors.

Given the findings of the study, it may be safe to conclude that Iranian language teachers are aware of the significance of the guidelines, but whether

they also use them to produce more effective MC items is another issue which needs further examination. The very fact that Iranian language teachers considered the majority of the guidelines to be either very important or important highlights the importance they attach to such guidelines. Such awareness may contribute to teacher autonomy in that they will be more autonomous in developing more effective MC item types. We, therefore, suggest that guidelines be carefully edited and proofread before they are handed down to language teachers through textbooks, research journals, or whatever medium best suited for this endeavor.

The findings may offer new insights into how they can help the Iranian EFL teachers to better understand the extent to which their MC tests conform to MC item-writing guidelines and shed light on their strengths and weaknesses in constructing MC items. Language teachers may benefit from the findings of this study and modify the method of constructing MC items based on MC item-writing guidelines. They can become aware of MC item-writing guidelines and try to follow these guidelines in order to enhance the quality of their MC tests. Moreover, the results of this study can raise teachers' consciousness about the significance of various MC guidelines in constructing MC tests.

Like other studies, the present study has its own limitations. The first limitation was that the present study was purely quantitative, so we cannot use the findings to help us know why certain guidelines were more important than others. The second limitation had to do with the sampling procedure used in the present study. We employed a nonprobability sampling procedure. Using a probability sampling procedure like stratified random sampling could better represent the sample from the population.

References

Ascalon, M. E., Meyers, L.S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*(2), 153-170.

Bailey, K. M. (2018). Multiple-choice item format. In *Encyclopedia of English Language Teaching* (pp. 1-8). John Wiley & Sons.

Brame, C. J. (2014). *Writing good multiple choice test questions*. Nashville, TN: Vanderbilt University Center for Teaching. Retrieved from http://cft.vanderbilt.edu/guides-subpages/writing-good-multiple-choice-test-questions/

Burton, J. S., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty.* Department of Instructional Science, Brigham Young University Testing Services. Retrieved August 16, 2009, from http://testing.byu.edu/info/handbooks/betterItems.pdf

Cohen, R. J., & Swerdlik, M. E. (1999). *Psychological testing and assessment: An introduction to tests and measurement* (4th ed.). Mountain View, CA: Mayfield.

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.

Dörnyei, Z. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed). New York, NY: Routledge.

Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice test items: What are textbook authors telling teachers? *The Journal of Educational Research*, *83*(5) 289-293. doi: 10.1080/00220671.1990.10885972

Farhady, H., Jafarpoor, A., & Birjandi, P. (1994). *Testing language skills: From theory to practice*. Tehran: SAMT Publication.

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

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.

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: The case of "none of the above." *Evaluation and the Health Professions, 17*(1), 123–126.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1(*1*),* 51–78.

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), 309-333.

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

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are multiple-choice items too fat? *Applied Measurement in Education, 32*(4), 350-364,

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

Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple – choice items. *Evaluation and the Health Professions, 15*(2), 198-220.

Holt, L. C., & Kysilka, M. (2006). *Instructional patterns: Strategies for maximizing student learning*. Thousand Oaks, CA: Sage Publication.

Kiss, H. J., & Selei, A. (2017). Do streaks matter in multiple-choice tests? *Education Economics, 26*(2), 179-193.

Kubiszyn, R., & Borich, M. (2013). *Educational testing and measurement-classroom application and practice.* Jefferson City, MO: Wiley.

Lennox, B. (2009). Multiple choice. *Medical Education, 1*(5), 340-344.

Linn, R., Baker, E., & Dunbar, S. (1991) Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8**)**, 15–21.

Liu, O. L., Lee, H. S., & Linn, M. C (2011) An investigation of explanation multiple-choice items in science assessment, *Educational Assessment, 16*(3), 164-184.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology.* New York: Harcourt Brace.

Moreno, R., Martı´nez, R. J., & Mun˜iz, 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.

Nitko, A. J. (1985). Review of Roid and Haladyna's a technology for test item writing. *Journal of Educational Measurement, 21*, 201–204.

Oermann, M. H. (2013). *Teaching in nursing and role of the educator: The complete guide to best practice in teaching, evaluation, and curriculum development.* New York, NY: Springer.

Osterlind, S. J. (2002). *Constructing test items: Multiple-choice, constructed-response,performance, and other formats.* New York, NY: Kluwer Academic.

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (6th ed.). Berkshire: McGraw-Hill Education.

Paxton, M. (2000). A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education, 25*(2), 109-119.

Rachor, R. E., & Gray, G. T. (1996, April). *Must all stems be green? A study of two guidelines for writing multiple choice stems.* Paper presented at the annual meeting of the American Educational Research Association, New York.

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.

Richichi, R.V. 1996. An analysis of test-bank multiple-choice items using item-response theory. Research report. Accessed 11 May 2010 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/5d/eb.pdf.

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155–1159.

Schrock, T. J., & Mueller, D. J. (1982). Effects of violating three multiple-choice item construction principles. *The Journal of Educational Research*, *75*(5) 314-318.

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.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed). New York: HarperCollins.

Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education*, *12*(5), 563-573. doi: 10.1080/0950069900120508

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *6*(6), 662-671.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198-206.

Torres, C., Lopes, A., Babo, L., & Azevedo, J. (2011). Improving multiple-choice questions. *US-China Education Review,Education Theory 1*(1), 1-11.

Vacc, N. A., Loesch, L. C., & Lubik, R. E. (2001). Writing multiple-choice test items. In G. Walz & J. Bleuer (Eds.) *Assessment: Issues and Challenges for the Millennium.* CAPS: Greensboro, NC.

Wainer, H.,Wadkins, J.R.J., & Rogers, A. (1983). *Was there one distractor too many?* Princeton, NJ: Educational Testing Service.

Williams. (1984). Multiple choice tests and the computer. *Journal of Further and Higher Education*, *8*(3), 53-73. doi: 10.1080/0309877840080305

Wilson, M., & Wang, W. C. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement*, *19*(1), 51–71.

Appendixes

Appendix A. Attitudinal questionnaire regarding Significance of Multiple-choice Guidelines

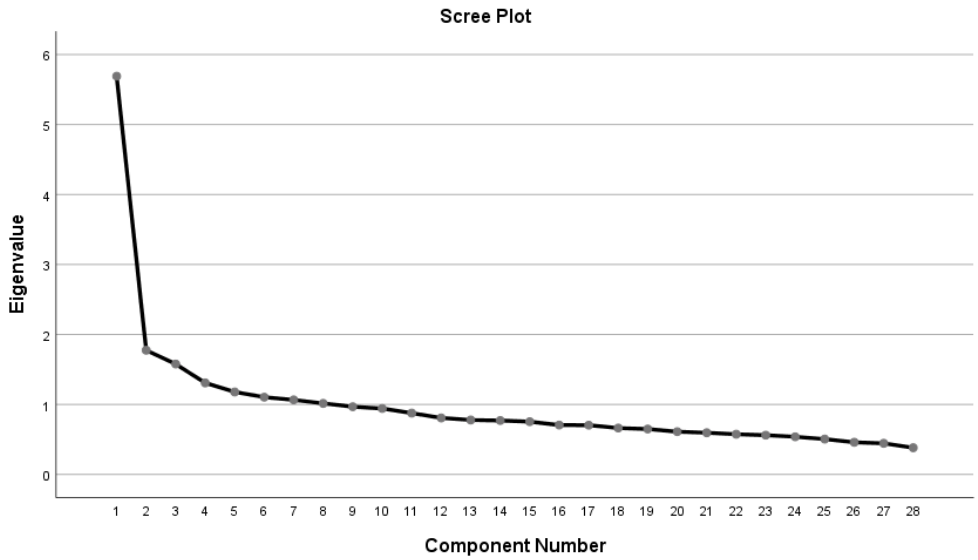
The following questionnaire consists of 28 guidelines for constructing multiple-choice items. It helps us to seek the opinions of Iranian English language teachers about the significance of writing effective multiple-choice items. Please, read each statement and provide your answers by clicking on the appropriate option. This is not a test, so there are no "right" or "wrong" answers, and you do not even have to write your name on it. The results of the questionnaire will be used for research purposes only, so please provide answers as thoughtfully as possible. Thank you very much for your help!

Gender: male female
Age:
Years of teaching experience:
Degree: BA holder MA holder PhD holder
Field of study:
English Language Teaching English Literature Translation studies Linguistics
Affiliation:
Ministry of Education
Ministry of science, research, and technology
Other (Please, indicate):

	Very important (5)	Important (4)	Moderately Important (3)	Slightly Important (2)	Not Important (1)
1. Grammar, punctuation, capitalization, and spelling should be correct.					
2. Content of each item should be independent of that of other items.					
3. Items should be edited before given to examinees.					
4. Items should be worded as simply as possible.					
5. Items should be as brief as possible.					
6. Choices should be arranged vertically instead of horizontally.					
7. The stem should be written in a way that, without referring to the options, examinees know immediately what the focus of the item is.					
8. Main idea should be in the stem instead of the choices.					
9. Irrelevant and extra information should be avoided.					
10. Stem and options should not contain negative words such as **NOT** or **EXCEPT**.					
11. If negative words are used, one of the following strategies, or a combination of them, should be used: capitalizing, italicizing, underling, and bold typing.					
12. The stem should not start with a blank.					
13. The stem can be in the form of a statement or a question.					
14. The number of options depends on the number of functional distractors, but research suggests three options are adequate.					

15. The position of the correct answer should be randomly assigned.
16. There must be one correct answer.
17. Choices should be homogeneous in content and grammatical structure.
18. Length of the choices should be equal.
19. **"*None of the above*"** as an option should be used carefully.
20. ***"All of the above"*** as an option should be avoided.
21. Clues to the right answer, such as using specific determiners (always, never, …) in choices or grammatical inconsistencies should be avoided.
22. All distractors (wrong options) should seem correct and plausible for examinees.
23. It is better to use typical errors of students in developing distractors.
24. Repeated words (words which are common in all options) should be included in the stem.
25. There should be only one blank in each stem.
26. The length of the blanks should be equal in all stems.
27. "***both a and b"*** or "***neither c nor d"*** in options should be avoided.
28. All parts of an item should appear on the same page.

Appendix B: Scree Plot for Factors



Appendix C. Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.690	20.322	20.322	5.690	20.322	20.322
2	1.775	6.338	26.660	1.775	6.338	26.660
3	1.577	5.634	32.294	1.577	5.634	32.294
4	1.308	4.671	36.965	1.308	4.671	36.965

Extraction Method: Principal Component Analysis.

Bibliographic information of this paper for citing:

Ganji, M., & Esfandiari, R. (2020). Attitudes of language teachers toward multiple-choice item writing guidelines: An exploratory factor analysis. *Journal of Modern Research in English Language Studies, 7*(3), 115-140.