Pazhuhesh-e Zabanha-ye Khareji, No. 27, Special Issue, English, 2006, pp 199-214

The Effect of Defective Options on Multiple-choice Items and Test Characteristics

Fattaneh Sajadi
University of Tehran, Faculty of Foreign Languages
e-mail: fattaneh-sajadi@yahoo.com

Abstract

Over a century or so, testing has been the most common way to determine the achievement of the learners in educational centers. Among many test formats, multiple-choice has been more common for which many rules and regulations have been suggested for preparing sound items. Many researchers, therefore, have focused on whether observing the rules would improve the quality of multiple-choice tests. So, the purpose of this study was to investigate the effect of faulty options on item characteristics. To accomplish this goal, a sample of 50 male and female participants was chosen from the students majoring in TEFL. Participants were given two multiple-choice tests in a week interval: (1) Michigan Test of English Grammar, and (2) Distorted Michigan Test of English Grammar. The results of this study revealed that in general, violation of item-writing principles did not significantly influence item characteristics. However, use of grammatical clues made the distorted items significantly easier.

Key Words: Item Characteristics, Item Facility, Item Discrimination, Choice Distribution, Reliability, Defective options, Multiple-choice.

1. Introduction

There are two main dimensions for the construction of an acceptable test item: psychometric and logical consideration. The psychometric aspect involves technical knowledge and skills regarding statistical operations in determining item discrimination, item facility, and test reliability. The logical aspect involves integrating theoretical, verbal and practical aspects into the process of test development.

In educational contexts, teachers make the largest population of test makers. Therefore, they should be well trained in implementing both psychometric and logical aspects of test construction. On the psychometric side, utilizing statistical techniques of items analysis would make decisions on item quality fairly objective. On the logical sides, however, teachers are recommended to attend to the suggestions made by testing experts. For example, it is claimed that including grammatical clues to the best choice would decrease the quality of an item. The point is whether such a distortion would influence item characteristics or not, and if so, what would be the extent of this effect. Although using statistical operations would remove much of subjectivity about the quality of the items, there has not been solid empirical evidence to support the validity of the suggestions made in testing books. The purpose of this study, therefore, was to evaluate the effects of violating item construction rules. More specifically, the following research questions were formulated.

1- Is there any relationship between the violation of item-writing principles and item characteristics of the Michigan test of grammar?

2- Is there any relationship between the violation of item-writing principles and the reliability of the Michigan test of grammar?

To put the issue in an appropriate context, first a brief overview of the advantages and disadvantages of multiple choice items will be provided. Then the details of this study will be presented, the findings will be

explained, implications and applications will be discussed, and suggestions will be made.

2. An Overview

Multiple-choice items are the most common type of selected-response items which usually consists of a problem and a list of suggested solutions. The problem may be stated as a direct question or an incomplete statement, and it is called the stem of the item. The list of suggested solutions may include words, numbers, symbols, or phrases, and are called alternatives, (choices, or options). The examinee is typically required to read the stem and the list of alternatives and to select the one correct, or best, alternative. The correct alternative is called merely the answer, and the remaining alternatives are called distracters. These incorrect alternatives receive their name from their intended function i.e., to distract those examinees that are in doubt about the correct answer (Gronlund and Linn, 1990).

Multiple-choice items are widely used due to their versatility in assessing a range of learning objectives. Multiple-choice items have many general positive characteristics mentioned in the literature. Some of which are summarized as follows: (Wiersma and Jurs, 1990; Gronlund, 2003, Popham, W.J. 2000; Thorndike and Hagen, 1969)

Versatility: MC items are adaptable to the measurement of a wide variety of learning outcomes including reasoning, making inferences, solving problems, exercising judgment and demonstrating knowledge of facts through interpretation and analysis of information.

Efficiency: Because of the large number of items that can be posed in a given length of time, MC items permit wide sampling and broad coverage of the content domain.

Scoring accuracy and economy: Expert agreement on the correct answer to MC items is easy to obtain, and they can be machine scored.

Reliability: Consistency in scoring and wide sampling of content provides test results that can be generalized to the domain of interest.

Diagnosis: Patterns of incorrect responses can provide diagnostic information about the learning of individual students or groups.

Control of difficulty: The level of difficulty of a test can be increased or decreased by adjusting the degree of similarity among the options to the items.

In addition, compared to other types of selected-response item family such as true false and matching, multiple choice items are free from the shortcomings specific to these types. Gronlund and Linn (1990) believe that multiple-choice items are superior to the true-false items in that pupils cannot receive credit for simply knowing that a statement is incorrect; they must also know what is correct. Further, multiple-choice items tend to be more reliable than the true-false item, because as the number of alternatives is increased from two to four or five, the opportunity for guessing the correct answer is reduced, and the reliability is correspondingly increased. Finally multiple-choice items are advantageous over the matching exercise in that the need for arranging homogeneous materials is avoided. The matching exercise requires a series of related ideas to form the list of premises and alternative responses. In many content areas it is difficult to obtain enough homogeneous material to prepare effective matching exercises. But this problem is avoided by multiple choice items because each item measures a single idea.

Despite its superiority, multiple-choice items do have certain disadvantages and limitations. Some of these shortcomings are summarized below:

1- Multiple-choice tests are difficult and time consuming to develop.

2- There is a tendency for items to focus on low level learning objectives. In other words, there is a tendency to write items requiring only factual knowledge rather than higher level skills and understanding.

3- Multiple-choice items, like other paper-and-pencil tests, measure

whether the examinee knows or understands what to do when confronted with a problem situation, but it cannot determine how the examinee actually will perform in that situation.

4- Performance on MC items can be influenced by student characteristics unrelated to the subject of measurement, such as reading ability, deductive reasoning, the use of context clues and risk taking.

5- As with other types of selection items, the multiple-choice item requires selection of the correct answer, and therefore, it is not well adapted to measuring some problem solving skills in mathematics and science or to measuring the ability to organize and express ideas.

6- Multiple-choice items have a disadvantage not shared by the other item types: the difficulty of finding a sufficient number of incorrect but plausible distracters. This problem is especially acute at the primary level because of the student's limited vocabulary knowledge in any particular area.

Hoffman (1962) also criticized multiple-choice items because of their strict concern with the answer and not with the quality of thought behind it or the skills with which it is expressed. However, to avoid such shortcomings and improve the quality of multiple choice items, many suggestions are made, some of which are mentioned below.

Suggestions for Constructing Multiple-Choice Items

Recognizing the fact that even skillful item writers might make mistakes and that these mistakes are multiplied by beginners, educational measurement texts usually provide a set of recommendations or guidelines designed to improve the effectiveness of multiple-choice items. A list of such guidelines extracted from several books on educational measurement is presented here. (Aiken, 2000; Chattererji, 2003; Farhady, Birjandi, Jafarpoor, 1994; Johnson D.W. & Johnson, R.T. 2002; ; Mckeachie, W.J. 1999; stanley and Hopkins, 1972; Trice, A.D. 2000; Wiersma and Linn,

1990). Considering these guidelines, one can conclude that in developing good multiple-choice items, test developers often focus on writing appropriate stems and logical options. The following is a brief list of frequently mentioned suggestions.

1. The stem of the item should clearly formulate a problem. The stem should be worded so that the test taker clearly understands what problem or question is being asked before he reads the choices. If the stem does not clearly specify the problem, the alternatives would serve as true-false items.

2. As much of the item as possible should be included in the stem and the options should be kept as short as possible. In the interests of economy of space, economy of reading time, and clear statement of problem, test makers should word and arrange the item so that the choices can be kept relatively short. Wordy stems increase the reading time at the expense of answering time, making the test inefficient.

3. Negative statements should be avoided or used sparingly in the stem of an item. Because negative statements are likely to be ignored by the examinees. Further, negative stems when combined with the answer choices not only present reading problems but also provide the teacher with little information concerning the knowledge that a student has. However, there are times when it is important for the student to know the exception or to be able to detect errors. For these purposes, a few items with "not" or "except should be underlined and/or capitalized to call the student's attention to it. Terraviova (1969) found that items with positive stems were easier than those with negative stems, but reliability seemed unaffected.

4. Novel materials should be used in formulating problems to measure understanding or ability to apply principles. Most teacher-made tests focus too closely on rote memory of the materials and neglect measurement of the ability to use. The multiple-choice item is well adapted to measuring understanding but a novel situation must be presented to the student if more

than rote memory is required to answer the question.

5. All distracters should be plausible. The purpose of a distracter is to distract the uninformed away from the correct answer. To the examinee who has not achieved the learning outcome being tested, the distracters should be as attractive as the correct answer and preferably more so. One factor contributing to the plausibility of distracters is their homogeneity. If all of the alternatives are homogeneous with regard to the knowledge being measured, the distracters are more likely to function as intended.

6. There should be one and only one correct or clearly best answer.

7. The stem should not provide any unintentional clues which might help the examinee find the correct response without understanding the item. Inexperienced test constructors frequently give away clues that permit the examinee to eliminate one or more of the distracters from consideration. Dunn and Goldstein (1959) have shown that items containing irrelevant cues or specific determiners, correct answers consistently longer than incorrect answers, and grammatical inconsistencies between the stem and the options are easier than items without such faults (Thorndike and Hagen, 1969).

8. Options such as "all of the above" or "none of the above" should be used sparingly if at all. Williamson and Hopkins (1967) compared the validity and reliability of a "none-of-these" option versus a homogeneous fifth option, and found no significant differences. Some studies of the use of "none of these" as an option have indicated that use of the option makes items more difficult and more discriminating, but other studies have failed to confirm this finding (Wesman and Bennett, 1946; Rimland, 1960).

9. The stem of a multiple-choice item should not be initiated with a blank. This recommendation originates from the concept of meaningful learning. According to the cognitive-code learning theory, language processes start with known information and move towards unknown information. Starting a stem with a blank means that the examinee should

move from unknown to known information. Thus, the flow of information is in the direction opposite to normal flow of information (Farhady, Birjandi, and Jafarpoor, 1994).

10. The relative length of the alternatives should not provide a clue to the answer. Since the correct answer usually needs to be qualified, it tends to be longer than the distracters unless a special effort is made to control the alternatives' relative length. If the correct answer can not be shortened, the distracters should be expanded to the desired length.

11. All distracters should be of similar level of difficulty, with the same token, similar length. If an alternative is exceedingly more difficult than others, either it will be ignored or selected erroneously as the correct response. In either case, such a selection is not based on the knowledge of the examinee but on some sort of wild guess.

12. The correct answer should appear in each of the alternative positions an approximately equal number of times but in random order. Some teachers bury the correct answer in the middle of the list of alternatives. As a consequence, the correct answer appears in the first and last positions far less often than it does in the middle positions. This, of course, provides an irrelevant clue to the alert examinee. Further, in placing the correct answer in each position approximately an equal number of times, care must be taken to avoid a regular pattern of responses. A random placement of correct answers is most desirable.

13. Verbal associations between the stem and the correct answer should be avoided. Frequently a word in the correct answer will provide an irrelevant clue because it looks or sounds like a word in the stem of the item. Such verbal associations should not permit the examinee who lacks the necessary achievement to select the correct answer. However, words similar to those in the stem might be included in the distracters to increase their plausibility. Pupils who depend on rote memory and verbal association will

then be led away from, rather than to, the correct answer.

14. It is recommended that the stem be a direct question. Although there is no research evidence to support the preferability of the direct question lead over the incomplete statement, it has been found in practice that the novice item writer will produce fewer weak and ambiguous items if the direct question lead is used.

15. All the alternatives must be grammatically correct and consistent with the stem. Distracters, however, should prove wrong when they are placed in the stem. Using wrong expressions as alternatives is quite useless because they will be automatically ignored by the examinees.

It should be mentioned that multiple-choice items should be used where most appropriate. Although the multiple-choice item has many valuable features, there are some subjects for which it is less suitable than other item formats. It should also be mentioned that any of the above mentioned suggestions can be ignored when the test writer has a good reason for doing so. Although these rules provide valuable guidelines for constructing multiple-choice items, test writers may encounter instances where an exception to the rule may improve the item.

3. METHOD

Participants

This study was conducted at Teacher Training University in Tehran. The participants were 50 male and female students majoring in TEFL. The sample selected for this study was a cluster sample consisting of two classes of freshman students.

Instrumentation

The data for this study was collected through two paper and pencil tests:

(1) A 40-item Michigan Test of English Grammar (MT).

(2) A distorted form of the same Michigan Test of English Grammar (DMT).

The reason for choosing the MT was that this test besides being standard was almost free from the flaws usually found in local tests. For constructing the second experimental test, faults were added to MT items which were originally free from item-writing faults. That is, the items were intentionally contaminated by adding to each item one of the following flaws:

(1) Implausible alternatives

(2) Grammatical clues

(3) Ungrammatical alternatives

(4) Blank at the beginning of the stem

(5) Unequal length of alternatives

(6) "all of the above" and "none of the above"

(7) Redundant wordings of the alternatives

(8) Grammatical inconsistency of the alternatives with stem.

Table 1 illustrates the rules which have been violated with regard to the number of the contaminated items. It is attempted to include only realistic and reasonably subtle faults.

Table 1: List of items with their corresponding distortion

	The Violated Rule	No. of Items
1	All alternatives should be plausible	3, 11, 21, 27.36
2	The stem should not provide any grammatical clue.	2, 5, 7, 18, 23, 26, 28, 35
3	All the alternatives must be grammatically correct.	6, 9, 13, 17, 20
4	Stem should not be initiated with a blank.	10, 16, 25, 33
5	All distracters should be of similar length.	8, 19, 39
6	Use of "all of the above" and "none of the above" is not recommended.	1, 4, 12, 24, 32, 38
7	The stem should include as much of the item as possible.	14, 22, 29, 31, 37, 40
8	All alternatives must be grammatically consistent with the stem.	15, 30, 34

Note: In the underlined items some violation of rules was observed in the original form of the test.

The first draft of the Distorted Michigan Test (DMT) was reviewed by testing experts and suggestions were implemented and necessary modifications were made. The following is an example of how the distortion was made to accommodate the purpose of this study.

Rule	MT	DMT
1	3. "Will the entertainment begin right away?" "No. John will give ..... speech before we begin the song." a. him b. himself c. his d. its	3. "Will the entertainment begin right away?" "No. John will give ..... speech before we begin the song." a. him b. himself c. his **d. fewer***
2	5. "Are you going to leave now?" "Unless you would prefer me .... here." a. stay b. will stay c. that I stay d. stay	5. "Are you going to leave now?" "Unless you would prefer me **to** .... here." a. stay b. will stay c. that I stay d. stay
3	9. "Where is your hat?" "It was windy yesterday, and it ---- into the river." a. blows b. to blow c. blown d. blew	9. "Where is your hat?" "It was windy yesterday, and it ---- into the river." a. blows b. to blow c. blown **d. blowed***

* Distorted Item

The resulting experimental instrument contained items matched on content. It should be noted that only seven items out of the 40-item MT were left intact. The reason was that they were originally contaminated with some faults such as ungrammatical, inconsistent and implausible alternatives.

Procedure

This study was conducted in two phases. In the first phase, 29 students received the original Michigan Test of English Grammar and 21 students received its distorted form. In the second phase, which was conducted one week later, each group received the type of the test which it has not received in the first phase (i.e. either MT or DMT). The reason for dividing the students into two groups (group A: MT first, DMT second; group B: MT second, DMT first) was to neutralize the possible test order effect. Since

there was no significant difference between group A and group B with regard to the administration time for each of the two tests, the two groups were then considered as one group for data analysis. That is, all participants had two scores: one on MT and one on DMT. The two tests were administered in the same fashion and the time allocated to each test was 20 minutes. Within this time limit the students were required to attempt 40 grammar items. It should be noted that after the administration of the first test, the students were not informed that they would be required to participate in another test a week later.

4. RESULTS

The data were analyzed based on the groups of items in which a particular rule of item construction was violated. Tables 2 represent item facility and item discrimination values for each item in both distorted and undistorted forms. The mean item facility and item discrimination values are also presented.

Table 2: IF and ID values of the items

Rule No.	Item No.	IF		ID	
		MT	DMT	MT	DMT
1	3	0.66	0.58	0.36	0.84
	11	0.56	0.52	0.72	0.56
	21	0.84	0.78	0.32	0.36
	27	0.78	0.78	0.36	0.36
	X'	**0.67**	**0.64**	**0.45**	**0.49**
2	2	0.92	0.92	0.16	0.08
	5	0.78	0.88	0.4	0.16
	7	0.7	0.88	0.24	0.16
	18	0.9	0.9	0.2	0.2
	23	0.88	0.92	0.08	0.16
	28	0.86	0.92	0.48	0.16
	35	0.56	0.9	0.56	0.04
	X'	**0.78**	**0.89**	**0.28**	**0.13**
3	6	0.6	0.94	0.15	0.6
	9	0.82	0.7	0	0.28
	13	0.48	0.52	0.48	0.4
	20	0.74	0.72	0.36	0.32
	X'	**0.67**	**0.73**	**0.28**	**0.37**
4	10	0.92	0.84	0.16	0.24
	16	0.96	0.94	0.08	0.04
	25	0.72	0.6	0.32	0.48
	33	0.46	0.74	0.04	0.28
	X'	**0.76**	**0.78**	**0.15**	**0.26**
5	8	0.92	0.94	0.6	0.12
	19	0.96	0.8	0.08	0.24
	39	0.66	0.66	0.52	0.52
	X'	**0.84**	**0.80**	**0.40**	**0.29**
6	1	0.94	0.56	0.12	0.64
	12	0.7	0.28	0.2	0.48
	24	0.86	0.32	0.2	0.48
	32	0.8	0.58	0.16	0.52
	38	0.24	0.52	0.36	0.48
	X'	**0.60**	**0.47**	**0.22**	**0.53**
7	14	0.94	0.92	0.15	0.16
	22	0.74	0.56	0.2	0.56
	29	0.84	0.6	0.15	0.72
	37	0.62	0.56	0.44	0.72
	40	0.48	0.48	0.16	0.56
	X'	**0.68**	**0.59**	**0.19**	**0.54**
8	15	0.9	0.8	0.2	0.32
	30	0.8	0.8	0.15	0.32
	X'	**0.75**	**0.71**	**0.22**	**0.41**
	Total	**0.71**	**0.70**	**0.27**	**0.37**

The table reveals different patterns for the violation of different rules. For example the violation of rule 2, (the stem should not provide any grammatical clue), resulted in a significant difference between the mean scores of item facility *indices* of MD and DMT at 0.05 level of significance. Further, the violation of rule 6, (use of "all of the above" and "none of the above") and rule 7, (the stem should include as much of the item as possible) resulted in significant difference between the mean scores of item discrimination indices of MD and DMT at 0.05 level of significance. Regarding the reliability coefficients, KR-21 formula was utilized. Table 3 presents the reliability estimates for MT and DMT. The difference between the reliability coefficient was checked using the formula by Glass and Hopkins, 1984, was not significant.

Table 3. The reliability indices of MT and DMT

Variable	N of items	Mean	Variance	Reliability
MT	40	29.96	36.61	0.81
DMT	40	29.46	45.56	0.84

5. Discussion and Conclusion

Although the overall performance of the test takers on the two forms of the test showed no significant differences between the mean scores of the subjects on the two forms, violation of certain rules had more effect on the characteristics than did the other rules.

However, it can be observed that there is an inconsistent pattern for individual rules. In some cases the violation of the rule has led to higher mean item facility, and in other cases the reverse has happened. For instance, violation of rules 2, 3 & 4 has lead to easier items which violation of other rules have resulted in more difficult items. Although the differences in either direction do not seem significant, the pattern shows an inconsistency. This implies that violation of item construction rules would lead to inconsistent

performance on the part of test takers. Therefore, it is recommended not to ignore the rules suggested for item construction.

Regarding item discrimination indices, it was revealed that the items did not become systematically more or less discriminating either. However, interestingly enough, only violation of Rule 6 (use of "all of the above" and "none of the above") and Rule 7 (The stem should include as much of the item as possible) caused the contaminated items to become more discriminating than fault-free items.

In general, the answer to the research question that there is a relationship between violation of item writing principles and item characteristics of a Michigan test of grammar is somehow negative. Although violation of rule 2 made some items significantly easier, the reliability indices of the two tests were essentially unaltered. The reason may be twofold: (1) The insertion of other faults throughout the test has nullified the effect of items violating rule 2 on total test, and (2) according to Payne, McMorris and Payne (1975), the reliability indices are not sensitive to small changes in means such as those noted in this study. Therefore, since no significant difference is noticed between the reliability indices of the two tests, the second null hypothesis is retained. That is, there is no relationship between the violation of item-writing principles and the reliability of the Michigan test of grammar.

6. Implications

The results of this study indicated that the performances of the examinees who know the answer to a particular question are not much influenced by the quality of the distracters. According to Payne, McMorris and Pruzek (1975) most people will answer on the basis of some relevant information, not solely on the basis of a set, a fault, or other extraneous information. However, teachers should know that testing students with appropriate items is not the same as testing with inappropriate items.

The logical reasons for application of some rules to item construction were already mentioned. Therefore, it is obvious that the value of such recommendations, despite their weak effectiveness, should not be overlooked to keep the face validity of the items high. For such a decision more comprehensive studies are required. Further, examining the tests such TOEFL and IELTS would enrich the available information on the quality of the distracters. More importantly it would be interesting to investigate the effect of item distortion on the recently developed communicative tests. This would reveal whether such tests are also robust to item constructions rules or not.

References

1- Aiken, L.R. ***Psychological Testing and Assessment*** (10th Edition). Boston, MA: Allyn and Bacon, 2000.

2- Chatterji, M. ***Designing and Using Tools for Educational Assessment. Boston***, MA: Allyn and Bacon, 2003.

3- Dunn, T. F., and Goldstein, L. G., ***Test Difficulty, Validity, and Reliability as Functions of Selected Multiple Choice Item Construction Principles***. Educational and Psychological Measurement, (19). Pp. 171-175, 1959.

4- Farhady, H., Birjandi, P. and Jafarpoor, A. ***Testing English as Foreign Language***. Tehran: SAMT, 1994.

5- Gronlund, N.E. ***Assessment of Student Achievement*** (7th Edition). Boston, MA: Allyn and Bacon, 1990.

6- Gronlund, Norman, E. and Linn, R.L. ***Measurement and Evaluation in Teaching***. (6th Edition). New York: Macmillan Publishing Co., 1990.

7- Hoffman, B. ***The Tyranny of Testing***. New York: Crowell Collier Co., 1962.

8- Johnson, D.W. & Johnson, R.T. ***Meaningful Assessment: A Manageable and Cooperative Process***. Boston, MA: Allyn and Bacon, 2002.

9- McKeachie, W.J. Teaching Tips: ***Strategies, Research, and Theory for College and University Teachers*** (10th Edition). Boston, MA: Houghton Mifflin Company, 1999.

10- Payne, D.A. and McMorris, R.F. ***Education and Psychological Measurement: Contributions to Theory and Practice***. 2nd ed. USA: General Learning Corporations, 1975.

11- Popham, W. J. ***Modern Educational Measurement: Practical Guidelines for Educational Leaders*** (3rd Edition). Boston, MA: Allyn and Bacon, 2000.

12- Rimland, B. ***The Effects of Varying Time Limits and Using "Right answer not given" in experimental forms of arithmetic Tests***. Educational and Psychological Measurement Journal (20) pp. 533-540, 1960.

13- Terraviova, C. ***The Effects of Negative Stems in Multiple Choice Test Items***. Unpublished Doctoral dissertation. State University of New York at Buffalo. Ann Arbor, Michigan: University Microfilms, 1969.

14- Thorndike, Robert L. and Elizabeth Hagen. ***Measurement and Evaluation in Psychology and Education***. 3rd ed. New York: John WHey and sons, Inc., 1969.

15- Trice, A. D. ***A Handbook of Classroom Assessment***. New York: Addison, 2000.

16- Wiersma, William and Stephen G. Jurs. ***Educational Measurement and Testing***. U.S.A.: Simon and Schuster, Inc., 1990.

17- Williamson, M. L. and Hopkins, K. D. ***The Use of "None of the above" verses homogeneous alternatives in multiple Choice Tests***: Experimental Reliability and Validity Comparisons. Journal of Educational Measurement, (4) pp. 53-58, 1967.