«نشریه علمی-پژوهشی آموزش و ارزشیابی»
سال هشتم ـ شماره ۲۹ ـ بهار ۱۳۹۴
ص. ص. ۹۲-۷۵

Effectiveness of Options in Multiple-choice Items: the case of "none of the above"

Reza Nejati[1]
Mohammad Morady[2]

تاریخ دریافت مقاله: ۱۳۹۳/۰۲/۱۵
تاریخ پذیرش نهایی مقاله: ۱۳۹۴/۰۶/۰۹

Abstract

The present study examined the effectiveness of 'none of the above' (NOTA) as a test alternative in multiple- choice items.It intended to estimate item fit, item difficulty, itemdiscrimination, guess factor of such a choice and the reliability of the whole test. To this end, the researchersselected five passages of reading section of the Cambridge Key English Test known as KET (2010) and developed a parallel form of that test; test one did not include NOTA, whereas the second test, administered two weeks later, included NOTA. The two tests, 32 items each, were given to 142 high-school third graders. The results, analyzed through 3-parameter logistic model of item response theory (IRT), revealed that multiple-choice questions including the alternative NOTA were easier than their counterparts. In addition, NOTA option did not threaten item fit and item discrimination but increased the guess factor, which, in turn may threaten the reliability and validity of the test.

Keyterms: none of the above, item fit, item difficulty, item discrimination, guess factor, item response theory

1. Assistant professor, Shahid Rajaee Teacher Training University. reza.nejati@srttu.edu
2. M.A. student, Shahid Rajaee Teacher Training University

1. Introduction

One of the most commonly used tests items for assessing language learners' abilities is multiple-choice (MC) items.MC tests are used for a variety of purposes: placement, selection, certification, achievement, proficiency, and diagnosis of what has (not) been learned.Therefore, it is justifiably important to understand how to write these items correctly.

Writing MC items is not an easy task. Item writers need to follow certain instructions. These instructions are generally compiled in books on (language) testing (e.g., Brown, 2005; Burton, Sudweeks, Merrill, & Wood, 1991; Farhady, Jafarpur & Birjandi, 1994; Gronlund, 1988; Haladyna, 1994, 2004; Haladyna, Downing & Rodriguez, 2002; Osterlind, 2002 ; Rodriguez , 2005).

Haladyna (2004) provides 31 general guidelines for developingtest items. He proposes thirteen guidelines specifically for writing options (alternatives). For examples, he states:

> Develop as many effective options as you can, but two or three may be sufficient.
> Keep the length of options about the same.
> None of the above should be used sparingly.
> Avoid using all of the above.
> Avoid negative words such as not or except.
> Avoid options that give clues to the right answer.
> Make distractors plausible.
> Use typical errors of students when you write distractors (pp.99-100).

From the examples above, the researcher examines the function of the rule: None of the above (NOTA for short) should be used sparingly.The aforementioned guidelines are mainly based on personal intuitions or, at best, on limited empirical research. That is why questions with notable faults remain common on most multiple choice items (Downing, 2002).Faulty questions are prevalent even in high-stake tests such as Konkur, university admission test in Iran. As such defective items are a major concern, as they may negatively influence student learning and unreasonably inhibit the performance of some of the students over others (Tarrant and Ware, 2008).

Test writers are divided in their views on the use of NOTAin multiple choice items. Haladyna et al. (2002) report that 48% of researchers hold that NOTA should never be used, while 44% hold that NOTA can be useful, if used considerately. According to Haladyna and Downing review (1989a) twenty- six of 33 authors believe that NOTA should be avoided.

Some empirical research studies have been carried out on the issue of multiple- choice item writing in the past decade (e.g. Haladyna, 2004; Haladyna, Downing & Rodriguez, 2002; Martínez, Moreno, Martín and Trigo 2009; Rodriguez, 2005). These studies report results of item analyses done on the basis of the so-called Classical Test Theory (CTT). This theory is said to be sample and context dependent, hence, the findings seem to be questionable. Therefore, there still remains a clear need for methodologically defendable empirical research. That is, item analysis should be done through a measurement theory which is independent of the sample and the context. Such a theory is called Item Response Theory (IRT). The present study intends to provide empirical evidence for one the guidelines of item writing, namely, none-of-the-above (NOTA). The data is analyzed through IRT.

2. Review of literature

The field of testing seems to be divided over the use of NOTA in multiple-choice items.Seyf (2009) suggests that NOTA may be useful in testing spelling, pronunciation and mathematics.Pashasharifi and Kiamanesh (1984) and Rodriguez (1997) agree with NOTA in testing mathematical and computational problems in that such an option demands more cognitive processing on the part of the test takers; making the test item more difficult.However, Payne (2003) is concerned with using NOTA even in mathematics and suggests that it be the incorrect choice all the time if it is supposed to function appropriately. This would be a strange practice, though.

Rodriguez (1997) states that NOTA can be useful in that it may motivate examinees to read each option more carefully. He further holds that NOTA option may prove fruitful in tests of mathematics on the ground that it may encourage more accurate calculation and discourage repeated attempts to find the correct answer.

Proponents of NOTA believe that these optionsmay be helpful for increasing item difficulty when the options do not assess estimations (Dudycha & Carpenter, 1973; Mehrens & Lehmann, 1991). Frary (1991)advocated the use of NOTA believing it would favorably increase the difficulty of the test item .Frary's study indicated that NOTA items for which NOTA were the answer were on the average minimally more difficult than NOTA items for which NOTA was a distractor. Some research studies report that using NOTA increases the difficulty level of test items (e.g., Boynton, 1950; Crehan and Haladyna 1994; Crehan et al. ,1993; Dudycha and Carpenter ,1973; Hughes and Trimble ,1965; Kolstad and Kolstad

,1991; Mueller ,1975; Rich and Johanson ,1990; Oosterhof and Coats,1984;Schmeiser and Whitney ,1975; Tollefson and Chen ,1986; Tollefson, 1987; Wesman and Bennett ,1946 ;Williamson and Hopkins ,1967). Item difficulty, however, may increase the discriminatory power of the items. These characteristics of items may or may not be desirable in and by themselves; rather they should be considered in line with the purpose of the test and ability level of the examinees. It is worth noting thatRimland (1960) believes that NOTAdoes not influence the difficulty level of MC items.

Mousavi (2009) agrees with the use of NOTA in MC tests. Not only for what he calls its *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. Mousavi, further, suggests that great care must be taken regarding the use of NOTA alternative. However, he points out that the option NOTA tends to limit the possibility for guessing a single correct answer from among the response alternatives because of their open-endedness and this can make them advantageous.Closely related to this open-endedness nature of NOTA option is that it may inspire examinees to reflect on each option more carefully (Frary, 1991; Oosterhof & Coats, 1984; Wesman & Bennett, 1946).

Williamson and Hopkins (1967) hold that NOTA decreases the amount of chance variance represented in test scores when the test taker does not have the knowledge and consequently increases test reliability and validity of the test.Tollefson and Tripp (1983) indicated that items having NOTA as the correct answer had a significantly higher mean discrimination index compared to other two item formats in one of which the choice was a distractor and in the other, it was not included. However, no significant difference in the mean item difficulties for the three item formats was reported.Kolstad and Kolstad (1991) argued for the use of NOTA and suggested that it improved discrimination by reducing the likelihood of guessing correctly and enhanced validity.

Odegard and Koen (2007) demonstrated that the positive testing effect was negated when the NOTA alternative was the correct response on the initial multiple-choice test, but was still present when the NOTA alternative was an incorrect response. Mehrens and Lehman (1984) agree with the use of NOTA when it is used as a correct option about 1/c times the number of items in which it appears, where c is the number of options per item.

Rich and Johanson (1990) maintain that using NOTA increases discrimination power of the test. However, Tollefson and Tripp (1983),

Frary (1991), Hughes and Trimble (1965), Tollefson, (1987), Crehan and Haladyna (1991), Crehan, et al. (1993) claimed that this option did not have significant effect on item discrimination power in MC items. Results of Crehan and Haladyna's (1991) experimental study offered limited evidence to caution against the option NOTA.

In their meta-analytic review of item discrimination and difficulty in MC items using NOTA, Knowles and Welch (1992) reported no significant effect size for discrimination and item difficulty. Their findings indicated that using NOTA, as a test item option did not result in items of lesser quality than in items not using this option.

On the other hand, Dudycha and Carpenter (1973), Wesman and Bennett (1946), Schmeiser and Whitney, (1975), Tollefson and Chen (1986), Mueller (1975) and Haladyna and Downing (1989b) suggest that using NOTA decreases item discrimination and test reliability. They conclude that the use of NOTA generally has a negative effect on item characteristics, making items about 4.5% more difficultIn fact, studies into the difficulty level of test items as Osterlind (2002) states, have reported mixed results for those items in which the response alternative NOTA has been used. NOTA has been misused in some tests because students have learned that they are almost always the right answer, Mueller (2011) proposes that if used at all, NOTA be used both as the correct answer and as a distractor. She adds that the common recommendation for NOTA is to limit their use.

By examining the effects of NOTA on difficulty and discrimination indices in light of optimal difficulty, Rich and Johnson (1990) found that: (1) difficulty tended to approach the optimal level; (2) discrimination tended to increase; and (3) reliability was unaffected.

Rodriguez (1997) addressed the potential dangers in using this option: "One danger that exists in using NOTA is that the examinee that chooses NOTA as the correct response may be given credit for a wrong answer "(p.20). According to Brown (2005) the use of NOTA as response alternatives in multiple choice test items is tempting to many novice item writers.In the same way, Osterlind (2002) states: "[it] appears to fit easily into many multiple-choice test items and superficially make the item writer's task simpler" (p. 151).

Burton, Sudweeks, Merill and Wood (1991) and Farhady et al. (1994) do not recommend using NOTA. They maintain: "[this] alternative is usually used when the test developers do not find appropriate choices" (p. 96). Burton et al. (1991) argue against the use of NOTA on the ground that NOTA measures the ability to recognize incorrect answers rather than correct answers especially in cases that NOTA is the correct answer and it

does not appear plausible to some students.

Test items that use NOTA cannot discriminate between the test taker who really knows the answer and the one who does not. For example, a test taker may choose 'none' in the following item simply because he knows that none of London, Paris, and Madrid is the capital of Russia, but whether he knows that Moscow is, is not evident.

The capital of Russia is ……..….. .

A. London B. Paris C. Madrid D. None

In this case the question is only testing the students' ability to rule out wrong answers, and this does not guarantee that they know the correct one (Gronlund, 1988; Zimmaro, 2010).

As one of the main guidelines of constructing MC items, it is recommended that all of the alternatives or options be of the same length or matched (Farhady et al. 1994). If not, weak test takers may easily choose the longest or the shortest alternative only because they think that the correct answer is a long one or a short one. NOTA option may be quite longer than other alternatives in many cases and this may influence the test takers' decision about the correct alternative.

As the literature shows the scholars hold different views towards the inclusion of NOTA as an alternative in MC items .The present study intends to investigate the issue in an Iranian context and view how Iranian high-school third graders function in MC items in which NOTA is used.

3. Purpose of the Study

As mentioned earlier in this paper, multiple choice items are among the most widely used formats especially in high-stake tests in which standardized tests of language proficiency are administered. Based on the results of such tests, every year, a large number of educational systems throughout the world make great decisions for proficiency, prognostic and evaluation of attainment purposes. However, one of the main shortcomings of MC items is the difficulty of constructing items. In many cases, the test developer faces difficulty, especially, in finding plausible 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. Such items may be neglected by the test takers. Paradoxically, these items may push test takers into reprocessing the test items and cause misinterpretation of the input.MC items are probably the most frequently used types of items for measuring the test takers' abilities. They are also among the most difficult types of items to develop if not the most difficult ones.

As mentioned before, some test developers include choices such as NOTA, while constructing multiple-choice items. These kinds of choices are not endorsed by the professionals and authorities (Farhady et al 1994). The view of the scholars on the inclusion of such items ranges from strong disagreement to moderate and even strong agreement in some cases.

Inclusion of these kinds of alternatives in test items or avoiding them is an open debate among teachers and test developers. These views are mostly based on personal experience, wisdom, and limited empirical research.Among the many choices included in such tests are NOTA that are not supported by authorities in our field. However, lack of empirical research in this regard in the Iranian context motivated the researcher to investigate the issue among the learners in question.

For the purpose of the study two forms of the KET test, detailed later in the paper, were developed. In form 1 NOTA is not used, but in form 2 NOTA is used. Therefore, the researcher attempts to provide empirical evidence to demonstrate the function of such items. The functions are examined in terms of item-fit, item difficulty, item discrimination and guess factor. The reliability and validity of these tests are examined, too. In fact, this study tries to answer the following research questions:

1. What is the item fit of items entailing NOTA as compared to items without NOTA?
2. What is the difficulty level of items containing NOTA as compared to items without NOTA?
3. What is the role of guess in items containing NOTA as compared to items without NOTA?
4. How discriminating are items containing NOTA as compared to items without NOTA?
5. Does the option NOTA influence the reliability of the test?

4. Method

4.1 Participants

The participants of the study were 142 high school third graders. These students would go to four different high schools. They were in seven classes and studied in different majors (humanities: 3 classes, sciences: 3 classes and mathematics: 1 class). Seventy nine of the participants (55.6%) were boys and sixty three (44.4%) were girls.

4.2 Materials

In order to investigate the problem of the inclusion of NOTA in multiple-choice item tests, two 32-item forms of a test based on reading

comprehension passages taken from KET (Cambridge Key English Test, 2010) were prepared by the researchers and checked by two professors. In the first form, NOTA option was not included among the alternatives of the test items. In the second form of the test, eight items included NOTA and the remaining items were free from such an alternative.The passages and the stems of the items were the same in the two forms. In addition, except for the inclusion of the choice in question, the remaining alternatives were held fixed as much as possible.

4.3 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 NOTA. The test takers were assured that the test results would not influence their classroom evaluation.

5. Data Analysis

The data obtained from the two tests were analyzed through a three-parameter logistic (3PL) model of item response theory. This model addresses item-fit, item difficulty, item discrimination; guess factor, reliability and empirical validity of the tests.

6. Results and discussion

The purpose of the present study was to examine the item-fit, difficulty, discrimination power, guess factor , reliability and validity of test that include NOTA and the test without NOTA. The items were calibrated with a three-parameter logistic (3PL) model. It should be reminded that eight items, as displayed in the following tables, included NOTA. Hence, data analysis was limited to these items. Here, research questions are dealt with one by one.

Research Question 1: What is the item fit of items with NOTA as compared to items without NOTA?

To answer the question, the Chi-square (x^2) index of the two forms of the test was generated. The results are displayed in Table 1.As Table 1 shows, all chi-square values in test 1 are insignificant. It follows that allitems in this test fit with each other. In contrast, item10 in Test 2 (NOTA) does not fit well with other items ($x^2 = 27.162, P = 0.002$). The remaining items, however, fit well with the test. Hence, it is safe to say that the option NOTA

does little harm to the overall fit of the test developedfor the purpose of the present study. Faulty items do not provide a good

Table 1
Item-fit Parameter for NOTA items

	Test1		Test2	
Items	(X^2)	P	(X^2)	P
2	1.352	0.998	10.020	*0.439*
5	2.395	0.984	1.261	1.000
7	0.960	1.000	5.692	*0.840*
10	0.889	1.000	27.162	*0.002*
12	1.922	0.993	2.821	*0.985*
14	3.007	0.964	1.769	0.998
21	0.343	1.000	0.388	1.000
28	1.159	0.999	13.589	*0.193*

Sample of the domain (ability) under study. That is to say such items may threaten the validity of the test. To do away with the faulty item, test makers can alter its wording or replace the options with more plausible options or replace the item with an alternative one.

Research Question 2: What is the difficulty level of items containing NOTA as compared to itemswithout NOTA?

To answer the question, following IRT practice, the threshold (*b*) of the two forms of the test was generated. The results are presented in Table 2. As the readers know, in IRT outputs, the threshold index ,i.e., difficulty, is usually reported within -3 and +3 standard deviations in Z score scale (-∞, +∞, *infinity*, cf., 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). As the table suggests items 5, *10, 12, 14, 21*, and *28* seem to have turned into easier items. With the exception of items *2* and *7*, items with NOTA tend to be easier than items without NOTA. All in all, the mean difficulty of test 1 was 1.54 (SD= .76) while that of test 2 (NOTA) was .91 (SD= 1.22). In other words, test 2 (NOTA) turned out to be somewhat easier than test 1. It follows that items with NOTA may favor low ability level test takers. Assuming low ability test takers tend to resort to guess, it seems reasonable to hold that items with NOTA may be inflated with chance factor.

Table 2
Threshold Parameter for NOTA items

	Test1		Test2	
Items	(b)	s.e.	(b)	s.e.
2	2.431	1.166	2.433	1.383
5	1.717	0.867	-0.516	0.110
7	1.870	0.972	2.753	1.454
10	0.936	0.202	0.000	0.000
12	0.992	0.183	0.241	0.116
14	1.267	0.249	0.582	0.128
21	0.476	0.186	0.084	0.108
28	2.675	1.141	1.766	1.507

s. e. stands for standard error of estimation

Readers agree that chance may cause poor students get good marks and ironically cause good students get bad marks. Hence, construct irrelevant information, systematic error in test scores, will be produced which, in turn, may threaten reliability and validity of the test. That is to say, these tests may fail to measure what they are supposed to measure. In such cases, the test makers should identify faulty test items and revise them.

Research Question 3: What is the role of guess in items containing NOTA as compared to items without NOTA?

To answer the question, the asymptote (*c*), to use IRT terminology, of the two forms of the test was supplied. The results are presented in Table 3. As it can be seen in Table 3, in three out of eight items in Test2 (NOTA) the asymptote index has increased. However, the mean guess index of test 1 is estimated to be .19 (SD=.03) and that of test 2 is .18 (SD=.04). The mean difference is negligible. It sounds reasonable to hold that NOTA option has little contribution to guess factor. Guess factor is further examined in terms of information function of the tests later in this paper.

Table 3
Asymptote Parameter for NOTA items

	Test1		Test2	
Items	(c)	s.e.	(c)	s.e.
2	0.259	0.005	0.150	0.020
5	0.145	0.016	0.116	0.007
7	0.183	0.016	0.262	0.000
10	0.193	0.001	0.200	0.000
12	0.239	0.000	0.209	0.000
14	0.178	0.000	0.219	0.000
21	0.166	0.007	0.146	0.000
28	0.180	0.001	0.149	0.038

s. e. stands for standard error of estimation

Research Question 4: How discriminating are items containing NOTA as compared to items without NOTA?

To answer the question, the slope (*a*), to use IRT terminology, of the two forms of the test was produced. The results are provided in Table 4 below.As Table 4 shows, the discrimination power of items 2, 10 and 28 in test 2 (NOTA) has decreased and in the remaining five items, the discrimination power has increased. However, the mean slope of test1 is estimated to be 1.22 (SD= .75) and that of test 2 is estimated to be 1.65 (SD= 1.2). However, due to the larger SD in NOTA test, the discriminatory power of (NOTA) items does not seem to have increased. This piece of finding may

Table 4
Slope Parameter for NOTA items

	Test1		Test2	
Items	(a)	s.e.	(a)	s.e.
2	0.701	0.436	0.279	0.140
5	0.437	0.207	1.452	0.290
7	0.443	0.226	1.480	0.662
10	1.737	0.499	1.000	0.000
12	2.555	0.612	2.946	0.591
14	1.826	0.568	3.398	0.627
21	1.130	0.307	2.548	0.516
28	0.981	0.582	0.154	0.100

s. e. stands for standard error of estimation

Helpus suspect that guess tactic has played a role in test 2. The guess factor may threaten the reliability and the construct validity of the test.

However, in the case of criterion-referenced tests in which test item need to be linked to an instructional objective such as diagnostic, achievement and proficiency tests, item discrimination may not be significant.

Research Question 5: Does NOTA influence the reliability of the test?

To answer this question, the information function of the test, i.e., precision of the measurement should be examined (Baker, 2001).Usually, information function is presented in a graph. To do so, the amount of test information is plotted against ability level. According to IRT, the larger the information functions is the more precise the measurement is. The information function of the two tests is presented in the following figures. Inspecting Figure 1, one can see that the amount of information has a maximum at an

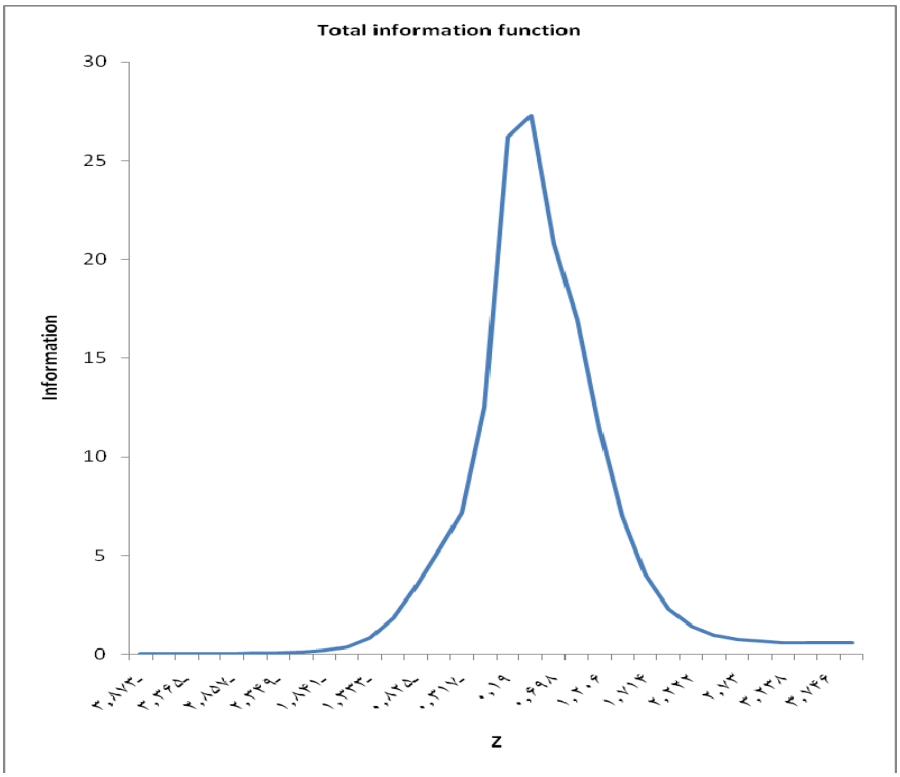

Figure1. Information functions of Test 1

Ability level of .44 and examining Figure 2, one can see that the amount of information has a maximum at an ability level of – .06. Since the information provided by test 2 (NOTA) is

Smaller than test 1, it seems reasonable to suggest that NOTA reduces the precision of estimation. That is to say, such items may fall short of assessing the 'ability' in question. In other words, these items may reduce the reliability of the test. Further examination of the following figures indicates that the higher end of the distribution is very close to the horizontal axis of the graph in Figure 2 (NOTA) as compared to Figure 1. This can be an indication that NOTA may increase the guess factor. However, in Figure 1, the lower end of the distribution touches the horizontal axis of the graph at ability level of -1.84 while in Figure 2 (NOTA), the lower end of the distribution touches the horizontal axis of the graph at ability level of -2.34.

Owing to the fact that in item response calibration the negative sign shows that the item is more likely to be easier for weaker students, test 1 has been more prone

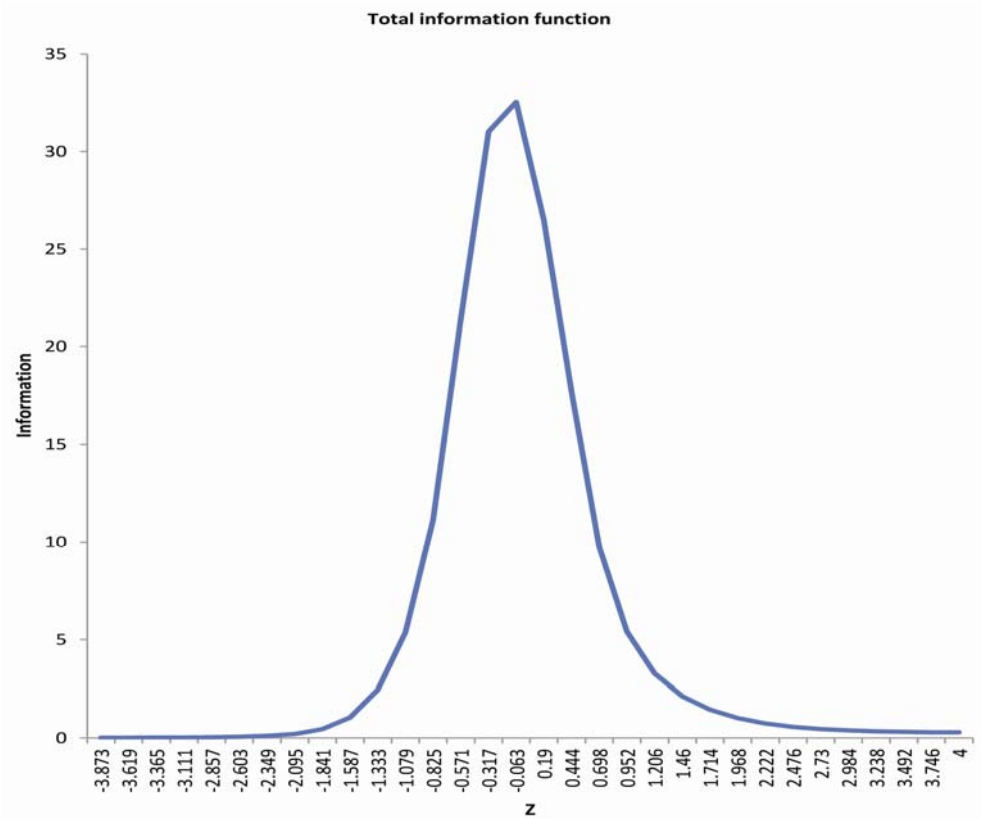

Figure2. Information functions of Test 2 (NOTA)

To guess factor at lower ability level while test 2 (NOTA) has been more prone to guess factor at a slightly higher level of ability. Hence, the option NOTA has brought about little chance for weaker students to use guessing tactic for answering questions.

Observing DeMars (2010:109), the following formula was used to estimate the empirical reliability of the two tests:

$$1 - s^{-2}_{e}$$

The reliability of test 1 turned out to be.73 and that of test 2 (NOTA) turned out to be .65. Hence, it stands to reason that the option NOTA negatively influences the reliability of the test.

7. Conclusion

This study provided some empirical evidence to claim that the inclusion of NOTA among alternatives makes the test easier for the test takers. Given that the test items, here, are unduly easy, the information they provide may be construct-irrelevant. Hence, the construct validity of such tests is in question. This piece of finding is at odds with Williamson and Hopkins (1967) who argue that NOTA increases test reliability and validity. As it was argued earlier, one reason for the easiness of the items including NOTA can be the test takers' tactic of guessing the correct answer as soon as they come across two incorrect options. In the present study, it was found that the students at the higher ability level, the higher end of the distribution (Figure 2), may have resorted to guess tactic. One way to control guessing may be to set a reasonably time limit for answering such test items. Hence, a speed test is recommended.Another alternative may be setting some punishment procedure such as cutting scores for random guesses.

As it was demonstrated in this study, NOTA has a negative influence on test reliability and validity. However, this piece of finding should be interpreted with great care because, as readers readily know, there is no clear-cut agreement among practitioners as to what coefficient of reliability and validity is acceptable or desirable.

To sum up, NOTA unduly increases item easiness and chance factor, perhaps due to test wise-ness of the test takers which is considered a source of construct-irrelevant information. It follows that NOTA option jeopardizes the reliability and validity of the test. Hence, with empirical evidence provided in this study, it may be safe to ask teachers and test developers to dismiss the use of NOTA option, at least in the case of assessing students' reading comprehension.Since this study was limited to a small sample of high school third graders, it would be safe to suggest that the study should be replicated with other groups of learners to provide results that are more reliable.

References:

Baker, F., B. (2001). ***The Basics of Item Response Theory.*** ERIC Clearinghouse.

Boynton, M. (1950). Inclusion of 'none of these' makes spelling items more difficult. ***Educational and Psychological Measurement***, 10, 431-432.

Brown, J. D. (2005). ***Testing in Language programs: a Comprehensive Guide to English Language Assessment.*** New York: McGraw-Hill.

Burton, S. Sudweeks, Merrill, P. & Wood, B. (1991). ***How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty***. Brigham Young University: Testing Services and the Department of Instructional Science.

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

Crehan, K. D., & Haladyna, T. M. (1991).The validity of two item-writing rules. ***The Journal of Experimental Education***, 59(2), 183-192.

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 University Press.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? ***Academic Medicine***, 77(10), 103-104.

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. & Birjandi, P. (1994). ***Testing language skills: from theory to practice,*** Tehran: the Center for Studying and Compiling University books in Humanities (SAMT).

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

Gronlund, N.E. (1988***). How to construct Achievement Tests***. Englewood Cliffs: Prentice Hall.

Haladyna, T. M. (1994). ***Developing and validating multiple-choice test items.*** Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

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

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

Haladyna, T. M., & Downing, S. M. (1989b).Validity of taxonomy of multiple-choice item writing rules. ***Applied measurement in education***, 2 (1), 51-78.

Haladyna, T. M., Downing, S. M., & Rodríguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. ***Applied Measurement in Education***, 15(3), 309-334.

Hughes, H. H., & Trimble, W. E. (1965).The use of complex alternatives in multiple-choice items. ***Educational and Psychological Measurement***, 25(1), 117-126.

Knowles, S. L., & Welch, C. A. (1992).A meta-analytic review of item discrimination and difficulty in multiple-choice items using none-of-the-above. ***Educational and psychological measurement***, 52, 571-577.

Kolstad, R. K., & Kolstad, R. A. (1991). The option "none of these" improves multiple-choice tests items. ***Journal of dental education***, 55(2), 161-163.

Martínez, R., J., Moreno, R., Martin, I., Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. ***Psicothema*** 21 (2), 326-30.

Mehrens, W. A., and Lehmann, I. J. (1984). ***Measurement and evaluation.*** Third edition. New York: CBS College Publishing.

Mehrens, W. A., and Lehmann, I. J. (1991). ***Measurement and evaluation in education and psychology***. Orlando, FL: Harcourt Brace Jovanovich.

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.

Mueller, J. (2011). ***Constructing good items***. Retrieved on January 2012 from http:// jfmueller.faculty.noctrl.edu/toolbox/index.htm.

Odegard, T.N., Koen, J.D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: implications for the testing effect. ***Memory***,15(8), 873-85.

Oosterhof, A. C., & Coats, P. K. (1984). Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple choice item formats. ***Applied psychological measurement***, 8, 287-294.

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

Pashasharifi, H. & Keyamanesh, A. (1984). ***Shivehaye Arzeshyabi az Amookhtehaye Danesh Amoozan*** [Methods of assessing students knowledge].Tehran: Sherkat e chap va nashre iran.

Payne, D. A. (2003). ***Applied educational assessment*** (2nd Ed.). United States: Wadsworth.

Rich, C. E., & Johanson, G. A. (1990). ***An item-level analysis of "none of the above."*** Paper presented at the annual meeting of the AERA, Boston, MA.

Rimland, B. (1960). The effects of varying time limits and of using Right answer not give in experimental forms of the U.S. Navy Arithmetic Test. ***Educational and psychological measurement***, 20(3), 533-539.

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

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

Schmeiser, C. B., & Whitney, D. R. (1975). ***The effect of incomplete stems and "none of the above" foils on test and item characteristics.*** Paper presented at the annual meeting of the NCME, Washington, DC.

Seyf, A. A. (2009). ***Andazehgiri sanjeshva arzeshyabi e amoozeshi*** [Educational measurement assessment and evaluation]. Tehran: Nashr e Dawran.

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

Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using none of the above and one correct response options. ***Educational and psychological measurement***, 47(2), 377-383.

Tollefson, N., & Chen, J., S. (1986). ***A comparison of item difficulty and item discrimination of multiple-choice items using none of the above options.*** Paper presented at the annual meeting of the Midwest AERA, Chicago, IL.

Tollefson, N., & Tripp, A. (1983). ***The effect of item format on item difficulty and item discrimination.*** Paper presented at the annual meeting of the AERA, Montreal, Quebec.

Wesman, A. G., & Bennett, G. K. (1946).The use of 'none of these' as an option in test construction. ***Journal of Educational Psychology***, 37, 541- 549.

Williamson, M. L., & Hopkins, K. D. (1967). The use of none-of-these versus homogeneous alternatives on multiple-choice tests: Experimental reliability and validity comparisons. ***Journal of educational measurement***, 4(2), 53-58.

Zimmaro, D. M. (2010). ***Writing good multiple-choice exams***. Retrieved from www.ctl.utexas.edu/Evaluation--Assessment/Writing Good Multiple Choice Exams-04-28-10-pdf.