International Journal of Language Testing
Vol. ٦, No. ١, March ٢٠١٦ ISSN ٢٤٧٦-٥٨٨٠

Properties of Single-Response and Double-Response Multiple-Choice Grammar Items

Purya Baghaei[١], Alireza Dourakhshan[٢]

Received: ٢١ October ٢٠١٥ Accepted: ٧ January ٢٠١٦

Abstract

The purpose of the present study is to compare the psychometric qualities of canonical single-response multiple-choice items with their double-response counterparts. Thirty, two-response four-option grammar items for undergraduate students of English were constructed. A second version of the test was constructed by replacing one of the correct replies with a distracter. The two test forms were analysed by means of the Rasch model. To score double-response items, two scoring procedures, dichotomous and polytomous, were applied and compared. Results showed that dichotomously scored double-response items were significantly harder than their single-response counterparts.Double-response itemshad equal reliability to the single-response items and had a better fit to the Rasch model. Principal component analysis of residuals showed that double-response multiple-choice items are closer to the unidimensionality principle of the Rasch model which can be the result of minimized guessing effect in double-response items.Polytomously scored double-response items, however, were easier than single-response items but substantially more reliable. Findings indicated that polytomous scoring of double-response items without the application of correction for guessing formulas results in the increase of misfitting or unscalable persons.

Keywords: *Multiple-choice items, Multiple-response multiple-choice items, Guessing effect*

١. Introduction

Multiple-choice (MC) method is a very popular test format in large scale and classroom testing. The reason for the popularity of MC tests is the advantages which are associated with them. Some of these advantages are:

a. MC items are very flexible in measuring different types of objectives.
b. Since examinees can answer the items by just ticking the correct option a lot of testing time is saved and therefore a large domain of the content can be included in the test. This increases the validity of test.
c. Scoring is very objective and reliable with MC items.
d. MC items are easy to score either by machines or human beings and easily lend themselves to item analysis.

[١] *English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran. Email:* baghaei@mshdiau.ac.ir.
[٢] *English Department, Farhangian University, Mashhad, Iran.*

Two major drawbacks of MC items which have been mentioned by many researchers are that (١) MC items trigger a high degree of guessing among examinees and (٢) they are easy to cheat. Guessing and cheating are two important threats to the validity and reliability of tests. If examinees answer questions without knowledge of the content of items and get them right just as a result of guessing and cheating then the assessment is not valid, reliable and fair.

In order to solve the two problems of guessing and cheating in MC tests researchers have suggested a type of MC items which is called multiple-response multiple-choice items (MRMC) (Kubinger&Gottschall, ٢٠٠٧). In this method MC items are constructed with several options and more than one correct response. Examinees are instructed to mark all the correct responses in each item. The chances that an examinee marks all correct replies by chance only considerably diminish in these types of items. Furthermore, the chances of copying all correct options from another examinee diminish too.

Multiple response multiple-choice items were first introduced by Orleans and Sealy (١٩٢٨) who called it *multiple-choice plural response* items. They argued that MRMC items can be used to test a wide range of abilities from rote knowledge to complex cognitive reasoning. They also discussed the problems involved in scoring such items.

Dressel and Schmid (١٩٥٣) compared the psychometric qualities of MRMC items with single response MC items. They found that MRMC items enjoyed higher reliability. They stated that MRMC items are superior to canonical single response MC items because they allow measurement of partial knowledge and therefore contribute to finer discrimination and enhance the validity of tests.

Ma (٢٠٠٤) argues that the call for authentic and valid measurement of students' abilities revived MRMC items as an alternative to performance assessment in the ١٩٨٠'s. It was argued that MRMC format has most of the advantages of MC items and at the same time allows for the economic assessment of higher order skills and therefore is an alternative to costly performance assessment. To this end, MRMC items were used in the Kansas State Assessment Program to test reading and mathematics. Pomplun and Omar (١٩٩٧) analysed the data of this test and concluded that MRMC technique had adequate reliability and validity and because of its ease of scoring is a promising technique.

Page, Bordage, and Allen (١٩٩٥) used MRMC items in the Canadian Qualifying Examination in Medicine (CQEM). CQEM is a licensing test taken by all who want to practice medicine in Canada. The test is composed of clinical scenarios followed by several questions. Page et al. (١٩٩٥) concluded that MRMC items are valid and reliable tests of clinical problem solving skills and should be considered by medical testing professionals. They also state that the technique has been employed by the American College of Physicians and other medical schools.

In MRMC items two scoring methods are possible. The first scoring method is dichotomous. In this method examinees are given a point on an item if all correct options and none of the distractersare marked. If an examinee marks one of the distracters then the item is scored as wrong, regardless of the number of correct options marked. In the second method, partial credit is given if one or more correct options are marked even if some incorrect options are marked too. When the latter scoring method is used some formulas are applied to penalize for guessing. Ma (٢٠٠٤) provides a list of the existing scoring techniques and formulas used currently. In one formula the number of incorrect options marked is subtracted from the number of correct options marked to penalize for guessing. In another, the number of correct options

marked is divided by the number of correct options. Or the number of correct options marked is added up with the number of incorrect options unmarked. And, as mentioned before, in the simplest scoring procedure the item is scored correct if all correct options and none of the incorrect options are marked.

It is possible to administer the MRMC tests in two ways: examinees can either be informed of the number of correct answers in each item in advance or they can be asked to mark all the correct answers without mentioning the number of correct answers in each item. Of course the result of these two distinct methods will be cognitively different since knowing or not knowing the number of correct responses can affect the guessing process.

Clearly, MRMC tests are more challenging than the canonical MC tests and the reason according to Dressel and Schmid (١٩٥٣) is that, "A student may be forced to see not only the relationship existing between the stem and the responses but also to reconstruct his thinking as he looks at each response in relationship to the other responses of the item" (p.٥٨١). That is why it is likely that MRMC tests are more reliable, valid and representative of examinees' real knowledge than single-response MC tests. Pomlun and Omar (١٩٩٧) investigated the psychometric properties of the multiple-mark items and concluded that there is adequate reliability (٠.٧٣ to ٠.٧٨) and validity evidence to support the use of MRMC format, and, because of its desirable features (e.g., allowing multiple correct answers and ease of scoring), it is a promising item format for use in state assessment programs.

In addition, the recent developments in computer based testing programs also consider MRMC testing formats as innovative. In their investigation of the feasibility of using innovations such as graphics, sound and alternative response modes in computerized tests, Pashall, Stewart and Ritter (١٩٩٦) devoted a section of their study to the evaluation of MRMC format and concluded that the psychometric functioning of the various item types appeared adequate. They suggest that future research on MRMC format should focus on the effects of guided instruction. (e.g., "select the best ٣"). In different test situations, whether classroom test or large scale standardized tests, or whether paper and pencil testing or computer- based testing, MRMC format can be used with ease and confidence.

MRMC items have been compared with single response format and constructed response format in a number of studies (Hohensinn&Kubinger, ٢٠٠٩; Kubinger&Gottschall, ٢٠٠٧; Kubinger, Holocher-Ertl, Reif, Hohensinn, &Frebort, ٢٠١٠). In these studies four test formats were compared: single response MC format with six options, multiple-response (MR) format with five options and two correct replies, MR format with five options and an unknown number of correct replies, and constructed response format.

Kubinger and Gottschall (٢٠٠٧) examined a type of multiple choice items, called, the "x of ٥" MC items. In this format the items have five options with multiple correct answers. In order to get a point for an item the test-takers have to mark all the correct options and none of the wrong ones. The number of correct options can vary across the items. Kubinger and Gottschall (٢٠٠٧) demonstrated that the 'x of ٥' format is significantly more difficult than single response MC items with six options. They also showed that 'x of ٥' is as difficult as the constructed response format. The lower difficulty of single-response MC items compared to MRMC items was attributed to the large guessing effect which is involved in replying them.

Hohensinn and Kubinger (٢٠٠٩) demonstrated that the three response formats of '١ of ٦', '٢ of ٥' and constructed response format measure the same construct and the response format

does not affect what the test measures. Kubinger, Holocher-Ertl, Reif, Hohensinn, and Frebort (٢٠١٠) compared two MC formats of (‘١ of ٦’) and (‘٢ of ٥’) with free-response format in a mathematics test. Only if examinees had marked both correct answers and none of the distracters in the latter format the items were scored as correct. Kubinger et al. (٢٠١٠) demonstrated that the constructed response format and the ‘٢ of ٥’ were significantly harder than the ‘١of ٦’. The free-response format was slightly harder than the ‘٢ of ٥’, but not statistically significantly. Kubinger et al. (٢٠١٠) conclude that the reason why ‘١ of ٦’ format was easier than the ‘ ٢ of ٥’ was the large degree of guessing that is involved in answering single response MC items even when there are five distracters and recommend double or multiple-response MC items to eliminate guessing effect.

The introduction of MRMC items was one pragmatic approach to solve the guessing problem in MC tests. There are some psychometric methods as well. To overcome guessing and improve model fit experts have focused on Item Response Theory (IRT).IRT has developed an approach aimed to overcome guessing effects by accounting for it by adding an extra parameter to the IRT model. The ٣-PL IRT model (Birnbaum, ١٩٦٨), provides a person ability parameter and an item difficulty parameter, as well as an item discrimination parameter and a guessing parameter. Kubingerand Draxler(٢٠٠٦) have recently devised the Difficulty plus Guessing-PL model, which is simply the ٣-PLmodel without an item discrimination parameter. The basic assumption behind IRT is that if a test taker fails to identify the correct option in a rather simple test but manages to identify the correct option in a rather difficult test, chances are that he has done it on the basis of a good luck. In order to estimate the examinee’s ability parameter, different IRT models have been presented, the most important of which include: the well-known ٣-PL and ٢-PL models (Birnbaum, ١٩٦٨) and the Rasch model (Rasch, ١٩٦٠). The ٣-PL model takes into account that any correct response to an item might be due to an item-specific guessing effect. Implementation of these models would obviously be the optimal approach from a psychometric perspective. The other approach is the investigation of person-fit indices which flag lucky guessers as unscalable respondents (Baghaei&Amrahi, ٢٠١١). However, these psychometric models are very complicated and not economical especially for medium-stakes tests.

Considering single-response and MRMC items the fundamental question which arises is whether MRMC items are equivalent to canonical single response items in terms of what they measure? Which one is psychometrically superior?Ispolytomous scoring of MRMC items superior to dichotomous scoring?The purpose of the present study is to address these questions.

٢. Method

٢.١ *Instrument*

Forty, four-option two-response multiple-choice(TRMC) grammar items were developed for freshmen undergraduate students of English as a foreign language. A parallel version of the test was constructed with four options and one correct responseby replacing one of the correct replies with a distractor. The stems of the items remained intact; only one correct response in the two-response test was replaced with a distractor to construct the single-response multiple-choice (SRMC) form. The following are examples of two TRMC items and their single-responsecounterparts.

١. The papers _____ by the instructor himself.

a. are being corrected b. will have corrected
c. are supposed to be corrected d. would often correct

٢. The head of the state _____ welcomed by the mayor.

a. had been b. has just c. was d. is going to

٢. The papers _____ by the instructor himself.

a. have just corrected b. are being corrected
c. will have corrected d. would often correct

٢*.The head of the state _____ welcomed by the mayor.

a. was being to b. had been c. has just d. is going to

The first ١٠ items in both tests were the same canonical single-response MC items. Ten items were kept identical in both forms to be used as anchor items to equate the two forms. The items in the two forms were ordered in the same sequence.

٢.٢ *Participants*

The two test forms were randomly distributed among ١٥٤ first year undergraduate students of English as a foreign language in two universities in Mashhad, Iran. The test results were used by the instructors as midterm exam scores for the summer semester of ٢٠١٢ and the test-takers were informed of this prior to testing. Seventy-nine students took the SRMC testand ٧٥ took the TRMC test.

٣. Data analysis

The two test forms, i.e., single-response and double-response multiple-choice tests were linked by means of ١٠ common items to make the application of concurrent common item equating possible. The data were analysed with WINSTEPS Raschprogramme (Linacre, ٢٠٠٩) in a one-step procedure to estimate the item and person parameters form the two forms on a common scale.

To make sure of the accuracy of the equating procedure the quality of anchor items were checked. The difficulty estimates of anchor items from separate calibrations of the wo forms should not change much after they are brought onto the same scale. Items whose difficulties change drastically should not be used as anchor items (Baghaei, ٢٠١١; Linacre, ٢٠٠٩). A graphical check was carried out to make sure of the stability of anchor item parameters. The difficulty estimates of the items from the two calibrations were cross-plotted. Items which fall far from an identity line are dropped from the analysis (Linacre, ٢٠٠٩). Examination of the

difficulty of anchor items across forms showed that all items had kept their difficulty estimates in the two analyses. The ١٠ common items were used for linking and equating purposes only and were dropped from further analyses and comparisons.

٣.١ *Dichotomous scoring*

Double-response items were scored dichotomously in the first phase of the study, i.e., an item was considered correct and was scored ١ if test-takers had marked both correct replies and none of the distractors. As explained above the two test forms were linked by means of ١٠ well-functioning common items. The connected data were analysed with WINSTEPS Raschprogramme (Linacre, ٢٠٠٩) in a concurrent common item equating design. Therefore, the difficulty estimates of all items from the two forms and the ability parameters of all the persons who had taken the two forms could be estimated on a single scale.

Fit statistics showed that all items in the SRMC test had acceptable infit and outfit mean square values between .٧٠ to ١.٣٠ (Bond & Fox, ٢٠٠٧). Only Item ٦٧ in the TRMC test misfitted the Rasch model with outfit mean square value of ١.٣٥.

Figure ١and Table ١show that the two-response items were harder than their one-response counterparts by about half a logit. An independent samples t-test showed that the mean of item difficulty parameters in TRMC form (M=.٣٠, SD=.٩٠) was significantly higher than the mean of item parameters in SRMC form (M= -.١٩, SD=.٧٥), *t* (٥٨) = -٢.٣٤, *P*= .٠٢, effect size=.٠٨ (etta squared). Separation reliability for items in SRMC test was .٨٨ and in TRMC testwas .٨٩.

Results indicated that examinees who took the TRMC form (M=-.٦٩, SD=.٧١) performed better than those who took the SRMC form (M=-.٨٠, SD=.٨١) after ability parameters were brought onto the same scale. However, the difference was not statistical, *t* (١٥٢) =.٨٦,*p*=.٣٨. Person separation reliability for both forms was .٦٣.Item separation reliability for SRMC and TRMC form were .٨٨ and .٨٩, respectively. The number of misfitting persons with infit mean square values above ١.٣٠ in both forms was one. Misfitting persons are those with random responses who do not conform to model expectation and are thereforeunscalable. Table ١ summarizes the statistics for the two tests.

Table ١. Test statistics across forms (dichotomous analysis)

	Person Rel.	Item Rel.	Person. Mean	Item Mean	# Misfit Persons	# Misfit Items
SRMC	.٦٣	.٨٨	-.٨٠	-.١٩	١	٠
TRMC	.٦٣	.٨٩	-.٦٩	.٣٠	١	١

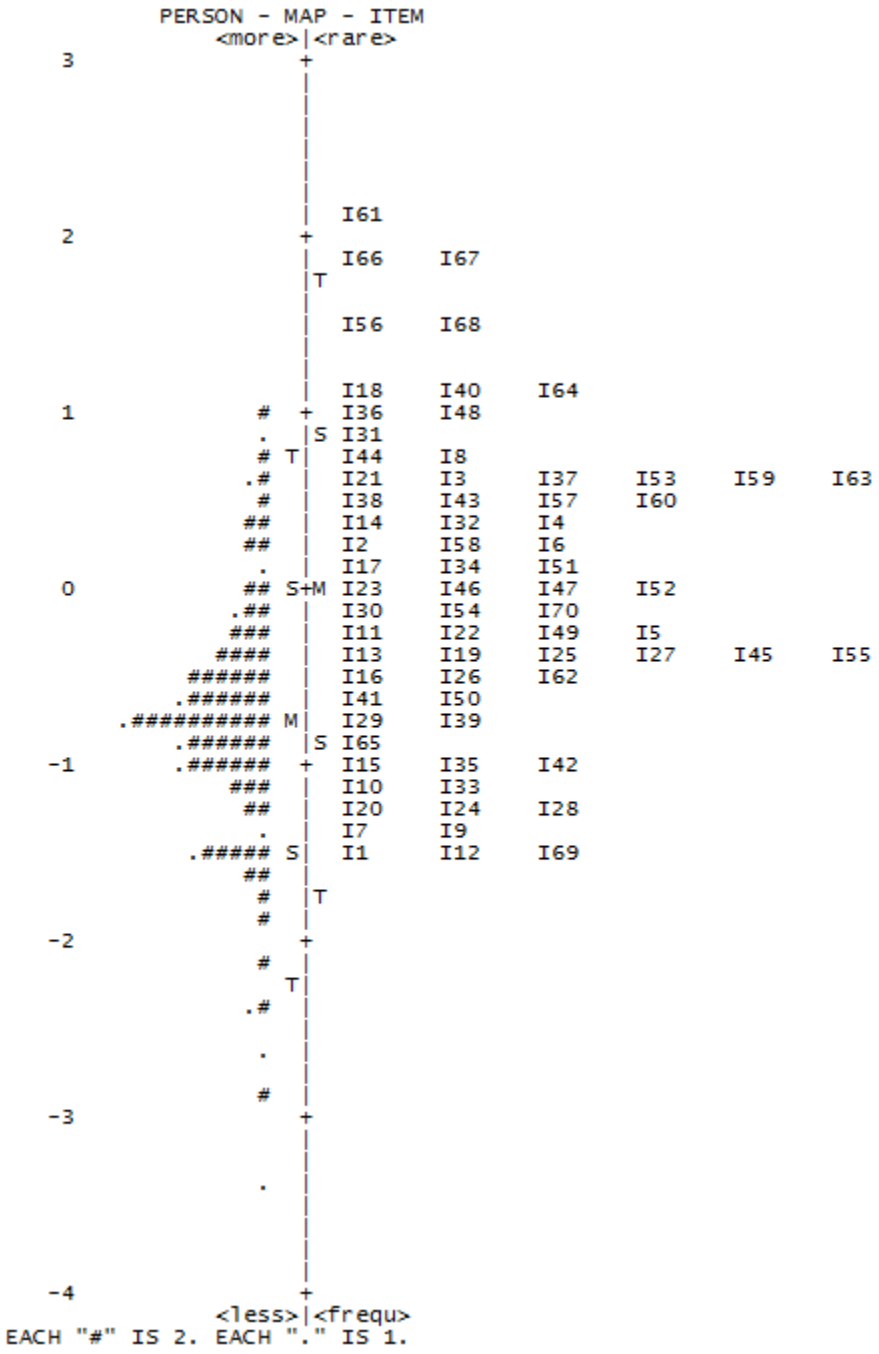
Rel.=Reliability

Table ٢. Item measures and fit statistics in the two forms (dichotomous analysis)

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.
11	29	74	-.28	.25	1.01	.2	1.01	.1	.31	.33
12	48	74	-1.46	.26	.99	.0	1.02	.2	.35	.35
13	31	77	-.34	.25	1.10	1.3	1.20	1.5	.18	.33
14	19	74	.41	.28	1.00	.1	.96	-.1	.29	.29
15	41	77	-.94	.24	.96	-.6	.91	-.9	.41	.35
16	35	78	-.55	.24	1.05	.7	1.04	.4	.28	.33
17	24	78	.13	.26	.87	-1.2	.79	-1.3	.46	.30
18	12	76	1.06	.33	.97	-.1	.94	-.1	.28	.24
19	31	78	-.32	.25	.93	-.9	.87	-1.0	.42	.32
20	46	78	-1.20	.25	.84	-1.9	.79	-2.0	.54	.35
21	17	75	.60	.29	.94	-.4	1.16	.7	.29	.27
22	29	78	-.19	.25	.90	-1.2	.86	-1.0	.44	.32
23	26	77	-.03	.25	1.05	.5	.96	-.2	.28	.31
24	47	78	-1.26	.25	.97	-.4	.90	-.8	.41	.35
25	31	77	-.34	.25	.99	.0	.95	-.4	.35	.33
26	34	75	-.56	.25	.88	-1.6	.84	-1.4	.48	.34
27	32	76	-.40	.25	1.08	1.1	1.09	.7	.23	.33
28	46	78	-1.20	.25	.99	-.1	.99	.0	.36	.35
29	38	77	-.76	.24	.87	-1.8	.83	-1.7	.50	.34
30	28	76	-.17	.25	1.02	.3	1.00	.1	.30	.32
31	13	72	.90	.32	.98	-.1	.85	-.4	.30	.25
32	20	76	.37	.27	.94	-.4	.86	-.6	.37	.29
33	45	78	-1.14	.24	.83	-2.2	.77	-2.2	.56	.35
34	24	78	.13	.26	1.10	1.0	1.21	1.2	.16	.30
35	41	75	-.98	.25	.86	-1.9	.80	-2.0	.52	.35
36	12	75	1.06	.33	.99	.0	.91	-.2	.27	.24
37	17	75	.60	.29	1.12	.8	1.15	.7	.13	.27
38	18	74	.51	.28	.92	-.6	1.03	.2	.34	.27
39	37	77	-.69	.24	1.24	3.1	1.24	2.2	.07	.34
40	11	75	1.18	.34	1.00	.1	1.21	.7	.21	.23
41	34	70	-.64	.25	1.02	.4	1.02	.2	.29	.32
42	41	73	-.99	.25	.98	-.2	1.00	.1	.33	.31
43	17	71	.55	.29	.92	-.5	.88	-.5	.38	.28
44	15	70	.76	.30	1.05	.3	1.13	.6	.19	.28
45	31	75	-.31	.25	1.05	.6	1.04	.5	.25	.31
46	26	73	-.04	.26	.97	-.3	.94	-.4	.35	.30
47	26	74	-.03	.26	.88	-1.3	.85	-1.2	.47	.31
48	13	74	1.00	.32	1.13	.7	1.14	.6	.09	.26
49	30	75	-.25	.25	1.01	.2	1.01	.1	.29	.31
50	37	75	-.67	.24	.85	-2.3	.82	-2.1	.52	.31
51	23	71	.13	.27	.99	.0	.95	-.3	.33	.30
52	25	74	.05	.26	.96	-.3	1.01	.1	.33	.30
53	16	71	.67	.30	.98	.0	.96	-.1	.30	.28
54	28	74	-.15	.25	.91	-1.0	.89	-1.0	.43	.31
55	31	73	-.37	.25	.94	-.7	1.03	.4	.37	.32
56	9	73	1.44	.37	1.02	.2	1.26	.8	.15	.23
57	19	73	.48	.28	.97	-.1	.94	-.3	.32	.28
58	23	74	.20	.26	.91	-.8	.88	-.8	.42	.30
59	17	73	.64	.29	1.17	1.1	1.24	1.2	.02	.28
60	19	71	.44	.28	1.09	.7	1.09	.5	.16	.29
61	5	70	2.12	.47	.98	.1	.92	.0	.21	.18
62	30	67	-.44	.26	.92	-1.0	.94	-.7	.41	.31
63	17	73	.64	.29	.99	.0	1.01	.1	.28	.28
64	11	69	1.17	.34	.97	-.1	.93	-.2	.31	.25
65	39	72	-.84	.25	.94	-.9	.94	-.7	.38	.29
66	6	70	1.91	.44	.86	-.3	.57	-1.0	.45	.20
67	6	72	1.94	.43	1.06	.3	1.35	.9	.03	.19
68	9	71	1.45	.37	1.08	.4	1.12	.5	.10	.23
69	39	58	-1.48	.29	1.07	.7	1.04	.3	.17	.26
70	21	56	-.13	.29	.87	-1.4	.85	-1.3	.48	.28

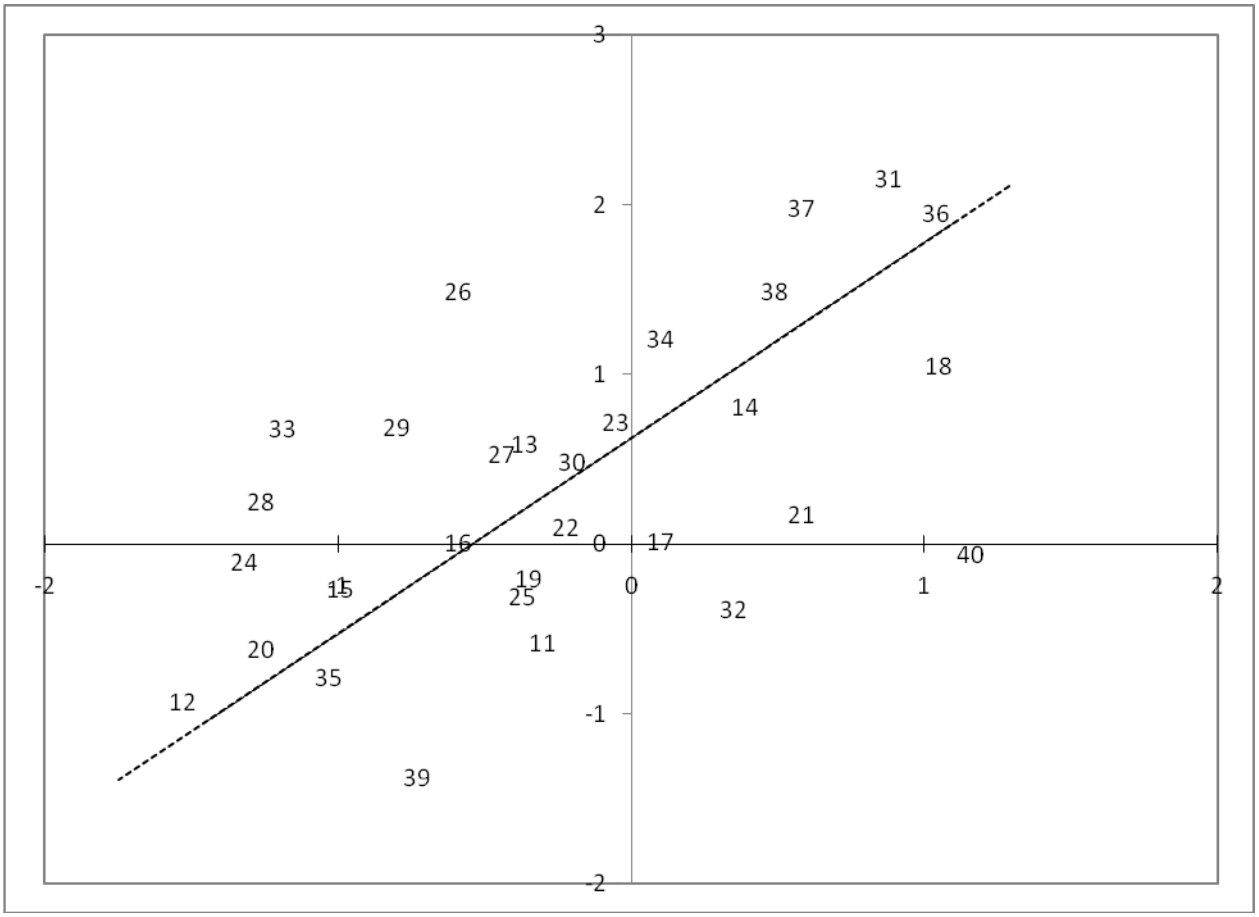
Note: Items ١١ to ٤٠ are in SRMC test and items ٤١ to ٧٠ are in TRMC test. Items ١ to ١٠ were anchor items and were dropped from further analyses after equating the two forms.

Figure ١. Distribution of items and persons on the Wright map (dichotomous analysis)



Since the items in the two forms were identical and were different only in one option, one can consider them as the same items and compare their difficulty parameters after they are brought onto the same scale. Figure ٢ which is the cross plot of item parameter estimates in the two test forms shows that the item parameters have changed considerably as items fall very far from the empirical line, indicating that the scoring procedure and the number of correct replies have notable impact on item estimates. The correlation between item measures estimated from the two forms after equating was .٥٦ which indicates that difficulty estimates have drastically changed depending on the number of correct options and scoring procedure.

Figure ٢.Cross plot of item parameters from SRMC against TRMC



٣.٢ Polytomous Analysis

In the second phase of the study the two-response items were scored polytomously, i.e., if examinees had marked one of the correct options they were given a score of ١ and if they had marked both correct options they were scored ٢. No correction for guessing formula was applied for scoring. TRMC data were analysed by means of Rasch partial credit model (Masters, ١٩٨٢). As in the dichotomous analysis, the two test forms were linked by means of ١٠ common items to estimate the difficulty and ability of all items and persons on a single scale. The difficulty estimates of the SRMC items were recalibrated in this analysis to be on a common scale with the polytomously scored two-response items. Therefore, the estimates of single-response items are different from their estimates in the dichotomous analysis. Cross plot of the difficulty parameters of the ١٠ common items from the two forms showed that all ١٠ items functioned well as they had identical estimates across forms.

Fit statistics in Table ٣ showed that polytomously scored two-response items had slightly better fit compared to their one-response counterparts. While none of the items in the two forms had infit and outfit mean square values above ١.٣٠, four items in the SRMC test and one item in the TRMC test had outfit mean square values above ١.٢. Examining person fit statistics showed that amongSRMC test-takers there was one person with an outfit mean squarevalue greater than ١.٣٠ and among TRMC examinees there were seven. The correlation between the item parameter estimates from the two forms was .٦١. Figure ٣ is a comparison of item difficulty parameters across the two forms when the two-response items were scored polytomously.

Figure ٣.Cross plot of item parameters from SRMC against TRMC in the polytomous analysis

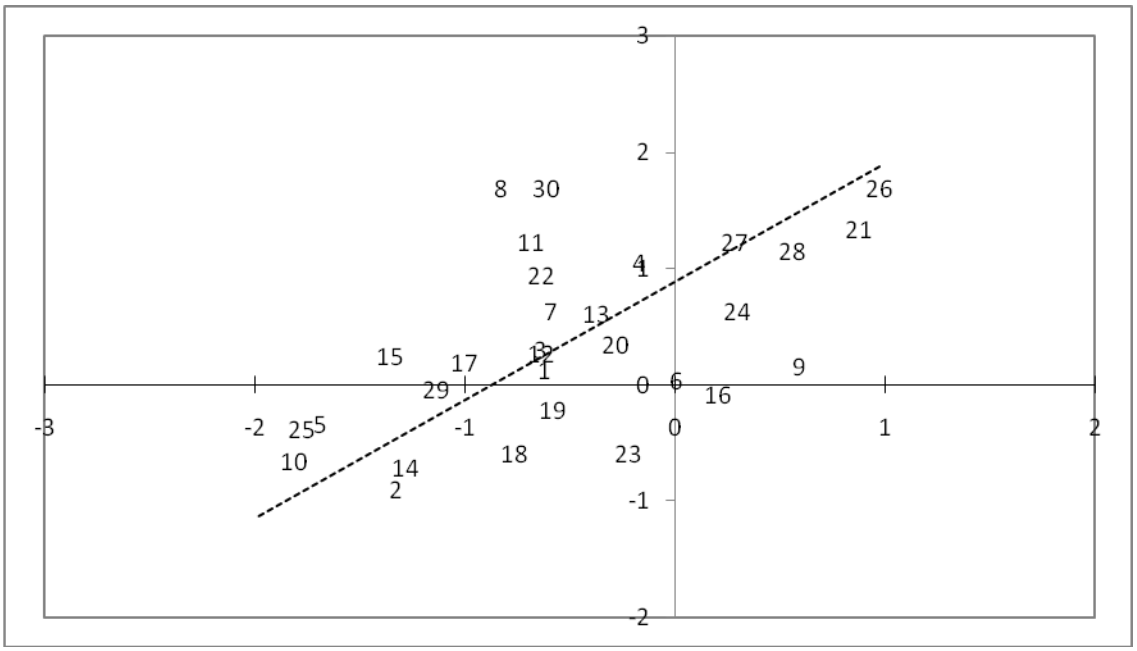


Table ٣.Item measures and fit statistics in the two forms (polytomous analysis)

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM	G
11	33	74	.13	.25	1.06	.9	1.09	1.0	.21	.30	56.8	63.1	I11	1
12	50	74	-.90	.26	1.03	.3	1.07	.6	.23	.29	71.6	69.6	I12	1
13	31	77	.31	.24	1.11	1.3	1.24	2.3	.11	.30	62.3	64.8	I13	1
14	19	74	1.05	.28	1.00	.1	.99	.0	.27	.28	74.3	74.8	I14	1
15	42	77	-.33	.24	.99	-.1	.96	-.4	.33	.30	55.8	62.6	I15	1
16	36	78	.04	.24	1.07	1.1	1.09	1.1	.19	.30	60.3	62.5	I16	1
17	26	78	.63	.25	.88	-1.3	.83	-1.4	.47	.29	74.4	69.1	I17	1
18	12	76	1.69	.32	.96	-.1	.97	.0	.28	.23	84.2	84.1	I18	1
19	34	78	.16	.24	.95	-.7	.93	-.8	.38	.30	65.4	63.3	I19	1
20	48	78	-.65	.24	.87	-1.7	.82	-1.8	.50	.30	69.2	65.6	I20	1
21	17	75	1.23	.29	.93	-.4	1.20	1.0	.28	.26	78.7	77.3	I21	1
22	32	78	.27	.24	.93	-.9	.92	-.8	.40	.30	71.8	64.4	I22	1
23	26	77	.61	.25	1.05	.5	.99	.0	.24	.29	63.6	68.8	I23	1
24	49	78	-.71	.25	.99	.0	.94	-.5	.33	.29	62.8	66.4	I24	1
25	32	77	.25	.24	.98	-.2	.96	-.4	.33	.30	67.5	64.2	I25	1
26	37	75	-.08	.24	.92	-1.2	.92	-1.0	.42	.30	73.3	62.0	I26	1
27	33	76	.19	.24	1.07	.9	1.10	1.2	.19	.30	60.5	63.2	I27	1
28	47	78	-.59	.24	1.00	.0	1.01	.2	.30	.30	65.4	64.9	I28	1
29	40	77	-.21	.24	.89	-1.7	.86	-1.8	.47	.30	68.8	62.2	I29	1
30	30	76	.35	.25	1.07	.9	1.09	.9	.19	.30	57.9	65.2	I30	1
31	15	72	1.34	.30	1.03	.2	1.05	.3	.20	.25	80.6	79.0	I31	1
32	21	76	.94	.27	.93	-.6	.87	-.8	.39	.28	73.7	73.2	I32	1
33	47	78	-.59	.24	.85	-2.0	.81	-2.0	.52	.30	73.1	64.9	I33	1
34	26	78	.63	.25	1.10	1.1	1.23	1.7	.10	.29	66.7	69.1	I34	1
35	42	75	-.38	.24	.86	-2.0	.82	-2.1	.51	.30	69.3	63.2	I35	1
36	12	75	1.69	.32	.98	.0	.93	-.2	.26	.23	84.0	83.9	I36	1
37	17	75	1.23	.29	1.11	.8	1.19	.9	.08	.26	78.7	77.3	I37	1
38	18	74	1.15	.28	.91	-.6	1.06	.4	.34	.26	81.1	75.8	I38	1
39	37	77	-.04	.24	1.24	3.4	1.28	3.2	-.05	.30	44.2	62.2	I39	1
40	12	75	1.69	.32	.98	.0	1.18	.7	.22	.23	84.0	83.9	I40	1
41	95	70	-.62	.18	1.04	.3	.99	.0	.34	.36	57.1	50.1	I41	0
42	111	73	-1.33	.21	1.00	.0	1.03	.3	.29	.31	63.0	58.8	I42	0
43	84	71	-.64	.24	.90	-.5	.89	-.5	.44	.29	73.2	71.1	I43	0
44	77	70	-.17	.22	1.02	.2	1.02	.2	.29	.32	70.0	67.6	I44	0
45	105	75	-1.69	.23	1.05	.5	1.04	.4	.21	.29	60.0	62.6	I45	0
46	79	73	.01	.16	1.07	.7	1.03	.3	.33	.39	43.8	42.6	I46	0
47	93	74	-.59	.20	.83	-1.2	.82	-1.3	.55	.33	63.5	59.0	I47	0
48	85	74	-.83	.28	1.09	.5	1.14	.7	.11	.25	79.7	79.8	I48	0
49	105	75	.59	.25	1.00	.1	.99	-.1	.28	.27	62.7	64.6	I49	0
50	111	75	-1.81	.23	.86	-1.4	.85	-1.4	.50	.29	68.0	59.6	I50	0
51	89	71	-.68	.22	1.04	.3	1.06	.4	.25	.32	60.6	63.0	I51	0
52	93	74	-.63	.21	.99	.0	.99	.0	.33	.32	59.5	61.1	I52	0
53	81	71	-.37	.23	1.03	.2	1.02	.2	.26	.31	69.0	69.6	I53	0
54	100	74	-1.28	.23	.88	-.9	.88	-.9	.49	.30	71.6	63.5	I54	0
55	102	73	-1.35	.23	.93	-.5	.96	-.3	.39	.30	68.5	61.0	I55	0
56	72	73	.21	.24	1.08	.5	1.07	.4	.17	.29	74.0	74.0	I56	0
57	90	73	-1.00	.25	.96	-.2	.97	-.1	.34	.27	72.6	72.1	I57	0
58	93	74	-.76	.22	.91	-.6	.90	-.6	.44	.29	71.6	65.4	I58	0
59	86	73	-.58	.24	1.20	1.1	1.21	1.1	-.06	.28	68.5	71.8	I59	0
60	82	71	-.28	.21	1.08	.6	1.08	.6	.19	.31	62.0	63.2	I60	0
61	58	70	.88	.23	1.04	.3	1.04	.3	.22	.28	68.6	68.5	I61	0
62	90	67	-.63	.20	.98	-.1	.98	-.1	.36	.32	61.2	52.6	I62	0
63	82	73	-.22	.21	.98	-.1	.98	-.1	.33	.30	67.1	66.4	I63	0
64	67	69	.30	.21	1.01	.1	1.01	.1	.30	.31	65.2	65.2	I64	0
65	110	72	-1.77	.23	.91	-.8	.90	-.8	.41	.26	70.8	59.6	I65	0
66	52	70	.98	.21	.96	-.2	.95	-.3	.37	.31	57.1	58.5	I66	0
67	71	72	.29	.29	1.04	.3	1.04	.3	.13	.23	81.9	81.9	I67	0
68	63	71	.56	.21	1.04	.3	1.04	.3	.24	.31	63.4	63.3	I68	0
69	93	58	-1.13	.22	1.02	.2	1.01	.1	.25	.28	60.3	64.6	I69	0
70	72	56	-.61	.23	.87	-.8	.86	-.8	.50	.29	66.1	56.8	I70	0
MEAN	59.2	84.6	.00	.23	1.00	.0	1.01	.0			68.1	67.1		
S.D.	30.6	27.1	.88	.04	.09	.9	.12	1.0			8.3	7.7		

Note: Items ١١ to ٤٠ are in SRMC testand items ٤١ to ٧٠ are in TRMC test. Items ١ to ١٠ were anchor items and were dropped from further analyses after equating the two forms.

Figure ٤ shows that the two-response items were easier than their one-response counterparts by about .٨٧ logits, when partial credit is given to them. An independent samples t-test showed that the mean of item difficulty parameters for SRMC items (M=.٣٧, SD=.٧٦) was significantly higher than the mean of TRMC item parameters (M= -.٥٠, SD=.٧٥), *t* (٥٨) = ٤.٤٧, *P*= .٠٠. Separation reliability for SRMC test was .٨٨ and forTRMC test was .٩١.

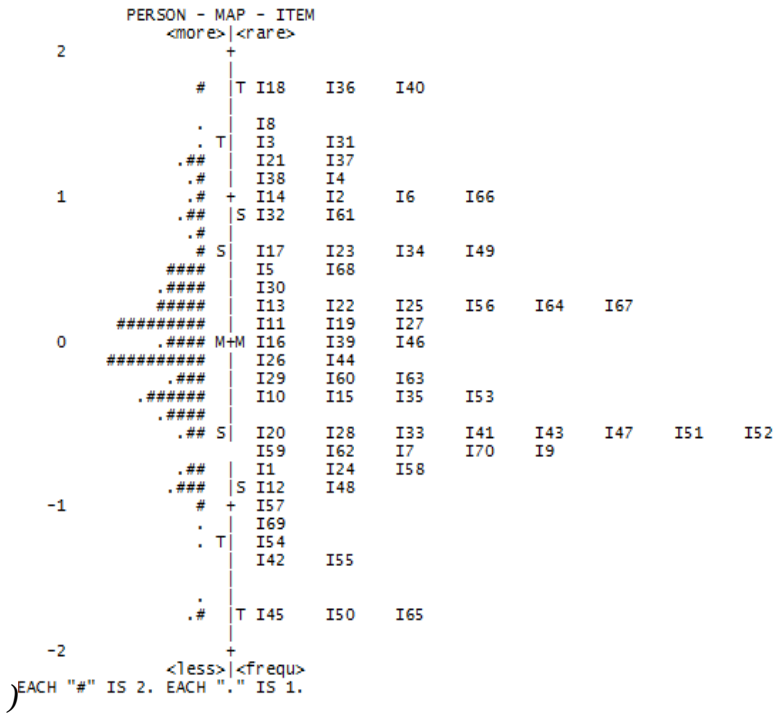
Results indicated that examinees who took the TRMC test (M=.١٥, SD=.٦١) performed better than those who took the SRMC test (M=-.١١, SD=.٦٧) after ability parameters were brought onto the same scale. The difference was statistical, t (١٥٢) =٢.٦٠ p=.٠١. Person separation reliabilities for single-response and two-response tests were .٦٣ and .٧١, respectively.

Table ٤. Test statistics across forms (polytomous analysis)

	Person Rel.	Item Rel.	Person Mean	Item Mean	# Misfit Persons	# Misfit Items
SRMC	.٦٣	.٨٨	-.١١	.٣٧	١	٠
TRMC	.٧١	.٩١	.١٥	-.٥٠	٧	٠

Rel.= Reliability

Figure ٤. Distribution of items and persons on the Wright map (polytomous analysis)



٣.٣ Dimentionality Analysis

Principal components analysis of standardized residuals (PCAR) was performed on both forms to compare the dimensionality of the two forms and the scoring procedures. In this approach to dimensionality assessment the residuals are subjected to principal components analysis (PCA). Since residuals are random noise in the data which are not explained by the Rasch model we do not expect to find a pattern among them. However, if a strong component is found in the residuals it is interpreted to be a secondary dimension in the data and evidence of departure from unidimensionality (Baghaei, &Cassady, ٢٠١٤; Linacre, ٢٠٠٩).

In SRMC test, the strength of the first component in the residuals was ٢.٧ in eigenvalue units. Linacre (٢٠٠٩) argues that the minimum eigenvalue for a component to be considered a distorting secondary dimension is ٢. SRMC test clearly departs from unidimensionality principle. In TRMC test the strength of the first component is ٢.٣ eignevalues in both dichotomous and polytomous scoring. Although the double-response form of the test is not strictly unidimensional, it is much closer to the measurement of a unidimensional construct.

٤. Discussion

The purpose of the present study was to compare the psychometric qualities of canonical single-response multiple-choice items with their double-response counterparts. To this end, ٤٠ four-option two-response multiple-choice(TRMC) grammar items for undergraduate students of English were constructed. A second form of the test was constructed by replacing one of the

correct options with a distractor to yield a canonical single-response four-option MC test (SRMC). The two test forms were linked by means of ١٠ one-response MC items which appeared at the beginning of both tests.

The two test forms were randomly distributed among ١٥٤ undergraduate students of English in their regular class time as their midterm exam. The data were analysed by means of Rasch measurement model to compare the psychometric qualities of the two test forms. Since the two test forms were linked with ١٠ common items and calibrated in a single analysis, person and item parameter estimates form the two separate forms were on a common scale.

The double-response items were scored in two different ways and both scoring methods were compared. First, they were scored dichotomously, i.e., items were scored as correct if examinees had marked both correct options and none of the distractors. In the second method they were scored polytomously, i.e., partial credit was given if examinees had marked only one of the correct options.

Results showed that double-response items were significantly harder than their single response counterparts when they were scored dichotomously. However, when they were scored polytomously, i.e., when creditwas given to partially correct responses, they turned out to be easier than single-response items. Comparing reliabilities showed that the single response test form was as reliable as the dichotomously scored double-response format. When double-response items were scored polytomously, their reliability surpassed that of single-response items. BaghaeiandSalavati (٢٠١٢) also demonstrated that polytomous scoring of such items results in easier but equally well fitting items with a slightly higher reliability.

Item fit statistics showed that almost all items fit in both forms. Person fit statistics however, showed that there are more misfitting persons among those who took the polytomously scored TRMC items than those who took theSRMC items. When double-response items were scored polytomously person fit deteriorated.

Examinees who took the TRMC items were more proficient that those who took single-response MC items as in both analyses they had higher means. However, in the dichotomous analysis their mean was not significantly different from the mean of single-response MC examinees (mean difference =٠.١٠ logits) but in the polytomous analysis their mean was significantly different (mean difference =٠.٢٦ logits). This is evidence that polytomous scoring which credits partial knowledge of the examinees results in finer discrimination among test-takers.

Dichotomous scoring of MRMC items has been criticized as it does not take into account partial knowledge of examinees. Test-takers who mark one correct reply are more knowledgeable than those who mark none of the correct options. Failure to account for examinees' partial knowledge threatens test validity and reliability. Applying a polytomous IRT model for scoring is one procedure to account for partial knowledge in MRMC items.

When double-response items were scored polytomously in this study, even without applying any formula to penalize for guessing the item and test statistics were satisfactory. Previous empirical research does not support the application of correction for guessing formulas either. For instance, Hsu, Moss, and Khampalikit (١٩٨٤) compared six scoring formulas in the context of College Entrance Examination in Taiwan. They found that giving partial credit improved reliability slightly but correction for guessing did not improve reliability and validity. Hsu, et al. (١٩٨٤) stated that the gain in penalizing for guessing, if any, is offset by the

complexity of scoring. Moreover, when correction for guessing formulas are used examinees are informed about this. Therefore, these formulas interact with examinees' personality (tendency to guess, etc.) and introduce additional sources of measurement error. Therefore, Ma (٢٠٠٤) does not recommend the application of correction for guessing formulas for scoring MRMC items.

Studies of Hsu, et al. (١٩٨٤) and Ma (٢٠٠٤) used classical test theory and compared statistics within this measurement model. This study, employing Rasch measurement model, showed that polytomous scoring without the application of correction for guessing formulas resulted in the increase of misfitting or unscalable persons. The reason for this might be that giving credit to partially correct items encourages guessing. If examinees know that two of the *n* number of options in an MC item are correct and they get a mark if they choose one of them then they are more encouraged to guess. In a four-option MC item with two correct responses the chances of selecting one correct reply is ٥٠ percent. If credit is given to partially correct items the application of some correction for guessing formula seems necessary otherwise it encourages guessing which result in unscalable persons. Therefore, the assertions of Ma (٢٠٠٤) and Hsu et al. (١٩٨٤) that correction for guessing is inconsequential sound questionable.

One limitation of this study is that the MRMC items of this study were double-response and examinees were informed of the fact that each item had two correct replies. This makes the test somewhat different from tests where there arean unknown number of correct replies in each item and examinees are instructed to mark as many correct replies as there are. These different configurations were not investigated and the results should be cautiously generalized to other MRMC items. Future research should studyand compare the efficiency of these response formats and their pertinent scoring formulas.

The results of the present study show that double-response multiple-choice items are promising alternatives to single response MC items. They have higher reliability than single-response items,if scored polytomoulsly, and fit the Rasch model better regardless of the type of scoring. Two-response MC items are closer to the unidimensionality principle of the Rasch model. Polytomous scoring of MRMC items requires the application of some correction for guessing formulas otherwise examinees' tendencies to guess lead to failure to scale lucky guessers. Future research should focus on the efficacy of scoring formulas and their effect on person fit.

References

Baghaei, P., &Cassady, J. (٢٠١٤).Validation of the Persian translation of the Cognitive Test Anxiety Scale.*Sage Open, ٤*, ١-١١.

Baghaei, P., &Salavati, O. (٢٠١٢). Double-track true-false items: A viable method to test reading comprehension? In Pishgadam, R. (Ed.).*Selected papers of the ١st conference on language learning and teaching: An interdisciplinary approach* (pp. ١٠٩-١٢٠). Mashhad: Ferdowsi University Press.

Baghaei, P., &Amrahi, N. (٢٠١١).The effects of the number of options on the psychometric characteristics of multiple choice items.*Psychological Test and Assessment Modeling, ٥٣*(٢), ١٩٢-٢١١.

Baghaei, P. (٢٠١١). Test score equating and fairness in language assessment.*Journal of English Language Studies, ١*(٣), ١١٣-١٢٨.

Birnbaum, A. (١٩٦٨). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. ٣٩٥-٤٧٩). Reading, MA: Addison-Wesley.

Bond, T. G., & Fox, C. M. (٢٠٠٧).*Applying the Rasch model: fundamental measurement in the human sciences*(٢nded.).Lawrence Erlbaum.

Dressel, P. L., &Schmid, J. (١٩٥٣).Some modifications of multiple-choice items.*Educational and Psychological Measurement, ١٣*, ٥٧٤-٥٩٥.

Hohensinn, C. H., & Kubinger, K. D. (٢٠٠٩). On varying item difficulty by changing the response format for a mathematical competence test. *Austrian Journal of Statistics, ٣٨*(٤), ٢٣١-٢٣٩.

Hsu, T. C., Moss, P. A., &Khampalikit, C. (١٩٨٤).The merits of multiple-answer item as evaluated by using six scoring formulas.*Journal of Experimental Education, ٥٢*, ١٥٢-١٥٨.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (٢٠١٠). On minimizing guessing effects on multiple- choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format.*International Journal of Selection and Assessment, ١٨*(١), ١١١-١١٥.

Kubinger, K. D., &Gottschall, C. H. (٢٠٠٧).Item difficulty of multiple choice tests dependent on different item response formats - An experiment in fundamental research on psychological assessment. *Psychology Science Quarterly, ٤٩*(٤), ٣٦١-٣٧٤.

Kubinger, K.D.,& Draxler, C. (٢٠٠٦). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier& C.H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models - Extensions and Applications* (pp. ٢٩٥-٣١٢). New York: Springer.

Linacre, J. M. (٢٠٠٩). A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs. Chicago, IL: winsteps.com.

Linacre, J. M. (٢٠٠٩). WINSTEPS® (Version ٣.٦٦.٠) [Computer Software].Chicago, IL: winsteps.com.

Ma, X. (٢٠٠٤).An investigation of alternative approaches to scoring multiple response items on a certification exam.Unpublished doctoral dissertation.University of Massachusetts.

Masters, G. N. (١٩٨٢). A Rasch model for partial credit scoring.*Psychometrika, ٤٧*, ١٤٩-١٧٤.

Orleans, J. S., & Sealy, G. A. (١٩٢٨).*Objective tests*. New York: World Book Company.

Page, G., Bordage, G., & Allan, T. (١٩٩٥).Developing key feature problems and examinations to assess clinical decision-making skills.*Academic Medicine, ٧٠*, ١٩٤-٢٠١.

Pashall, C.G., Stewart, R., & Ritter, J. (١٩٩٦, April).*Innovations: Sound, graphics and alternative response modes*. Paper presented at the annual meeting of the National Council on Measurement in Education. New York.

Pomplun, M., & Omar, M. H. (١٩٩٧). Multiple-mark items : An alternative objective item format. *Educational and Psychological Measurement, ٥٧*, ٩٤٩-٩٦٢.

Rasch, G. (١٩٦٠/١٩٨٠).*Probabilistic models for some intelligence and attainment tests* (Expanded Ed.). Chicago: University of Chicago Press.