choice items: implications for reliability and validity of objective achievement tests. **Educational and Psychological Measurement**, 32, 1035-1038.

Ebel, L. R. (1969). Expected reliability as a function of choice per item. **Educational and Psychological Measurement.** London: Taylor and Francis (Printers) LTD.

Grier, B. (1975). The number of alternatives for optimum test reliability. **Journal of Educational Measurement**, 12 (2), 109-112.

Guilford, J. ,& Fruchter, B. (1978). **Fundamental Statistics in Psychology and Education** (6th Ed.). New York: McGraw-Hill, Inc.

Lord, M. F. (1977). Optimal number of choices per item: A comparison of four approaches. **Educationl and Psychological Measurement**, 14, 33-38.

Lord, M. F. (1990). Reliability of multiple-choice tests as a function of number of choices per item. **Journal of Educational Psychology**, 35 (3), 175-180.

McNammara, T. (1996). **Measuring Second Language Performance.** Longman.

Meherans, W. A. , & Lehmann, I. J. (1978). **Measurement and Evaluation in Education and Psychology** (2nd Ed.). New York: Library of Congress.

Oller, W., Jr., & Perkins, K. (1978). **Language in Education: Testing the Tests.** Mass.: Newbury House Publishers, Inc.

Popham, W.J. (1978). **Criterion Referenced Measurement.** New Jersey: Prentice Hall.

Straton, G. R., & Catts, M. R. (1980). A comparison of two, three, and four-choice item tests given a fixed total number of choices. **Educational and Psychological Measurement**,

Thorndike, L., & Hagen, E. (1961). **Measurement and Evaluation in Psychology and Education.** New York: John Wiley & Sons, Inc.

Tversky, A. (1964). On the optimal number of alternatives at a choice point **Journal of Math ematical Psycho** , 1, 386-391.

Valette, R. V. (1977). **Modern Language Testing: A Handbook.** New York: Macmillan.

Wayne, S., & Lioyd, G. (1953). Item reliability as a function of the omission of misleads. **The American Psychologist,** 8, 460-461.

test in using it for different purposes. That is, no single test would be appropriate for all purposes. With the same token, it should be warned that one single fixed pattern (whether three, four, or five choice items) may not serve the purpose of measurement in all contexts. For instance, research on language assessment has demonstrated that the level of language ability of the test takers is an important factor in the testing process. That is, there are significant differences between elementary and advanced students regarding factors such as test form, item form, and the way tests are constructed. There is ample evidence that as the level of proficiency increases, the effect of such ability-independent factors decreases.

Therefore, the level of language proficiency of the expected test takers might be a significant factor in deciding on the number of choices. Although it is an open question, it can be claimed that the number of options should correspond to the language ability levels of the testees. That is, at elementary levels, due to the limitations of language elements, three option tests may serve the purpose better than four or five option tests. However, for advanced students, test developers can maneuver because they have more language data at their disposal to include in test items which makes constructing four, or five options a manageable task. Furthermore, having three choice items would decrease the time requirement allowing test developers to include more items in the test. This is an important point to be taken into account because when there is ample time for testing, teachers can move towards assessment, i.e., using other tasks to measure students' achievement. Even if the teachers spend the extra time they gain on just increasing the number of items of the test, it would cover more materials to be tested and would make the test more comprehensive which would automatically improve the content validity of the test.

Thus, the findings of this research can be summarized in the following practical suggestions:

Teachers can safely use three-option test items instead of four, or five-option tests because:

1. increasing the number of options does not necessarily improve the quality of multiple choice language tests, and

2. the extent of guessing the correct choice when the student does not have sufficient knowledge would not influence the quality of multiple choice language tests.

REFERENCES

Bachman, L., & Palmer, A. (1996) **Language testing in practice.** OUP. Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: some empirical evidence for a mathe- matical proof. **Educational and Psychological Measurement**, 30, 353-358.

Costin, F. (1972). Three-choice versus four-

From practical point of view, the findings of this research may have more remarkable significance. First, most teachers find 5- and 4- choice items quite difficult to constuct. Specially, at elementary levels, where the teachers are bound to the limited amount of lexicon and structures that students have covered, finding four plusible choices seems a demanding job. Thus, allowing for 3-choice items would be an acceptable alternative form for multiple choice items and would relieve the burden. As a matter of fact, in some cases insisting on four choices would either lead to language-wise wrong or semantically irrelevant options. For example, consider the following item:

Mary is this student. She ..... a pen and a pencil in her bag.

a. have b. has c. had d. .....

Through this item, a teacher wants to measure students' ability in recognizing and comprehending the meaning of the word "have" as "possession". Clearly, constructing four choices for such an item is almost impossible because the verb "have" includes only the three forms of "have, has, and had". Therefore, if the teacher is forced to prepare the fourth choice, it would be either remote in structure or obvious in deviation. Since a three choice item is as good as a four choice item, the teachers should feel comfortable with developing MC items with three choices.

Second, a major concern of test organizations is the economy of tests regarding both administration and scoing the tests. Since three option tests seem as good as, if not better, than other types of tests, utilizing three-option tests would save a considerable amount of time and energy. This is particularly important in high stake test situations where a large number of people take the test and expect quick announcement of the results. It is also important in classroom situation because teachers would spend less time on preparing the tests and the students would need less time to perform on the tests.

Third, allowing for three-option tests would increase the quality of the tests because test developers are not forced to use implausible distractors just for the sake of having four-choice items. This is even more important where the test developes are not professionals and do not possess a solid background in measurement and psychometrics. This is particularly useful if new trends in testing and measurement are taken into account. Recent developments in measurement, especially in language education, emphasize assessment of students' performance utilizing multiple sources of information. Assessment no longer takes place through single one-shot case testing, no matter how reliable and valid a test may be (McNammara, 1996). Thus, using three-option tests, teachers would be able to spend the saved time and energy on developing and using other instruments to complement test scores. Certainly multiple sources of information would lead to a better assessment of the students' achievement.

It should be noted that as Bachman and Palmer (1996) claim, one should not be deceived by psychometric characteristics of a

With chance score					Without chance score				
	St	Voc	RC	Cloze		St	Voc	RC	Cl
St	1				St.	1			
Voc	.65	1			Voc	.67	1		
RC	.53	.61	1		RC	.56	.65	1	
Cloze	.76	.60	.61	1	Cloze	.79	.63	.62	

Table 7. Correlation Matrix of Tests Form B

With chance score					Without chance score				
	St	Voc	RC	Cloze		St	Voc	RC	Cl
St	1				St.	1			
Voc	.65	1			Voc	.67	1		
RC	.53	.61	1		RC	.56	.65	1	
Cloze	.76	.60	.61	1	Cloze	.79	.63	.62	

Table 8. Correlation Matrix of Tests Form C

With chance score					Without chance score				
	St	Voc	RC	Cloze		St	Voc	RC	Cloze
St	1				St.	1			
Voc	.61	1			Voc	.54	1		
RC	.62	.62	1		RC	.61	.58	1	
Cloze	.68	.61	.66	1	Cloze	.68	.61	.67	1

The purpose of the last analysis was twofold 1) to compare the reliability indexes of the three forms of the tests with each other and 2) to examine the differences between the validity indexes. To compare the reliability indexes, the F ratio formula (Guilford, 1978, p. 165)was used. The results revealed that none of the obtained F values except for the F ratio of tests of vocabulary forms A and B, could exceed the critical F, which was 1.39 at the .05 level of significance.

To examine the differences among the correlation coefficients of the tests, the Fisher's transformation to Z formula (Guilford, 1978, p. 163) was utilized. In this section also, no significant difference was observed. In other words, none of the correlation coefficients could exceed the Z critical, which was 1.64 at the .05 level of significance. From the results obtained in this analysis it can be concluded that the number of options does not have any significant effect on the test qualities.

Discussion, Implications, and Applications

The study yielded no significant differences among reliability and validity coefficients of the three forms of the tests. The findings are in line with those of the previous studies; namely, increasing the number of options from three to four or five does not necessarily improve the test quality. Thus, certain theoretical implications can be drawn and practical applications can be made on the basis of the findings of this research.

Theoretically speaking, the common belief about the multiple choice items that as the number of options increases the extent of guessing decreases and thus test quality improves may not be taken for granted. In other words, removing the guessing factor, which has been one of the strong motivations for increasing the number of alternatives in MC tests, does not necessarily make significant changes in test quality. This implies that efficiency of the test can be determined without being very much concerned about the number of options. Furthermore, employing a strict convention that 5 choice items are superior to four- and three-option items may not be warranted. Deciding on the number of choices could be a matter of practicality rather than a theroretical requirement.

with chance score and once removing chance score. The descriptive statistics for the study measures are presented in Tables 1, 2, and 3.

Table 1. The Descriptive Statistics of Tests Form A (5 option tests)

	With Chance Score			Without Chance Score	
	X	S		X	S
St	24.95	8.75	St	21.49	7.90
Voc	17.09	4.48	Voc	14.42	5.22
RC	14.77	5.32	RC	12.72	5.83
Cloze	21.07	6.39	Cloze	18.49	7.24
Total	56.82	14.43	Total	48.63	16.55

Table 2. The Descriptive Statistics of Tests Form B (4 option tests)

	With Chance Score			Without Chance Score	
	X	S		X	S
St	26.45	6.62	St	22.67	8.63
Voc	17.76	4.20	Voc	14.63	5.44
RC	16.26	5.30	RC	13.60	6.23
Cloze	21.68	5.59	Cloze	18.03	6.82
Total	60.50	13.80	Total	50.90	17.59

Table 3. The Descriptive Statistics of Tests Form C (3 option tests)

	With Chance Score			Without Chance Score	
	X	S		X	S
St	26.89	6.99	St	21.30	9.57
Voc	18.16	3.90	Voc	13.69	5.27
RC	16.00	4.90	RC	12.42	5.83
Cloze	22.28	5.38	Cloze	17.40	7.04
Total	61.05	13.70	Total	47.41	17.62

The second analysis was done to determine the reliability coefficients of the tests. The K-R 21 formula was utilized, once with chance score and once without chance score. Table 4 represents the reliability indexes of the three forms of the test scores before removing the chance score. The reliability estimates of the three forms of the test scores corrected for chance score are shown in Table 5.

Table 4. Reliability Coefficients of the Study Measures With Chance Score

Sub-test	Form A	Form B	Form C
Structure	.82	.82	.84
Vocabulary	.65	.46	.53
Reading Comp.	.76	.76	.71
Cloze	.82	.76	.74
Total TOEFL	.90	.88	.75

Table 5. Reliability Coefficients of the Study Measures Without Chance Score

Sub-test	Form A	Form B	Form C
Structure	.87	.89	.91
Vocabulary	.83	.77	.75
Reading Comp.	.81	.83	.81
Cloze	.85	.83	.85

The third analysis was conducted to determine the validity indexes. After obtaining reasonably high reliability indexes for the three forms of the tests, the concurrent validity of the tests was investigated. The concurrent validity indexes of the tests were computed through the correlations between the subtests of the TOEFL and cloze. The correlation coefficients among the subtests of TOEFL, cloze in Form A, Form B, and Form C are reported in pairs of Tables 6,7 and 8 respectively. In each pair of tables, one represents the correlation coefficients with chance score and the other without chance score. The correlation coefficients range from moderate to high which means that the three forms of the tests were as valid as the criterion measure.

Table 6. Correlation Matrix of Tests

With chance score					Without chance score			
	St	Voc	RC	Cloze		St	Voc	RC
St	1				St	1		
Voc	.70	1			Voc	.69	1	
RC	.60	.60	1		RC	.61	.60	1
Cloze	.72	.51	.59	1	Cloze	.75	.55	.62

the psychometric characteristics of the individual items as well as the whole test when the number of alternatives ranges from three to five. Second, since no empirical study has been reported in which the guessing factor was considered as a variable and no research was conducted on the language skills, the present study attempts to provide empirical evidence for the effect and appropriacy of the chance factor interacting with the number of alternatives in a language skills test with MC items. Thus, the following research questions were posed:

1. Is there any relationship between the number of options and the characteristics of language proficiency tests?

2. Is there any relationship between the guessing factor and the characteristics of language proficiency tests?

METHOD

Subjects

A total of 431 senior English majors from differnt universities of Tehran participated in this study. All subjects were within a similar age range. They were from different parts of the country being admitted to the universities in Tehran. These universities included Tehran University, Shahid Beheshti University, University for Teacher Education, Allameh Tabatabai University, and Azad University.

Instrumentation

Two tests of language proficiency were used in this study:

1. An original TOEFL

2. An already established MC cloze test.

These tests were designed in three forms labeled as: Form A, tests with 5-option items, Form B, tests with 4-option items, and Form C, tests with 3- option items. All the forms included four subtests, i. e., structure (40 items), vocabulary (30 items), reading comprehension (30 items), and a cloze test (35 items).

Procedures

The administration of the tests was accomplished through three phases. At the first stage, tests with 5-option items (labeled Form A) were administered. Through item analysis, the least effective distractor was eliminated from each item to develop the tests with 4-option items (labeled Form B). Then the tests with 4-option items were administered to a group similar to the first group. Then again, through item analysis, the least effective alternative was eliminated from each item to form the tests with 3-option items (labeled Form C). The three forms of the tests were administered using a counter balanced procedurd, i,e,. each student taking only one form of the test on a random basis. The time required for test forms A, B, and C was 125, 100, and 80 minutes respectively. Along with these tests, every subject took the cloze test. The scores on the cloze test were used for the purpose of validation.

RESULTS

To answer the research questions, different analyses were conducted. It should be mentioned that all the tests were scored twice: once

alternatives decreases.

In this regard, measurement textbooks typically recommend four or five alternatives for multiple-choice items. Some scholars believe that the greater the number of plausible options, the higher the reliability of the test would be (Thorndike and Hagen, 1961; Meheren and Lenhem, 1978). Furthermore it is believed that with the greater number of options per item, the probability of getting an item correctly without sufficient knowledge, i. e., the guessing factor, would decrease. Although, the effect of the guessing ability could be compensated for through the application of the correction for guessing formula, it has, as yet, remained an unresolved issue.

Contrary to the above mentioned suggestions, some scholars have made theoretical arguments (Lord, 1977, 1990; Tversky, 1964; Ebel, 1969; Grier, 1975; to name a few) and some have reported empirical evidence (Costin, 1970, 1972; Straton and Catts, 1980) that an increment in the number of options does not necessarily lead to an increase in test quality.

On the theoretical dimension, Grier (1975) presents evidence that the expected reliability of the tests with 3-option items is higher than those with 2- 4- and 5-option items. In another theoretical study, Ebel (1969) has predicted, according to the formula he developed, that an appreciable increase will occur in the reliability of an objective test when the number of choices is increased from two to three, a smaller increase when 4-choice items are used, and a still smaller increase beyond that point. Also, Tversky (1964), in a theoretical paper, has demonstrated the superiority of 3- option items over other types of items.

On the empirical dimension, several studies have suggested the superiority of the 3- option items over 4-and 5-option items. The results of research project conducted by Wayne, et al, (1953) revealed that the efficacy of an MC test can be improved by deleting the misleads which item analysis discloses to have either weak discrimination or facility indexes. In addition, the findings of another research project based on Tversky's mathematical model indicated that the 3-option items enjoyed higher item characteristics than 4-or 5-option items. In another attempt, Trevisan and Sax (1991) compared the reliability indexes of tests with 2-3-4- and 5- option items. Their findings revealed no significant difference in the reliabitlity indexes of these tests. And finally, Frowman (1987) supported the efficacy of MC tests with 3-option items over tests with 4 or 5 option items.

Although most of the research cited above support the superiority of 3-option items over other types of items theoretically and empirically, they do not answer all the questions related to the issue. First of all, they do not address the validity of the findings with regard to non native speakers. Second, they do not offer comparative results on both item and test characteristics. More importantly, they do not consider the effect of the guessing factor on the results. Therefore, this paper addresses the issue. More specifically, the purpose of this paper is twofold. First, it attempts to compare

It should be mentioned that there is sometimes a good relationship between practicality and other characteristics of a test. For instance, oral interview and composition type tests, though highly valid, do not often show high reliability. At the same time, they do not show high practicality, either. Of course, this does not imply that an impractical test is unreliable, or an unreliable test is impractical. Rather, in such cases, one should pay attention to the source of unreliability or impracticality. One factor which contributes to both practicality and reliability is the objectivity of scoring a test. And one major type of item which is scored quite objectively is the multiple choice (MC) type. In its traditionl form, an MC item includes a stem which is intended to elicit information and a number of choices. The test taker is required to read the stem and choose one of the choices as the most suitable one to complete the stem. The number of choices varies from two (the minimum number) to five (the common maximum number), depending on the purpose of the test and the intention of the test developer.

The MC item is undoubtedly one of the most widely used item formats. As with many other types of items, however, MC items, have certain strengths and weaknesses. Regarding the strengths of MC items, Eble (1969) states that MC tests are adaptable to the measurement of most important educational outcomes. Meherans and Lenhem (1978) mention versatility as another merit of MC tests. Thorndike and Hagen (1965) believe that an MC item is effective for measuring vocabulary, the degree of understanding, the application of principles, and the ability to interpret data. Oller (1979) claims that MC tests can fulfill the requirements of pragmatic and integrative tests of language proficiency. Still another advantage of MC tests refers to their ease of scoring, i.e., MC items are scored easily, rapidly, accurately and objectively by teachers, scoring machines, and computers. These advantages have caused MC tests to be used in large scale administrations. Further, the accuracy of scoring MC tests leads to consistency of scores which in turn contributes to a higher degree of reliability (Popham, 1978).

In spite of the above-mentioned advantages, MC tests have not escaped the criticisms of the scholars. In this regard, Oller (1978) asserts that preparation of sound MC items is challenging and technically difficult. In addition, testwise students tend to perform better on MC tests than non-testwise students do (Meheran & Lenhem, 1978). Furhermore, Vallette (1977) complains of another drawback of MC items stating that the testees can benefit from guessing. Finally, it is commonly believed that preparing an MC item with sound and plausible alternatives is not an easy task.

Not only would developing reasonable alternatives but also making a decision on their number create some problems for the test makers. On the one hand as the number of alternatives increases, the extent to which testees can guess the correct choice without having the required knowledge decreases. On the other hand, as the number of alternatives increases, the chance of preparing many good

Number of Options and Economy of Multiple Choice Tests

Hossein Farhady Ph.D(TEFL)
Iran University of Science and Technology
& Shirin Shakery
Tehran University

ABSTRACT

A major concern in testing in general, and in language testing in particular, is economy in preparing, administering, and scoring procedures of the tests. In this study, the psychometric characteristics of a language proficiency test comprising four sections, (vocabulary, structure, reading comprehension, and cloze), were investigated as a function of the number of options per item. Parallel tests of 3-4-and 5-option items were administered to 431 seniors majoring in English. The findings revealed no significant difference in the mean item facility, the mean item discrimination and total test qualities. The results supported the findings of previous research in other areas of education that suggest the efficacy of the three-option item tests. The findings have significant implications to test construction, specially in large scale administrations. The findings are also applicable to areas of testing other than language proficiency.

INTRODUCTION

Scholars in the field of measurement have often been concerned with three major characteristics of the testing devices, namely, reliability, validity, and practicality. Though many arguments have been made for and against preferring one characteristic over the others, it does not seem reasonable to sacrifice one for the other two characteristics. The ideal case is to have a balance regarding the three characteristics. Of cours, reliability and validity can be computed through certain statistical techniques. Practicality, however, is a relative term and depends on many factors such as the nature and form of the items, and above all, on the administrative facilities. That is, a particular test may be quite practical in one occasion or for a particular institution but not in another occasion or for another institution.