# Utility of Complex Alternatives in Multiple-Choice Items: The Case of *All of the Above*

*Reza Nejati*[1] *& Mohammad Moradi*[2]

[1]Corresponding author, Shahid Rajaee Teacher Training University, reza.nejati@srttu.edu
[2]Shahid Rajaee Teacher Training University, Mohammadmoradi79@yahoo.com

*Received: 25/09/2014* *Accepted: 11/04/2015*

## Abstract

This study investigated the utility of *all of the above* (AOTA) as a test option in multiple-choice items. It aimed at estimating item fit, item difficulty, item discrimination, and guess factor of such a choice. Five reading passages of the *Key English Test* (*KET*, 2010) were adapted. The test was reconstructed in 2 parallel forms: Test 1 did not include the abovementioned alternative, whereas Test 2, administered 2 weeks later, included such an alternative. The 2 tests, 32 items each, were administered to 142 high school third-graders. Results, analyzed through 3-parameter logistic model, indicated that the multiple-choice questions, including the alternative *all of the above*, were easier. Results also revealed that the option *all of the above* increased the guess factor. Because guess factor is a source of measurement error, it may threaten test validity and reliability.

***Keywords:*** Multiple-Choice Questions, *All of the Above*; Item Fit; Item Difficulty; Item Discrimination; Guess Factor; Item Response Theory

## 1. Introduction

There are many techniques for testing language skills, namely, short answer, true/false, matching, multiple-choice, essay, and so on. Readers readily agree that multiple-choice items are widely used mainly because they are easy to administer and inexpensive to score and presumably objective.

The major disadvantage of multiple-choice items is that their construction is somewhat demanding. Perhaps the most difficult stage of constructing multiple-choice questions is finding plausible choices to function as distractors. To find plausible alternatives, language test developers need to use the findings of linguistics and other related disciplines such as educational psychology and assessment. For example, they should use typical errors of students to create plausible distractors. To do so, test developers need to have access to typical errors. It follows that we should collect such errors in a systematic way and save them for later use. That is to say, we need to have a collection of errors for purpose of item writing. To our best knowledge, such a collection is not available to L2 teachers in Iran; hence, they are left to their personal intuitions and experiences of item writing. Probably, they sometimes fail to use plausible choices.

It is sometimes observed that test constructors include choices such as *all of the above, none of the above* (known as complex alternatives) or *both* . . . *and* . . . (known as combination complex alternatives) in some items. These kinds of choices are not recommended by professionals in the field (Farhady, Jafarpur, & Birjandy, 2011; Osterlind, 2002). These alternatives are abbreviated as AOTA, NOTA, and BOTH, respectively, in the current paper. However, the present paper deals with AOTA.

Inclusion of these kinds of alternatives in test items or avoiding them is a matter of debate among teachers and test developers. The position of language testing scholars on the inclusion of such items ranges from strong disagreement to strong agreement in some cases. This opinion divide may be attributed to the scarcity of consistent empirical evidence in the field of item development. Hence, we felt motivated to investigate the issue.

## 2. Literature Review

Among the common alternatives included in test items are AOTA, NOTA, and BOTH. The former two are usually referred to as complex alternatives. and the latter is called combination complex alternative (Mueller, 1975; Tripp & Tollefson, 1985). The latter is also referred to as grouped options (Rossi, McCrady, & Paolino 1978). A test item that does not include any of these alternatives is referred to as *substantive response alternative test* (Mueller, 1975). Osterlind (2002) recommended the use of AOTA. Among the advantages he stated for the inclusion of AOTA is that:

> They can provide an appropriate discrimination between examinees who know an answer to an item and those who do not. The fact of their open-endedness tends to limit the possibility for guessing a single correct answer from among the response alternatives. (p. 151)

An argument for the use of AOTA is that the item format tends to be more difficult than those items in which it does not occur (Dudycha & Carpenter, 1973) and can, therefore, better discriminate between low and high achievers. Farhady, Jafarpur, and Birjandy (2011) and Osterlind (2002) do not recommend using complex alternatives, that is, AOTA or NOTA. They believe that such alternatives are usually used when the test developers do not find appropriate choices. Burton, Sudweeks, Merrill, and Wood (1991) argue against the use of AOTA as the correct answer and distractor. They hold that as AOTA, the correct answer can be identified by noting that two of the other alternatives are correct, and as a distractor, it can be eliminated by noting that one of the other alternatives is incorrect. It follows that in both cases, the possibility of chance increases, and this may affect the reliability and validity of the test.

On the other hand, Bruno and Dirkzwager (1995) and Owen and Freeman (1987) have shown that the optimal number of alternatives per item in multiple-choice items is three. Even Crehan, Haladyna, and Brewer (1993) have found in an experimental study with repeated measures design that items with three options were more difficult than those with four options. Therefore, it does not seem necessary to add an AOTA, NOTA, or BOTH as the fourth alternative in cases where the test taker really cannot find a plausible fourth alternative.

Candidates assume that each test item requires only one correct response. In the case of AOTA, when they see that two choices are correct, they even need not look at choice “C” and immediately find the correct choice and, as Osterlind (2002) states, “the perceptive examinee will automatically select the . . . [AOTA] option on the basis of only partial, rather than complete, knowledge of the item” (p. 154). On the other hand, some argue that the AOTA tends to be easier for test-wise students (Haladyna, Downing, & Rodriguez, 2002; Harasym, Leong, Violato, Brant, & Lorscheider, 1998) as test takers who can identify, at least, one option that is incorrect, logically eliminate the AOTA option, will find this item format easier than others who cannot.

Some believe that in cases where the test takers are sure of the correctness of a choice, but unsure of other choices, they may avoid giving any answer or select the choice which is correct by itself. In such cases, their knowledge of the choice which they are certain about is not assessed and not given any score. In other words, the students’ ability may be underestimated. Pashasharifi and Kiamanesh (1984) argue that “these phrases [AOTA & NOTA] do not result in meaningful and grammatical sentences when put in the blank” (p. 83). They add that “if a student finds that one of the alternatives is not correct, he or she immediately ignores AOTA [and again the item functions as a three choice item]” (p. 83). They also believe that using AOTA or BOTH contradicts the basic principle underlying the construction of multiple-choice items that there should be only one correct answer for each question.

Osterlind (2002) cautions test developers when he says, “the response alternative . . . [AOTA] should not be used with the best-answer type of multiple-choice item. With only cursory consideration, one realizes that “best” is inherently contradictory to [AOTA]” (p. 155). Musial, Nieminen, Thomas, and Bruk (2009) object to complex and combination complex alternatives saying that some students may find the first alternative as the correct answer and supposing that there is only one correct answer for each item they may never even look at the second alternative let alone the last alternative which is AOTA. Mousavi (2009), however, agrees with the use of AOTA and NOTA in multiple-choice items. Not only for what he calls their flexibility and ease of construction, but also for items involving logic skills or rote memory, such as spelling English mechanics and particular facts like historical dates and events.

The field of L2 testing seems to be divided over the use of AOTA. Some studies have provided mixed results and recommendations. The different designs, study limitations, and contrasts do not allow for definitive or careful statements about the use of AOTA. What has emerged from these studies is that student knowledge or ability is a factor in the use of AOTA; yet, almost no study systematically investigated this possibility. In addition, there may be differences in item performance when AOTA is the correct answer versus when it is used as a distractor. To date, to our best knowledge, almost no study has systematically investigated this possibility, either.

As it was previously stated, multiple-choice items are among the most widely used formats, especially in high-stake cases in which standardized tests of language proficiency are administered. Based on the results of such tests, a large number of educational systems throughout the world make great decisions every year for proficiency, prognostic, and evaluation of attainment purposes. However, one of the main shortcomings of multiple-choice items is the difficulty of constructing such items. In many cases, the test developer faces difficulty, especially in finding the suitable distractors. In such cases, he or she may include choices whose appropriateness for the situation may come strictly under question. He or she may also include choices that are not plausible and easily neglected by test takers.

Among the many choices included in such tests are AOTA. Educators and measurement specialists seem to be divided as to the usefulness of such choices. Therefore, it is worthwhile to provide empirical evidence for or against this argument, that is, the appropriateness of including such distractors in multiple-choices. In order to achieve the purpose of the study, the following research questions were formulated.

1. What is the item-fit of items entailing AOTA as compared to items without AOTA?
2. What is the difficulty level of items entailing AOTA as compared to items without AOTA?
3. What is the role of guess in items entailing AOTA as compared to items without AOTA?

4. How discriminating are items entailing AOTA as compared to items without AOTA?

## 3. Method

The design of the study was of a within-group nature in which the same participants took two forms of the test.

### *3.1 Participants*

The participants were 142 high school third-graders in seven classes in four schools. They majored in humanities (3 classes), sciences (3 classes) and mathematics (1 class). Seventy-nine of the participants were boys and 63 were girls.

### *3.2 Instrumentation*

In order to collect the data, five reading comprehension passages from the *Cambridge Key English Test* (*KET*, 2010) with 32 items were used. The items did not include the option *all of the above* (AOTA). To serve the purpose of the study, however, we developed a parallel form of the test to include the option AOTA. In the second test, eight items included AOTA, and the remaining items were free from AOTA.

The passages and the stems of the items were the same in the two forms. Also, except for the inclusion of the choices in question, the remaining alternatives were held fixed as much as possible.

### *3.3 Analysis and Procedure*

The first form of the test was administered to 142 students. After a two-week interval, the second form of the test was administered to the same students in order to investigate the possible effects of the inclusion of the AOTA.

Because the purpose of the study was to assess the qualities of individual items, not the students (to avoid sampling bias), item response theory (IRT) was used for data analysis. Because IRT data analysis is not dependent on the sample of respondents, it provides generalizable results. IRT provides item parameters such as difficulty, discrimination, and chance factor. The data obtained from the two tests were analyzed via Excel for Item Response Theory (EIRT)—a software developed by Germain, Valois, and Abdous (2007). It generates Baye's model estimator and can be mounted on Microsoft Excel.

## 4. Results

The present study was designed to compare the item-fit, difficulty, discrimination power, and guess factor of test items that include AOTA with items without AOTA. The items were calibrated with a three-parameter logistic (3PL) model. Here, research questions are taken one by one:

The first research question was: What is the item-fit of items entailing AOTA as compared to items without AOTA? In order to answer the question, the chi-square ($x^2$), to use IRT terminology, of the two forms of the test was generated. The results are displayed in Table1:

Table 1. *Item-Fit Parameter*

| Items | Test 1 | | Test 2 | |
|---|---|---|---|---|
| | ($X^2$) | *p* | ($X^2$) | *p* |
| 3 | 1.185 | 1.000 | .390 | 1.000 |
| 9 | .708 | 1.000 | 2.328 | .985 |
| 15 | .843 | 1.000 | 2.134 | .989 |
| 17 | .407 | 1000 | . 493 | 1.000 |
| 19 | 5.325 | .868 | 39.182 | .000 |
| 23 | 1.523 | .999 | 14.010 | .122 |
| 24 | 5.083 | .888 | .273 | 1.000 |
| 30 | 5.535 | .853 | 1.163 | .999 |

As Table 1 shows, item 19 in Test 2 (+AOTA) does not fit well with other items. The item-fit of items 9, 15, and 23 in Test 2 has decreased as compared to Test 1.The remaining items fit with the test. Because half of the items suffer the item misfit, it may be safe to hold that the option *all of the above* should not be recommended.

The second research question was: What is the difficulty level of items entailing AOTA as compared to items without AOTA? In order to answer the question, the Threshold (b), to use IRT terminology, of the two forms of the test was generated. The results are displayed in Table 2.

Before reporting the results, it should be noted that in IRT outputs, difficulty, threshold index, is reported within -3 and 3 standard deviations in *Z* score scale ( -∞, +∞, *infinity*, Baker, 2001). The smaller the index, the easier the item. The item difficulty identifies the ability level at which about 50% of the examinees are expected to answer the item correctly (DeMars, 2010). Item 24 seems to have turned into the easiest item. With the exception of items 9 and 17, items with AOTA tend to be easier than items without AOTA. It follows that items with AOTA would favor low ability level test takers. Assuming low ability test takers resort to guess, then items with AOTA are inflated with chance factor. Hence, constructing irrelevant information will be produced which, in turn, may disturb reliability and validity of the test:

Table 2. *Threshold Parameter*

| Items | Test 1 | | Test 2 | |
|---|---|---|---|---|
| | (b) | s.e. | (b) | s.e. |
| 3 | .405 | .243 | -.091 | .106 |
| 9 | -.018 | .138 | .025 | .136 |
| 15 | .769 | .164 | -.158 | .199 |
| 17 | .129 | .230 | .446 | .132 |
| 19 | .939 | .243 | .000 | .000 |
| 23 | 4.062 | 5.353 | .000 | .000 |
| 24 | 4.863 | 9.492 | -.525 | .092 |
| 30 | .171 | .111 | .032 | .137 |

*Note. s.e. stands for standard error of estimation.*

The third research question was: What is the role of guess in items entailing AOTA as compared to items without AOTA? In order to answer the question, the asymptote (c), to use IRT terminology, of the two forms of the test was supplied. The results are displayed in Table 3:

Table 3. *Asymptote Parameter*

| Items | Test 1 | | Test 2 | |
|---|---|---|---|---|
| | (c) | s.e. | (c) | s.e. |
| 3 | .130 | .011 | .172 | .000 |
| 9 | .213 | .007 | .234 | .001 |
| 15 | .179 | .000 | .132 | .017 |
| 17 | .168 | .019 | .136 | .011 |
| 19 | .195 | .000 | .200 | .000 |
| 23 | .139 | .004 | .200 | .000 |
| 24 | .223 | .015 | .128 | .002 |
| 30 | .245 | .000 | .261 | .001 |

*Note. s.e. stands for standard error of estimation.*

As it can be seen in Table 3, in five out of eight items in Test 2 (+AOTA) the asymptote index has increased. It sounds reasonable to hold that AOTA option may contribute to guess factor.

The fourth research question was: How discriminating are items entailing AOTA as compared to items without AOTA? In order to answer the question, the slope (a), to use IRT terminology, of the two forms of the test was produced. The results are displayed in Table 4 below.

Table 4. *Slope Parameter*

| Items | Test 1 | | Test 2 | |
|---|---|---|---|---|
| | (a) | s.e | (a) | s.e |
| 3 | .862 | .253 | 2.432 | .491 |
| 9 | 1.354 | .342 | 1.909 | .453 |
| 15 | 2.676 | .607 | .802 | .204 |
| 17 | .763 | .231 | 1.189 | .255 |
| 19 | 2.135 | .602 | 1.000 | .000 |
| 23 | .469 | .569 | 1.000 | .000 |
| 24 | .210 | .352 | 1.973 | .378 |
| 30 | 3.157 | .596 | 2.002 | .478 |

*Note. s.e. stands for standard error of estimation.*

As Table 4 shows, the discrimination power of items 15, 19, and 30 has decreased and, in the remaining five items, the discrimination power has increased. Because five out of eight items have contributed to chance factor (see Table 3 ), and six items of Test 2 are easier than those of Test 1 (see Table 2), the increased power of item discrimination may be attributed to chance.

## 5. Discussion

As noted earlier, six out of eight items of Test 2 (+AOTA) turned to be easier than their counterparts in Test 1. One of the main reasons for the easiness of the items including AOTA can be the test takers' tactic of guessing the correct answer as soon as they come across two correct options. They do not need to know the third choice and can choose the correct answer easily. The results, then, support Osterlind (2002) who believes that the test takers answer the AOTA-including alternatives only based on partial knowledge.

The results corroborate Harasym et al. (1998) and Haladyna, Downing, and Rodriguez (2002) that AOTA as a distractor can be eliminated by noting that one of the other alternatives is incorrect, hence making the test item seem easier for the test takers especially the test-wise ones. Test-wise and/or risk-taking students may perform better on these items. This would underscore item discrimination. Item discrimination, in IRT, is an item characteristic that describes the item's ability to measure differences of ability level. Because guess factor, prevailing AOTA items, is at odds with knowledge, such items tend to be biased. Bias is a threat to valid interpretation or use of test scores because bias favors one group of test takers over another. Bias also has dual meanings. Bias is a term that suggests unfairness or an undue influence. In statistics, bias is systematic error as opposed to random error.

When AOTA option is the distractor, test takers must know about the correctness of an alternative and the incorrectness of all remaining alternatives (at least 2). A main reason for this difference in performance may be attributed to the fact that when AOTA is the correct response, the test taker needs to know only about the correctness of two alternatives. Items in which AOTA was the correct response fell on the left side of the distribution curve. They tend to be very easy and helpful for low achievers. It seems that low ability is associated with guessing strategy. Thus, teachers and testers may not support the results of such items.

The results obtained, however, are not in line with Musial et al. (2009). They hold that items with AOTA as the correct response are difficult for the test takers. Some test takers may find the first alternative as the correct answer, and supposing that there is only one correct answer for each item, they may never pay attention to the second alternative and choose the first choice that is not necessarily the best response. The readers may readily agree that in cases which AOTA is the best answer, other options are also reasonably correct answers. Ironically, however, students who have chosen other (correct) options and have some information are penalized due to an all-or-none system of scoring. This flawed scoring system may compromise the reliability and validity of the test simply because the scoring scheme has ignored the fact that ability is relative not absolute. In other words, the scoring system has failed to count what counts. Test writers are advised to develop a partial credit scoring system for multiple-choice items. Such a system may be effective if the distractors can provide differential information. This type of information may be provided by items developed on the basis of a clear domain specification and the distractors are chosen from among choices generated out of the analysis of conceptual maps of reading. That is to say, test developers need to take account of what happens in the mind of the students during reading a passage. Testers should deal with reading processes such as text processing, syntactic analysis, vocabulary access, reasoning ability, and so on. Distractors based on testers' intuitions alone may not be helpful.

The results indicated that the test items in which one of the alternatives was AOTA-included were more discriminating than the ones in which they were not included. However, this increase in discrimination power cannot be helpful because it may be contaminated with guess and undue easiness.

## 6. Conclusion

We showed that the inclusion of AOTA among alternatives make the test easier for the test takers. Because test items here are unduly easy, the information they provide is construct-irrelevant. That is to say, such test items threaten the reliability and construct validity of the test. We also found that the alternative AOTA makes an item more discriminating. However, due to the construct-irrelevant information mentioned before, discrimination power of the present test may not be of any relevance. Moreover, when developing an achievement test, item discrimination may not count. Readers may agree that achievement tests are criterion-referenced. Each item needs to be linked to an instructional objective.

In sum, AOTA unduly increases item easiness and chance factor (perhaps due to test-wiseness of the test takers) and increases item discrimination that may not serve classroom teachers. Hence, with empirical evidence collected in this study, it may be safe to advise teachers and test developers to dismiss the use of AOTA option, at least, for classroom tests.

## References

Baker, F. B. (2001). *The basics of item response theory.* Maryland: RIC Clearinghouse on Assessment and Evaluation.

Bruno, E. J., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*(6), 959-966.

Burton, S. Sudweeks, Merrill, P., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty.* Utah: Brigham Young University Testing Services and Department of Instructional Science.

*Cambridge Key English Tests* (2010). Cambridge: Cambridge University Press.

Crehan, K. D., Haladyna, T. M., & Brewer, E. 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.

DeMars, C. (2010). *Item response theory*. Oxford: Oxford University Press.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item formats on item discrimination and difficulty. *Journal of Applied Psychology, 58,* 116-121.

Farhady, H., Jafarpur, A., & Birjandy, P. (2011). *Testing language skills: From theory to practice.* Tehran: Center for Studying and Compiling University books in Humanities (SAMT).

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,* 309-334.

Harasym, P. H., Leong E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of *all of the above* on the reliability and validity of multiple-choice test items. *Evaluation Health Professional, 21*(1), 120-133.

Mousavi, S. A. (2009). *An encyclopedic dictionary of language testing.* Tehran: Rahnama Publications.

Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple-choice achievement test items. *Educational and Psychological Measurement, 35,* 135-141.

Musial, D., Nieminen, G., Thomas, J., & Bruke, K. (2009). *Foundations of meaningful educational assessment.* New York: McGraw-Hill.

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

Owen, S. V., & Freeman, R. D. (1987). What is wrong with three option multiple items? *Educational and Psychological Measurement, 47,* 513-22.

Pashasharifi, H., & Kiyamanesh, A. (1984). *Shivehaye arzeshyabi az amookhtehaye danesh amoozan.*Tehran: Sherkat-e Chap va Nashre Iran.

Rossi, J. S., McCrady, B. S., & Paolino Jr., T. J. (1978). A and B but not C: Discriminating power of grouped alternatives. *Psychological Reports, 42*(2), 13-46.

Tripp, A., & Tollefson, N. (1985). Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options? *Journal of Nursing Education, 24*(3), 92-98.