arab journal for the humanities

Validating a Multiple-Choice Cloze Test to Assess the Proficiency of EFL Learners' Writing Skills

Ibrahim S. Al-Fallay*

* Ph.D. in Language Testing, New Mexico, U.S.A, 1994
* Lecturer, Dept. of English, King Saud University, Riyadh, Saudi Arabia.

 1026-9576/99/17/65/274-308

Abstract

Assessing learners' ability in EFL writing skills has been a major component of most conventional, standardized placement and proficiency tests. In both types of tests, learners are usually given a prompt or prompts to which they are expected to respond in a composition format. Their responses are usually marked manually, either analytically or holistically. Since such tests are normally administered to a huge number of students, and since raters are usually given a short period of time to report students' scores, the need for a tool to assess EFL writing ability quickly, efficiently, and practically is persistent. This study, therefore, aims at developing and validating a Multiple-choice Cloze (M-C) test, devised according to the rational procedure in deletion, and geared to assess the writing ability of EFL learners. The components around which the test's items are built are those generally regarded by EFL instructors as essential and to be included in marking grids. 52 intermediate EFL subjects took the M-C test. They were also asked to respond to a prompt in a composition format. The study found that the reliability of the M-C test is similar to that of a composition analytically marked. Furthermore, five different types of validity related to the M-C test were also investigated. High correlations between the M-C test and the composition, marked either impressionistically or analytically were found, which confirm the criterion-reference of concurrent validity of the instrument. The content of the M-C test was analyzed. It was found that the test's items were representative. The unfamiliarity of the subjects with the Multiple-choice Cloze test format and the peculiarity of the self-assessment technique led the subjects of this study to rate the M-C test to be low in its face validity. Furthermore, the M-C test was able to predict learners' future performance in EFL writing tasks.

To examine the construct validity of the M-C test, the "Internal Construct Validation" technique corrected for "part-whole overlap" correlations among the components of the M-C, and the analytically marked composition ANOVA, of the difference among the subjects' means on the M-C test and the composition were calculated. The results indicate that the M-C test's items enjoy a high construct validity. The study concludes by stating that although the M-C test shows high reliability, validity, and practicality, its use should not be overgeneralized until further investigation is conducted. However, the test seems to be a useful tool in assessing EFL writing ability when integrated with other components usually found in placement and proficiency tests.

I. Introduction

Teaching the strategies, principles, and techniques of writing, and enhancing it as a linguistic skill are a major objective of most, if not all, English as a second/ foreign language (ESL/EFL) programs around the world. The writing skill is usually ranked as the fourth element of ESL/EFL programs after the skills of listening, speaking, and reading. This kind of arrangement is not based on the importance of the skills but rather on the natural sequence of first language acquisition by children. Hence, most ESL/EFL teaching methods and approaches have adopted this sequence in skills presentation, where the skills of writing and the active counterpart of reading are often introduced after skills of listening, speaking, and reading have been introduced. Moreover, writing is the most difficult skill that ESL/EFL learners have to deal with. Bereiter and Scardamalia wrote "Writing... is probably the most complex constructive act that most human beings are ever expected to perform" (1983:20). In addition to the difficulty of learning and mastering the skill of writing, its marking or scoring is always deemed to be subjective, no matter what marking method (analytic or holistic) is adopted. On the contrary, the scoring of multiple-choice format tests used to assess the achievement or proficiency of ESL/EFL learners in the skills of listening and reading comprehension, along with their ability to competently manipulate English grammatical structures, is usually done with the help of a machine that reduces the time and efforts needed to grade learners' writing. The adoption of either the analytic or the holistic methods in marking learners' writing, though reflecting different degrees of subjectivity, forces the instructors to abandon the use of the machine in marking; and hence, subjective manual scoring becomes a must.

Bearing in mind the importance of teaching the skill of writing in ESL/EFL programs, the difficulties that face ESL/EFL learners in learning and mastering the strategies, principles, and techniques of writing, and the problems that ESL/EFL instructors experience in marking learners' writings, this study seeks to achieve the following goal: developing and validating a test that can be used in assessing ESL/EFL learners' proficiency in the writing skill. The proposed test has to be valid and reliable and must also be easy to mark. Hence, this test is developed following the multiple-choice format in test construction, enabling learners' compositions to be machine-scored. Such an approach to marking learners' writings has some advantages over the traditional hand-scored approach in marking, in that it is more reliable, objective, and efficient, besides facilitating marking.

II. Review of Selected Literature

Despite its status as an important and essential part of any ESL/EFL program, the skill of writing has not enjoyed the same amount of research devoted to other language skills; namely listening, speaking, and reading (McDonough, 1995). What little research there has been, however, has had diverse objectives: to study the transfer of L1 writing ability to L2 writing tasks (Boyle & Peregoy, 1990; Canale, et al., 1988; Edelsky, 1982; Eisterhold, 1990; Raimes, 1985; Uzawa & Cumming, 1989; Valdés, et al., 1992), to investigate the relationships between the skills of reading and writing in L1 and L2 (Carson, et al., 1990; Clarke, 1978; Cummins, 1981; Flahive & Bailey, 1988; Janopoulus, 1986; McLanghlin, 1987), to study the mechanism of the writing process in either L1 and/or L2 (Arndt, 1987; Hayes & Flower, 1983; Jensen & DiTiberio, 1989; Perl, 1980, 1981), or to investigate the assessment techniques and marking procedures employed in testing L1 and L2 writing abilities of native speakers as well as ESL/EFL learners (Benton & Blohm, 1986; Hamp-Lyons, 1990; Heaton 1988; Oller, 1979; Pollitt & Hutchinson, 1987; Ruth & Murphy, 1988).

A typical writing assessment procedure begins by giving subjects one of several prompts, and the students' task is to respond to the stimulus with a certain number of lines. Although prompt selection may seem an easy and straightforward process, it is in fact very complex because other extraneous factors such as cultural considerations, prior knowledge, prompt type (i.e., expository, narrative, argumentative, etc.), and prompt difficulty level play a role in determining students' performance, not only in L1 writing ability assessment (Brossell, 1986; Freedman, 1983, Hamp-Lyons & Prochnow, 1991; Hoetker & Brossell, 1986, 1989), but also when L2 writing ability is being tested (Carlson et al., 1985; Park, 1988; Reid, 1990; Spaan, 1989). Hamp-Lyons and Prochnow (1991), for instance, investigated the relationship between prompt type and difficulty, and ESL performance in the writing part of the Michigan English Language Assessment Battery (MELAB), a proficiency test similar to the TOEFL administered worldwide to screen prospective foreign students seeking admission to universities where English is the medium of instruction. They found that expository topics were easier for ESL students to write on than argumentative types. Furthermore, they claim that more competent ESL learners usually choose difficult topics as they push them to perform better, whereas less competent writers usually select easier topics of an expository nature. They also believe that raters usually sympathize with students who choose difficult prompts, as reflected in more lenient marking.

When students' compositions are collected, they are always manually marked. Two manual marking methods are usually described in the literature. First, the holistic or impressionistic scoring "involves the assignment of a single score to a piece of writing on the basis of an overall impression of it" (Hughes, 1989:86). In such a method, raters find difficulty in stating the used scales or marking criteria (Casanave, 1995; Charles, 1990); and even when they are able to determine them, their scales and criteria are inconsistent and uneven (Brown, 1991; Leki, 1995; Prior, 1995). However, by using this method, a rater can mark a large number of compositions in a short period of time (Hadley, 1993; Hughes, 1989). Hughes claims that a rater can mark a one-page essay by using this method in less than two minutes. This method, however, is not without disadvantages. It is usually described as a subjective method. Kassen (1990), for instance, purports that when teachers use this method, they do not usually choose to correct the same errors. Furthermore, Harris (1969) reported a low inter-rater reliability coefficient of 0.25 which questions the method's reliability and validity.

The second marking technique involves the analytic method. This method "involves the separation of the various features of a composition into components for scoring purposes" (Hadley, 1993:343). The mark given to each separate feature or criterion varies from one assessor to another, and it is not uncommon for these features and criteria to be allocated different scores. Usually, the final grade given to a student's composition is a composite of the individual estimates granted to the various criteria. Furthermore, the entities of these features are not completely agreed upon by researchers and ESL/EFL instructors. Hughes (1989), for instance, reported a scheme for writing assessment devised by John Anderson, which was based on a scale used to assess oral ability, inaugurated by Harris (1969). Anderson's scheme consists of five features, each with a six-point scale. These features are grammar, vocabulary, mechanics, fluency (style and ease of communication), and form/organization. However, Gaudiani (1981), cited in Hadley (1993:344), proposed a marking scheme that included the following features: grammar/vocabulary (composition comprehensibility), stylistic technique (composition syntax), organization, and content. A five-letter scale (A, B, C, D, and F) is used, which would later be translated into numbers (4, 3, 2, 1, and 0, respectively) to estimate the points given to each component or feature. The sum of these points in then divided by 4, the number of the features. The product is the score granted to the examinee's composition.

Two less common methods employed in writing assessment have also been reported in the literature. Heaton (1988:148) suggested an error-count method which he defines as "counting the errors made by the testee and deducting the number from a given total". Heaton's technique is similar to Oller's method in writing assessment where "essay score = [(the number of error-free words in the student's protocol) minus (the number of errors in the student's protocol)] divided by (the number of words in the rewritten text)" (1979:387). In primary trait scoring, another marking method reported in Perkins (1983), marks are assigned holistically based on the student's performance in a specific feature, such as content, organization, or grammar. For example, if the purpose of the writing task is to assess a student's ability to persuade the reader, the total marks will go to the number and types of arguments that he provides. Needless to say, primary trait scoring would not be a popular choice for teachers teaching and assessing writing in ESL/EFL classroom situations.

What is the best technique of the three, namely the holistic, analytic, and error-count methods? Is it better if the rater employs either one of the two norm-reference techniques: the holistic and analytic methods; or should he use the criterion-reference type, i.e. the error-count technique? The use of each method has opponents and proponents. Holistic marking is certainly faster. To overcome the problem of low inter-rater reliability, Hughes (1989) devised a scale for scoring according to the holistic method which, if used, would increase the reliability of this method up to .90, on condition that every student's composition is scored by two different assessors. However, Perkins (1981) believes that the major drawback of this method is its subjectivity.

Zughoul and Kambal (1983) sought to develop a detailed analytic scoring scale by comparing the writing scores of EFL learners, who came from three different proficiency levels: beginning, intermediate, and advanced, when their proposed scale is used, with the scores obtained by the same learners when the holistic or impressionistic method is employed. They reported that 41.2, the average score of the three levels when the impressionistic method was used, decreased by about 33% when the analytic method was employed. Surprisingly, the inter-rater reliability indices of the impressionistic method ranged from .85 to .81, higher than those of the analytic method; the indices of which were between .78 and .83. They contended that although "analysis of variance and the estimate of inter-rater reliability... showed no significant differences between the two methods... [,] the analytic method has its pedagogical advantages over the impressionistic method" (1983:100).

Brown and Bailey (1985) had a similar objective to that of Zughoul and Kambal, i.e., the development of an analytic scoring grid. However, their results favor the use of the analytic method over the impressionistic, due to the limited reliability and validity of the latter. Furthermore, Cleary (1988) compared the analytic method to the error-count technique in validity and reliability, using the composition of 49 EFL learners that were scored by two scorers. He found that when the error-count method was used, the two scorers came to almost complete agreement which yielded a high reliability index. However, when the analytic method was employed, the scores given by the two raters differed significantly, which indicates that the use of the error-count method is superior to the use of the analytic method. It seems that Cleary's results were in part due to the use of only two scorers.

Whatever method is used, the problem of subjectivity persists. Although the impressionistic or holistic method is invariably accused of being subjective, the analytic method is also subjective. For example, Hughes (1989) reported a six-point analytic scale for five proposed scoring criteria, whereas Gaudiani's (1981) grid has a five-point scale for her four proposed criteria. Furthermore, Zughoul and Kambal (1983) used an analytic scoring grid consisting of five criteria, each with a scale between 0-35 and 0-15. It is obvious that whatever points are given, the estimate is still subjective, and Cleary's results confirm such an observation. In addition, the number of rateas used plays a key role in obtaining a high inter-rater reliability Zughoul and Kambal obtained high reliability indices when the number of raters was five, whereas Cleary's study yielded low reliability indices as a result of using only two raters. The implication is that in order to achieve a reliable score by the analytic method, the number of raters has to be increased. Such a condition is difficult to impose when learners' compositions are scored in EFL writing classrooms where the instructor is solely responsible for such a process, but also in scoring the writing section of proficiency tests. For example, it was reported in the TOEFL Test and Score Manual (1992:37) that the TWE of the TOEFL is marked by only two raters. All in all, there is clearly a need to develop a multiple-choice writing test which could assess ESL/EFL learners' writing ability. Hopefully, such a test would turn out to be promising and helpful, especially in situations where the number of compositions is large and the time given to raters is short.

III. The Present Study

A. The Research Problem

Assessing ESL/EFL learners' writing ability has become an essential component of most reputable proficiency and placement tests. For example, the Educational Testing Service (ETS), producers of the Test of English as a Foreign Language (TOEFL), has introduced a subpart in 1986 (Test of Written English, TWE) which aims at assessing writing proficiency of EFL students (ETS, 1992). The TWE is integrated into six of the twelve annually administered TOEFL tests. Furthermore, Hamp-Lyons and Prochnow (1991:59) claim that the MELAB (Michigan English Language Assessment Battery) "is administered in the US and in 120 countries and over 400 cities around the world". The MELAB has a component especially developed to assess the English writing proficiency of ESL learners. Weigle (1994) mentioned that the English as a Second Language Placement Examination (ESLPE), administered quarterly by the University of California at Los Angeles (UCLA) to assess English proficiency of newly admitted foreign students, has a 50-minute composition writing component, in which students are required to respond to either one of two given prompts.

Given then that the importance of assessing ESL/EFL writing ability is of such status, this study aims at developing and validating a multiple-choice cloze test devised according to the rational procedure in deletion (henceforth, the M-C test) geared to assess that ability. In developing the M-C test, points usually looked for by experienced raters were taken into consideration (see material section). Five types of validity (criterion-reference or concurrent, content, face, predictive, and construct) and reliability indices were also calculated and investigated. The huge number of examinees and the short period of time usually given to raters to turn in examinees' scores make the use of the M-C test in assessing writing proficiency a must, since a large number of responses could be machine-marked in a very short period of time. However, it is noteworthy here that, even if the M-C writing test under investigation turns out to be highly valid and reliable, there is no claim made by this study, that it should replace traditional composition writing tests, especially in achievement attainment where instructors are interested in measuring and checking their students' hand-writing legibility, spelling ability, and/or knowledge of English grammar. Rather, the proposed M-C test, if its validity and reliability are accepted, should be integrated into large scale proficiency and placement tests where manual scoring of students' writings is the only possible procedure.

B. Questions of the Study

Among the major questions that this study attempts to tackle are the following:

1. Is it possible to develop an M-C test that is able to assess EFL learners' ability in the writing skill?
2. Will that M-C test enjoy a reliability index similar to those of compositions written by EFL students and analytically and holistically marked?
3. Is it possible to develop an M-C writing test that contains components usually contained in marking grids used and developed to assess EFL learners' compositions?
4. Do the subjects of this study believe the M-C test to be one that is capable of assessing their EFL writing ability?
5. Will the M-C test be capable of predicting future performance of EFL learners in writing tasks?
6. Will the M-C test be capable of assessing the same constructs assessed by traditional methods such as the responses of EFL learners to a prompt, which are then marked either analytically or holistically?

C. The Hypotheses

The study hypotheses are nondirectional and set at $p < .05$: They are the following:

H1: There are no statistically significant correlations among the subjects' means on the M-C writing test, and their scores in the composition analytically and holistically marked.

H2: There are no statistically significant correlations between the subjects' self-assessment and their performance on the M-C writing test, and the analytical and holistically marked composition.

H3: There is no statistically significant correlation between the subjects' scores on the M-C writing test, and their scores in the composition written by the subjects in the next proficiency level' and which is marked analytically.

H4: There are no statistically significant correlations among the components of the M-C writing test (grammar, vocabulary, mechanics, and unity and organization) and the components of the analytically marked composition

(grammar, vocabulary, mechanics, and unity and organization).

H5: There are no statistically significant differences among the subjects' means on the M-C writing test, and the analytically and holistically marked composition.

D. Method

1. Subjects

The subiects of this study were 52 Saudi Arabian EFL learners (the scores of 17 subjects were discarded becaue they failed the final achievement test). When the study was conducted, they were in the intermediate level, the third level of the Intensive English Program (IEP) of the Institute of Public Administration (IPA), Riyadh, Saudi Arabia. The intensive program consists of four levels, each of eight weeks' duration (one quarter). Prior to their enrollment in the IPA's private sector program, they had obtained diplomas awarded at the secondary level, where they had studied English in high and intermediate schools as a school subject for six years. When new students are admitted to the IPA's programs, their proficiency is assessed by a placement test developed in-house. Those who score 90 points or above (100 points is the highest possible score) are directly admitted to the private sector program where English is the medium of instruction. But students who score below 90 points are placed in the appropriate level of the IEP. Furthermore, the subjects of the study had passed the final achievement tests of levels 1 and 2, which assess the four skills of English, besides Knowledge of English grammar. Finally, all subjects were male students.

2. Materials

(i) M-C Cloze Test

A text, entitled 'Auction Sale', was adopted from Mosback and Mosback (1976:32-33). The Flesch (1984) readability score was 50. The text was then transformed into a 45 M-C Cloze test using the rational deletion method, where only items of interest were deleted regardless of their position. The reliability and validity of the M-C cloze test as a testing procedure were reported, according to the fixed ration deletion method (Al-Fallay, 1997) or the rational deletion procedure (Bensoussam &

Ramraz, 1984). From a review of the studies reported in the literature, the following components were deemed to be important and feasible in terms of measurement: grammar, vocabulary, mechanics, and unity and organization. The first three components and part of the fourth component (cohesion: reference and conjunctive and coordinate elements) were assessed by giving the subjects four alternatives from which they had to choose the most suitable one. However, part of the fourth component was assessed by five M-C questions each of which had four alternatives as well. Hence, the total number of items was 50.

The grammar component was aimed at assessing the subjects' mastery of the following grammatical points: English tense/ aspect forms, subject-verb agreement rules, passive, gerund, infinitive, and word form (noun, adjective, verb, and adverb). The vocabulary component was assessed by giving the subjects four lexical items from which they had to choose the word that best fitted the context. Moreover, mechanics assessment refers to testing the ability of the subjects in using English spelling and punctuation. This ability is, hopefully, reflected by the correctness of their choice from among the alternatives. Finally, the unity and organization component was developed to assess the subjects' familiarity with the following concepts: cohesion, topic sentence, thesis sentence, and summary sentence(s). The concept of cohesion is deemed to be "the relations of meaning that exist within the text, and that define it as a text" (Halliday and Hasan, 1976:4). Hence, cohesion includes, but is not limited only to, the following concepts: ties or reference (mainly anaphora) and coordinate and conjunctive elements. The M-C cloze test used appears in appendix (1).

(ii) The Attitude Questionnaire

To assess the face validity of the M-C cloze test used, a questionnaire developed by Oller and Conrad (1971), used by Lado (1986), and modified by Al-Fallay (1997) was used to investigate whether subjects believe that the M-C cloze test is capable of assessing EFL writing ability. The modified version of the questionnaire appears in appendix (3), with the Arabic version of the questionnaire, given to the subjects, in appendix (4). It was thought that giving the questionnaire to the subjects in their native language would eliminate any extraneous factors, such as the inability to understand

items due to language proficiency limitation, and/or the inability to complete the questionnaire due to time limitation.

(iii) The Production Test

To assess the students' writing ability using the traditional procedure, the subjects were asked to write a 25-line composition as a response to the following prompt: "Describe your first semester at the IPA's (Institute of Public Administration) IEP (Intensive English Program)." Their writing scores on the final achievement test of the fourth level (the one following the level of their current proficiency level when the study was conducted) were also obtained.

3. Procedure

The study took place in February and March 1997. In the sixth week, the subjects were asked to write a 25-line composition as a response to the topic: "Describe your first semester at the Intensive English Program of the Institute of Public Administration." They were given 90 minutes to do the task. In the seventh week, they were given the M-C test. They were given 90 minutes to finish the test. The compositions were rated twice by four EFL writing instructors of the IEP. First, the compositions were holistically rated, and after 30 days, the raters were given the marking scale and asked to remark the compositions analytically. This arrangement was followed to avoid any influence on the raters caused by their familiarity with the marking scale. To guide raters towards an analytic marking of the writing of the subjects, four components of the marking grid for the EFL intermediate level developed by Zughoul and Kambal (1983:96) were adopted: structure, vocabulary, organization, and mechanics. Zughoul and Kambal allocated 25, 20, 20, and 15 points to the assessment of the four components, respectively. This gives a total of 80 points, with 20 points left to the assessment of content, their fifth component, which is not assessed by the M-C cloze test. To calculate the marking out of 100 points, the 20 points devoted to the content was evenly distributed over the other four components. This modified grid appears in appendix (2). The M-C test was manually scored using a punched answer key. A value of (0) was given to wrong answers whereas two points were assigned to correct responses. Finally, the subjects' responses to the writing component of the final achievement test of level 4 (advanced) were obtained. The same

raters were given the marking scale for advanced EFL learners, developed by Zughoul and Kambal (1983:97) and asked to rate the compositions analytically. The same four components of the intermediate scale were included (the content component was ignored). The points given to each component were as follows: structure, 20 points; vocabulary, 15 points; organization, 25 points; mechanics, 10 points.

E. Results and Discussion

The first step in the analysis was to obtain the means ($\bar{X}$), standard deviations (SD), and reliability indices of the subjects' scores on the M-C test, and their scores in the composition marked analytically and holistically. The M-C test reliability was obtained by using Cronbach a statistic whereas the reliability indices of the analytic and holistic scoring were estimated by employing Winer (1971:666), cited in Zughoul and Kambal (1983:99). First Θ value is calculated by using the following formula:

$$\Theta = \frac{\text{MS Between subjects - MS Within subjects}}{(K)\,(\text{MS Within subjects})}$$

where MS is the mean squares and K is the number of raters. Then (r), the reliability index, is obtained by employing the following formula:

$$r4 = \frac{4\Theta}{1 + 4\Theta}$$

Table (1) Means, standard deviations, and reliability indices of the subjects' scores on the M-C test and their scores on the compositions marked analytically and holistically

Test/Method	*X*	*SD*	α./r
M-C test	44.23	11.88	.85
Analytic	42.73	13.93	.83
Holistic	51.33	13.26	.87

When subjects' compositions were scored holistically, they were given higher scores than when the same compositions were analytically rated (8.60 points difference). The mean of the subjects' scores on the M-C was higher than their mean when the holistic method was used, but lower than the mean of their compositions when the analytic method was employed. However, the standard deviation of the subjects' scores on the M-C was the lowest, 11.88, followed by

the standard deviation of the holistically rated compositions. The standard deviation of the subjects' scores when the analytic method was used was the highest. This implies that the M-C reflected little variation among the subjects and a conformity among them towards a more homogeneous group. Although the reliability index of the holistic method was the highest, the other two reliability indices of the M-C test and the analytic method were not far off.

To assess the suitability of the M-C test, its difficulty level for the subjects of the intermediate level, and its ability to distinguish among the EFL subjects, the following were calculated: item facility (IF), the proportion of the EFL subjects who chose the correct responses, and item discrimination (ID), the correlation between each item in the M-C test and the total of the test, as displayed in Table (2)

Table (2) Item analysis (facility and discrimination) of the M-C test items

Item No	IF	ID	Item No.	IF	ID
1	.35	.49	26	.48	.36
2	.57	.40	27	.32	.54
3	.44	.45	28	.42	.49
4	.47	.52	29	.37	.42
5	.60	.37	30	.45	.34
6	.39	.33	31	.60	.36
7	.49	.54	32	.55	.52
8	.56	.51	33	.30	.60
9	.37	.45	34	.49	.40
10	.24	.41	35	.60	.45
11	.38	.37	36	.48	.49
12	.46	.30	37	.31	.39
13	.39	.45	38	.54	.50
14	.60	.55	39	.41	.41
15	.38	.46	40	.47	.56
16	.38	.40	41	.44	.54
17	.55	.53	42	.50	.43
18	.41	.32	43	.41	.48
19	.36	.53	44	.41	.40
20	.42	.34	45	.60	.57
21	.54	.30	46	.35	.47
22	.34	.50	47	.33	.42
23	.36	.42	48	.31	.45
24	.60	.39	49	.37	.55
25	.41	.34	50	.40	.52

The values of IF and ID ranged from .24 to .60 and from .30 to .60. This range is within the recommended range of difficulty proposed by Oller (1979) and Bensoussan and Ramraz (1984). It is also within the preferred range of discrimination proposed by Hughes (1989). A conclusion which could be drawn here is that the M-C test assessing the writing ability of EFL learners is reliable, reasonable in its level of difficulty, and capable of distinguishing between strong and weak EFL learners.

Besides examining the reliability and suitability of the M-C test, the study had another objective: to ensure the validity of the M-C test. The first type of validity examined in this study was the criterion-reference or concurrent validity, which Weir (1990:27) defines as the extent to which "test scores correlate with another measure of performance, usually an older established test". Hence, it was deemed that the scores of the subjects in the compositions when rated analytically and holistically form an older established measure against which the subjects' scores on the M-C test can be concurrently validated. The Pearson product-moment correlation statistic was run, and the coefficients are given in Table (3) below:

Table (3) The Pearson product-moment correlation coefficients among the M-C test and the subjects' scores in the compositions marked analytically and holistically*

Test/Method	M-C Test	Analytic Method	Holistic Method
M-C Test			
Analytic Method	.82		
Holistic Method	.78	.74	

* All correlations are significant at $p < .05$

The concurrent validity of the M-C test is moderately high, as reflected by the correlation values between the test and the two marking methods. However, the correlation coefficient between the M-C test and the analytic method was higher than the coefficient between the test and the holistic method. In addition, the correlation between the analytic and the holistic methods was the lowest, at .74. Such results seem plausible since both the M-C test and the analytic method had similar marking components; on the contrary, in the holistic method, the scoring is conducted impressionistically and it is left to the rater to estimate the subjects' scores. Hence, the first null hypothesis of the study is rejected and it might be concluded that there are statistically significant correlations among the subjects' scores on the M-C test and their scores in the analytically and holistically marked composition.

The second type of validity investigated is the content validity. Hughes (1989:22) believes that "a test is said to have content validity if its content constitutes a representative sample of the language skills, structure, etc. with which it is meant to be concerned".

Hence, for the M-C test to have content validity, its 50 items should reflect a representative sample of the components of a marking grid for writing reported in other studies and for which content validity was investigated. For example, Zughoul and Kambal (1983) propose structure, content, vocabulary, organization, and mechanics to be the most important components which any marking grid should include. Hughes (1989) believes that any marking scale geared to assess EFL learners' writing ability should take into consideration the assessment of grammar, vocabulary, mechanics fluency (style and ease of communication), and form/organization. Gaudiani (1981) suggests grammar/vocabulary, stylistic technique, organization, and content as the most indispensable components of any marking scale developed to assess the writing ability of EFL learners. Taken together and since it is extremely difficult, if not impossible, to assess the content of students' writings by means of an M-C test, the following components were deemed feasible and essential in terms of testing: grammar, vocabulary, mechanics, and unity and organization. The points allocated to each component in the M-C test are in accordance with Zughoul's and Kambal's (1983) suggestions for the marking scale of the writing of EFL intermediate learners. Table (4) below gives the components and the subcomponents along with their counts and percentages with reference to their total, and the total of the test.

Table (4) Counts and percentages of components and subcomponents of the M-C test with reference to their sub-total and the total of the test

	Sub-total		M-C total	
Component/subcomponent	Count	%	Count	%
Grammar			15	30
Tense/aspect	4	26.67		8
Subject-Verb agreement	4	26.67		8
Passive	2	13.33		4
Gerund	1	6.67		2
Infinitive	1	6.67		2
Word Form	3	20.00		6
Vocabulary			12	24
Mechanics			10	20
Spelling	3	30		6
Punctuation	7	70		14
Unity & organization			13	26
Thesis sentence	1	7.69		2
Topic sentence	2	15.39		4
Summary sentence	2	15.39		4
Cohesion (ties/reference)	4	30.77		8
Cohesion (coordinate/conjunctive)	4	30.77		8

The M-C test seems to be balanced in the assessment of the major components of the marking scales of writing. Furthermore, the subcomponents also appear to be representative. The assigned items for each subcomponent are also acceptable when the importance and difficulty of the subcomponent is considered. For example, in the grammar component, 53.34% of the items were devoted to the assessment of the subjects' mastery of English tense and aspect, and the subject-verb agreement rules (26.67% each) - two major topics in English grammar; whereas, only 13.33% of the items were directed to an investigation of the subjects' familiarity with the gerund and the infinifive, two less common topics in English grammar. The ability of the subjects to distinguish among English word forms was assessed by 20% of the grammar component items. Furthermore, 70% of the items of the mechanics component was devoted to the assessment of the subjects' mastery of English punctuation. Thesis, topic, and summary sentences were assigned 38.46% of the unity and organization component whereas the mastery of cohesion was assessed by 51.54% of the component items.

A test has a face validity "if it looks as if it measures what it is supposed to measure" (Hughes, 1989:27). To examine the face validity of the M-C test, two methods were used. First, the subjects were surveyed in order to find out whether they believed that the test is one geared to assess their writing ability. Second, and following Wall, et al., (1994), face validity could also be derived from the subjects' own rating of their performance on a test. As can be seen in appendices (3) (the English version) and (4) (the Arabic version) the questionnaire consists of six questions developed in accordance with the Likert scale and two multiple-choice questions, each with four alternatives. Table (5) displays the means and the standard deviations of the six questionnaire questions which aim at surveying the test's (1) difficulty, (2) fairness, (3) representativeness, (4) completeness, (5) appropriateness and (6) adequacy, along with the mean and standard deviation of the average.

Table (5) The subjects' means and standard deviations on the six questions of the modified questionnaire

Question	(1)		(2)		(3)		(4)		(5)		(6)		Average	
	x	SD	x	SD	x	SD	x	SD	x	SD	x	SD	x	SD
	2.3	1.3	2.6	1.1	2.8	1.2	2.9	1.9	2.9	0.9	2.3	1.3	2.9	1.1

The means of the subjects' responses were toward the "disagree" and "do not know" end of the scale. They ranged from 2.37, the mean of the adequacy question, to 2.90, the mean of the appropriateness question. In general, the subjects' views on the test are moderately negative. However, such an attitude is expected when the subjects are unfamiliar with a test format (Wall, et al., 1994). The subjects whose answers to the questions on fairness and adequacy were "strongly disagree", "disagree", or "do not know" were asked to answer questions (3) and (8) on the questionnaire in order to elicit their points of view. Questions (3) and (8) are two multiple-choice questions, each with four alternatives. Table (6) shows the count and percentages of the subjects' responses to these two questions.

Table (6) The subjects' means and standard deviations on the six questions of the modified questionnaire

Question	Fairness					Adequacy				
Alternative	(1)	(2)	(3)	(4)	Total	(1)	(2)	(3)	(4)	Total
Count	10	4	4	20	38	8	17	13	2	40
40 %	26.32	10.53	10.53	52.63	73.08	20.00	42.50	32.50	5.00	76.92

52.53% of those who claimed that the test was unfair ascribed such claims to their unfamiliarity with the test's format. 26.32% of the subjects who gave negative responses to question (3) in the questionnaire reported that such a test should only be employed to screen prospective university students. Furthermore, 42.50% of the subjects who criticized its adequacy believed that the test assesses reading comprehension and grammar whereas 32.50% of them contended that the test assesses reading comprehension, grammar, and writing ability. It seems that training EFL learners to deal with various testing techniques and formats is advisable. When such training is conducted, the allegation of the test's unfamiliar format will not stand up anymore.

The subjects were also asked to rate their performance on the M-C test and the composition. They were given a six-point scale ranging from (5) [excellent] to [very poor]. Then Spearman rank-order correlation (Spearman's rho) was calculated as indicated in Table (7):

Table (7) Spearman's rho correlations between the subjects' self-assessment and their scores on the M-C test and the composition

Test/Method	r
M-C	.56**
Analytic	.38*
Holistic	.41*

* $p < .05$

** $p < .01$

Although the reported correlations are low, they are similar to those reported in conventional self-assessment studies. Criper and Davies (1988) reported a correlation of .40 in their evaluation of a proficiency test. Fok (1981) found a correlation equal to .30 between subjects' self-assessment and their scores on a placement test. Wall, et al, (1994) reported correlations ranging from .30 to .51. The explanations which these studies gave to account for such low correlations seem plausible. They claim that the unfamiliarity of the subjects with the self-assessment technique plays the key role in such results. EFL learners are not usually trained to assess their performance and, consequently, they do not, in fact, know where they stand on the EFL proficiency continuum. As a conclusion, although the M-C test received a low rating on the basis of its face validity, 48.08% of the sample (those who chose the "agree" and "strongly agree" responses in question (7)), in addition to the subjects who selected alternative (3) of question (8) believe that the test is geared to an assessment of their writing ability. The low rating of the rest of the sample should not be regarded as a drawback of the test. The novelty of the M-C test's format and the unfamiliarity of the subjects with self-assessment of their performance on tests seem to be the major factors causing such a low rating. In conclusion, the second null hypothesis of the study is thus rejected: A conclusion that there are statistically significant correlations between the subjects' self-assessment and their performance on the analytically and holistically marked composition could be formalized.

Predictive validity is the "indication of how well a test predicts intended performance. A university admission test is said to have predictive validity if its scores correlate highly with a performance criterion such as university grades" (Henning, 1987:196). Hence, the predictive validity is deemed to be the correlation between the subjects' scores on the M-C test which was administered when they

were at the intermediate level, and their scores on the writing test, marked analytically, when they were at the advanced level. Table (8) presents the Pearson product-moment correlation coefficients among the M-C test, the scores of the subjects on the compositions at the intermediate level, which were analytically and holistically marked; and their scores on the compositions at the advanced level which were marked analytically.

Table (8) Pearson product-moment correlations between the scores of the compositions at the advanced level and the subjects' scores on the M-C test and their scores on the compositions at the intermediate level

Test/Method	r*
M-C	.61
Analytic	.62
Holistic	.55

* All correlations are significant at $p < .001$

The predictive validity of the M-C test is high. This implies that the M-C test can predict the performance of students in writing, at the next proficiency level. The third null hypothesis tested in this study is thus rejected and it might be concluded that there is a statistically significant correlation between the subjects' scores on the M-C test and their scores in the composition written by them at next proficiency level.

Finally, the construct validity of the M-C test was investigated. A test is said to have construct validity if "it can be demonstrated that it measures just the ability which it is measure" (Hughes, 1989:26). Henning (1987:99) recommends a technique called "Internal Construct Validation" to assess construct validity. In this technique, the correlations among individual items in the test and the various component totals are calculated. If the value of the correlation between an item and the total of its components exceeds the values of the correlations between the item and the other component totals, the item is said to have construct validity. Furthermore, Henning insists that the correlations obtained from The "Internal Construct Validation" should be corrected for "part-whole overlap" since part of the amount of the total is due to the presence of that item. Table (9) gives the correlation coefficients among the items of the M-C test and the totals of the four components, after correction for part-whole overlap.

Table (9) Pearson product-moment correlations among the M-C test's items and the totals of the four components of the M-C test

Item No	Grammar total	Vocabulary total	Mechanics total	Unity/Orgn.total
Grammar				
1	.54	.49	.40	.20
2	.70	.13	.52	.44
4	.60	.36	.38	.43
6	.57	.30	.28	.43
8	.46	.37	.29	.60
9	.66	.17	.49	.47
11	.82	.64	.32	.14
13	.64	.47	.17	.29
17	.45	.17	.33	.43
18	.61	.54	.47	.38
22	.53	.30	.42	.13
25	.42	.23	.39	.28
28	.63	.46	.16	.44
30	.60	.06	.25	-.01
39	.68	.36	.17	.31
Vocabulary				
3	.38	.85	.06	.52
15	.47	.61	.54	.24
19	.20	.48	.36	.38
23	.48	.67	.20	.31
24	.33	.44	.25	.41
27	.31	.66	.41	.50
33	.39	.42	.26	.14
36	.38	.73	.15	.26
37	.08	.19	.33	.30
43	.17	.61	.42	.18
44	.23	.67	.28	-.15
45	.21	.59	.32	.19
Mechanics				
5	.61	.29	.84	.26
10	.26	.17	.64	.52
14	.39	.21	.78	.35
16	.28	.25	.35	.32
20	.34	.14	.51	.29
29	.40	.37	.71	.35
31	.39	.21	.61	.29
35	.36	.03	.52	.35
40	.18	.13	.29	.21
42	.31	.21	.42	.07
Unity & Orgn.				
7	.11	.31	.17	.39
12	.25	.37	.43	.55
21	.24	.19	.12	.55
26	.32	.37	.24	.49
32	.13	.52	.30	.63
34	.24	.25	.53	.67
38	.17	.30	.37	.70
41	.23	.33	.12	.59
46	.38	.44	.23	.54
47	.23	.35	.20	.80
48	.28	.11	.38	.71
49	.46	.15	.09	.51
50	.34	.23	.17	.43

The correlations between the grammar items and the total of the grammar components are higher than the correlations among these items and the other three component totals, with the exception of item no. (8). By the same token, the correlations between the items of the vocabulary component and their total are higher than the correlations among these items and the totals of grammar, mechanics, and unity and organization, except for the correlation between item (37) and the vocabulary total. Furthermore, the values of correlations between the mechanics items and their total exceed the values of the correlations among the mechanics items and the totals of the other three components. The same observation can also be extended to the correlations of the unity and organization items.

In addition, two other procedures were employed to ensure the construct validity of the M-C test. First, the correlations among the totals of the M-C test's components (such as grammar, vocabulary etc.) and the components' totals of the analytically marked compositions of the intermediate level were also calculated. The correlations are given in Table (10).

Table (10) Pearson product-moment correlations among the totals of the M-C test's components and the components' totals of the analytically marked compositions of the intermediate level

	Analytically marked composition components			
Test Components	Grammar	Vocabulary	Mechanics	Unity & Organization
Grammar	.71	.26	.46	.39
Vocabulary	.21	.68	.33	.42
Mechanics	.54	.33	.73	.54
Unity&Organization	.41	.21	.43	.63

It is obvious that the correlations of the subjects' scores on the M-C test's components with their scores on similar components of the composition marked analytically are higher than the correlations among the dissimilar components of the M-C Test and the compositon. For example, the correlation between the grammar component of the M-C test and the grammar component of the composition was .71 whereas the correlations between the M-C grammar component and the other three components of the composition, namely, vocabulary, mechanics, and unity and organization, were .26, .46, and .39,

respectively. Moreover, the vocabulary component of the M-C test correlated relatively high with the vocabulary component of the composition, .68; but the correlations between it and the other components of the composition were .26, .23, and .21. The correlations among the other components of the M-C test and the composition components underline the construct validity of the test. The fourth null hypothesis of this study is thus rejected and a conclusion that there are statistically significant correlations among the components of the M-C test and the analytically marked composition could be formalized.

Finally, the three means of the subjects on the M-C test and the analytically and holistically marked composition were compared. It is noteworthy here that the constructs assessed by the M-C test and the composition marked analytically are the same. Table (11) represents the summary table of the Analysis of Variance (ANOVA) of the three means.

Table (11) Analysis of variance for the difference between the subjects' means on the M-C test and the analytically and holistically marked composition

Source of Variance	DF	SS	MS	F
Between subjects	2	2192.65	1096.33	6.43*
Within subjects	153	26076.91	170.44	
Total	155	28269.56		

* $p < .0001$

With a significant F, Scheffe's post hoc comparisons were conducted to detect the significant differences. Table (12) gives Scheffe's F-test values and the differences between the means.

Table (12) Values of Scheffe F-test for the difference between the subjects' means on the M-C test and the analytically and holistically marked composition

Comparisons	Means diff.	Scheffe F-test
M-C test vs. Analytic method	-1.500	0.752
Holistic method vs. Analytic method	-8.596	24.697**
M-C test vs. Holistic method	-7.096	16.830*

* $p < .001$

** $p < .0001$

There was no statistically significant difference between the subjects' scores on the M-C test and their scores in the composition analytically marked. However, there were significant differences between the subjects' scores on the M-C test and their scores in the composition holistically marked, and between their scores in the composition when analytically and holistically marked. The results as revealed in the above table confirm the construct validity of the M-C test. It seems valid here to conclude that the M-C test has high construct validity indices. The part of the fifth null hypothesis concerning the statistically significant difference between the M-C test and the holistically marked composition is thus rejected whereas the other part concerning the difference between the M-C test and the analytically marked compositions is retained. It might be concluded that there is a statistically significant difference between the subjects' means on the M-C test and the holistically marked composition. However, there is no statistically significant difference between the subjects' means on the M-C test and the analytically marked composition.

F. Summary and Conclusion

The study aimed at developing and validating an M-C test which could be utilized in assessing the writing proficiency of EFL learners. The test was developed according to the cloze procedure with the rational deletion method, where only lexical items of interest are deleted and substituted by four alternatives, one being the original lexical item of the text. The task of 52 intermediate EFL learners (the sample of the study) was to choose the alternative that best fitted the context. The appropriateness of the text difficulty level from which the M-C test was developed was ensured by using Flesch's readability index. In order to compare the M-C test with the traditional method of assessing the writing ability of EFL learners, the subjects were given a prompt to which they had to respond in a 25-line composition. The compositions were then marked analytically and holistically. In the holistic method of marking, four raters were asked to assign one score to each composition depending on their impression of the subjects' writing following the analytic approach in marking. The four raters were given a marking grid in which different scores were assigned to various components (grammar, vocabulary, mechanics, and unity and organization) and asked to assign a separate score to each component. The score assigned to each composition was the sum of the subjects' scores in the various components.

The findings of this study revealed that the M-C writing test enjoys a high reliability index which is higher than that of the subjects' composition scores when holistically marked and almost equal to the reliability index of subjects'

composition scores if analytically rated. In addition, the M-C test has a high concurrent or criterion-reference validity. It correlated significantly with the analytically and holistically marked composition. The content of the test was similar to the marking grid proposed by some studies reported in the literature. It was representative of the most important elements of English grammar such as tense and aspect, subject-verb agreement rules, and word form. Furthermore, the test contained elements geared to assess EFL learners' mastery of mechanics and familiarity with the acceptable organization of paragraphs in English. The test was capable of predicting EFL learners' future performance in writing tasks, and it also has a high degree of construct validity. This indicates that the test was capable of assessing the same or similar constructs assessed by the traditional method which requires learners to write compositions. However, the test's rating in its face validity was low. This could be ascribed to the unfamiliarity of the learners with the testing format and their inability to deal properly with self-assessment techniques. Training EFL learners to deal with various testing techniques and improving their ability to assess their performance on tests might solve this problem.

The M-C writing test seems to be a valuable tool in assessing EFL learners ability in writing. It does not require the time-consuming efforts usually devoted to the manual marking of learners' compositions, since learners' responses could be marked quickly and efficiently by machine. Moreover, the subjectivity which is usually a major drawback of composition marking is absent in the M-C test. Correct responses will continue to be marked correct no matter who marks learners' responses. The M-C test also contains an element for the assessment of learner's proficiency in spelling and familiarity with English vocabulary. However, the M-C test is not able to assess learners' hand-writing legibility or their competence in developing their own ideas. But these considerations should not discourage instructors and language testing specialists from considering the M-C writing test as a reliable and valid tool which could be employed for proficiency assessment, diagnostic, and/or screening purposes in writing, especially if the number of examinees is large and the time available to raters is short. Since there is no perfect testing procedure, shortcomings are to be expected. These shortcomings, however, are not due to a deficiency in the M-C test but rather due to the use of a passive test (for example, the M-C test) to assess an active skill (i.e., writing.) Finally, the use of the M-C writing test should not be overgeneralized. It should be integrated into proficiency and placement tests where other skills are assessed.

References

Al-Fallay, I. 1997. Investigating the Reliability and Validity of the Fixed Ratio Multiple-Choice Cloze Test. **Dirasat**,

Arndt, V. 1987. Six Writers in Search of Texts: A Protocol Based Study of L1 and L2 Writing, **English Language Teaching Journal**, 41, 257-67.

Bensoussam, M. and Ramraz, R. 1984. Testing EFL Reading Comprehension Using a Multiple-Choice Rational Cloze. **The Modern Language Journal**, 68, 230-39.

Benton, S. and Blohm, P. 1986. Effect of Question Type and Position on Measures of Conceptual Elaboration in Writing. **Research in the Teaching of English**, 20, 98-108.

Bereiter, C. and Scardamalia, M. 1983. Does learning to write have to be difficult: In: A. Freedman, I. Pringle and J. Yalden (Eds.), **Learning to Write: First Language/Second Language** (pp. 20-33). London: Longman.

Boyle, O. and Peregoy, S. 1990. Literacy Scaffolds: Strategies for First- and Second- Language Readers and Writers. **Reading Teacher**, 44, 194-200.

Brossell, G. 1986. Current Research and Unanswered Questions in Writing Assessment. In K. Greenberg, S. Harvey, S. Weiner and R. Donovan (Eds.), **Writing Assessment: Issues and Strategies** (pp. 168-82). New York: Longman.

Brown, J. 1991. Do English and ESL Faculties Rate Writing Samples Differently. **TESOL Quarterly**, 25, 587-603.

Brown, J. and Bailey, K. 1985. A Categorical Instrument for Scoring Second Language Written Skills. **Language Learning**, 34, 21-42.

Canale, M., Frenette, N., and Belanger, M. 1988. Evaluation of a Minority Student Writing in First and Second Language. In J. Fine (Ed.), Second Language **Discourse: A Textbook of Current Research**. Norwood, NJ: Ablex.

Carlson, S., Bridgeman, B. and Waanders, J. 1985. The Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English. **TOEFL Research Report 19**. Princeton, NJ: Educational Testing Service.

Carson, J., Carrell, P., Silberstein, S., Kroll, B., and Kuehn, P. 1990. Reading-Writing Relationships in First and Second Language. **TESOL Quarterly**, 24, 245-66.

Casanave, C. 1995. Local Interactions: Constructing Context for Composing in a Graduate Sociology Program. In D. Belcher and G. Braine (Eds.), **Academic Writing in a Second Language: Essays on Research and Pedagogy** (pp. 83-100). Norwood, NJ: Ablex.

Charles, M. 1990. Responding to Problems in Written English Using a Student Self-monitoring Technique. **ELT Journal**, 44, 286-93.

Clarke, M. 1978. Reading in Spanish and English: Evidence from Adult ESL Students. **Language Learning**, 29, 121-50.

Cleary, C. 1988. Testing Lower Intermediate Writing: A Comparison of Two Scoring Methods. **British Journal of Language Teaching**, 26, 75-80.

Criper, C. and Davies, A. 1988. **ELTS Validation Project Report**. London: British Council and University of Cambridge Local Examinations Syndicate.

Cummins, J. 1981. The Role of Primary Language Development in Promoting Educational Success for Language Minority Students. In **Schooling and Language Minority Students: A Theoretical Framework** (pp. 3-49). Los Angeles, CA: Evaluation, Dissemination and Assessment Center of California State University.

Edelsky, C. 1982. Writing in Bilingual Program: The Relation of L1 and L2 Texts. **TESOL Quarterly**, 16, 211-28.

Educational Testing Service. 1992. **TOEFL: Test and Score Manual**. Princeton, NJ: Educational Testing Service.

Eisterhold, J. 1990. Reading-Writing Connections: Toward a Description for Second Language Learners. In B. Kroll (Ed.), **Second Language Writing: Research Insights for the Classroom** (pp. 88-102). Cambridge: Cambridge University Press.

Flahive, D. and Bailey, N. 1993. Exploring Reading/Writing Relationships in Adult Second Language Learners. In I. Leki and J. Carson (Eds.), **Reading in the Composition Classroom: Second Language Perspectives** (pp. 128-140). Boston, MA: Heinle and Heinle.

Flesch, R. 1948. A New Readability Yardstick. **Journal of Applied Psychology**, 32, 321-33.

Fok, C. 1981. **Reliability of Self-assessment**. Paper presented at BAAL Seminar of Language Testing. Reading, England.

Freedman, S. 1983. Student Characteristics and Essay Test Writing Performance. **Research in the Teaching of English**, 7, 313-25.

Gaudiani, C. 1981. **Teaching Composition in the Foreign Language Curriculum**. Language in Education: Theory and Practice Series, No. 43. Washington, DC: Center for Applied Linguistics.

Hadley, A. 1993. **Teaching Language in Context**. Boston, MA: Heinle & Heinle.

Halliday, M. and Hasan, R. 1976. **Cohesion in English**. London: Longman.

Hamp-Lyons, L. 1990. Second Language Writing: Assessment Issues. In K. Barbara (Ed.), **Second Language Writing: Issues and Options**. New York: Macmillan.

Hamp-Lyons, L. and Prochnow, S. 1991. The Difficulties of Difficulty: Prompts in Writing Assessment. In S. Anivan (Ed.), **Current Developments in Language Testing** (pp. 58-76). Singapore: SEAMEO Regional Language Center.

Harris, D. 1969. **Testing English as a Second Language**. New York: McGraw-Hill.

Hayes, J. and Flower, L. 1983. Uncovering Cognitive Processes in Writing: An Introduction to Protocol Analysis. In P. Mosenthal, S. Tamor and S. Walmsley (Eds.), **Research on Writing: Principles and Methods** (pp. 207-19). Harlow: Longman.

Heaton, J. 1988. **Writing English Language Tests**. London: Longman.

Henning, G. 1987. **A Guide to Language Testing: Development, Evaluation, Research.** Boston, MA: Heinle & Heinle.

Hoetker, J. and Brossell, G. 1986. A Procedure for Writing Content-Fair Essay Examination Topics for Large Scale Writing Assignment. **College Composition and Communication**, 37, 328-35.

Hoetker, J. and Brossell, G. 1989. The Effect of Systematic Variation in Essay Topics on the Writing Performance of College Freshmen. **College Composition and Communication**, 40, 414-21.

Hughes, A. 1989. **Testing for Language Teachers**. Cambridge: Cambridge University Press.

Janopoulus, M. 1989. The Relationship of Pleasure Reading and Second Language Writing Proficiency. **TESOL Quarterly**, 20, 764-68.

Jensen, G. and Ditiberio J. 1989. **Personality and the Teaching of Composition.** Norwood, NJ: Ablex.

Kassen, M. 1990. Responding to Foreing Language Student Wrting: A Case Study of Twelve Teachers of Beginning, Intermediate, and Advanced Level French. Unpublished Ph. D. Dissertation. TX: The University of Texas.

Lado, R. 1986. Analysis of Native Speaker Performance on a Cloze Test. **Language Testing**, 3, 130-46.

Leki, I. 1995. Good Writing: I Know It When I See It. In D. Belcher and G. Braine (Eds.), **Academic Writhing in a Second Language: Essays on Research and Pedagogy** (pp. 23-46). Norwood, NJ: Ablex.

McDonough, S. 1995. **Strategy and Skill in Learning a Foreign Language**. London: Edward Arnold.

McLaughlin, B. 1987. Reading in a Second Language: Studies with Adult and Child Learners. In S. Goldman and H. Trueba (Eds.), **Becoming Literate in English as a Second Language** (pp. 57-70). Norwood, NJ: Ablex.

Oller, J. 1979. **Language Tests at School**. London: Longman.

Oller, J. and Conrad, C. 1971. The Cloze Technique and ESL Proficiency. **Language Learning**, 21, 183-94.

Park, Y. 1988. Academic and Ethnic Background as Factors Affecting Writing Performance. In A. Purves (Ed.), **Writing Across Languages and Cultures: Issues in Cross Cultural Rhetoric** (pp. 38-58). Newbury Park, CA: Sage Publications.

Perkins, K. 1983. On the Use of Compositon Scoring Techniques, Objective Measures, and Objective Tests to Evaluate ESL Writing Ability. **TESOL Quarterly, 17, 651-71.**

Perl, S. 1980. Understanding Composition. **College Composition and Communication**, 31, 363-9.

Perl, S. 1981. **Coding the Composing Process: A Guide for Teachers and Researchers**. Washington, DC: National Institute of Education. (ED 240609).

Pollitt, A. and Hutchinson, C. 1987. Calibrating Graded Assessment: Rasch Partial Credit Analysis of Performance in Writing. **Language Testing**, 4, 72-92.

Prior, P. 1995. Redefining the Task: An Ethnographic Examination of Writing and Response in Graduate Seminars. In D. Belcher and G. Braine (Eds.), **Academic Writing in a Second Language: Essays on Research and Pedagogy** (pp. 47-82). Norwood, NJ: Ablex.

Raimes, A. 1985. What Unskilled ESL Students Do as They Write: A Classroom Study of Composing. **TESOL Quarterly**, 19, 229-58.

Reid, J. 1990. Responding to Difference Topic Types: A Quantitative Analysis. In B. Kroll (Ed.), **Second Language Writing Assessment: Issues and Options** (pp. 154-170). Norwood, NJ: Ablex.

Ruth, L. and Murphy, S. 1988. **Designing Writing Tasks for the Assessment of Writing**. Norwood, NJ: Ablex.

Spaan, M. 1989. **Essay Tests: What's in a Prompt?** A paper presented at the 1989 TESOL Convention, San Antonio, TX.

Uzawa, K. and Cumming, A. 1989. Writing Strategies in Japanese as a Foreign Language: Lowering or Keeping up the Standards? **Canadian Modern Language Review**, 46, 178-194.

Valdés, G., Haro, P., and Echevarriarza, M. 1992. The Development of Writing Abilities in a Foreign Language: Contributions Toward a General Theory of L2 Writing. **Modern Language Journal**, 76, 333-52.

Wall, D., Clapham, C., and Alderson, J. 1994. Evaluating a Placement Test. **Language Testing**, 1, 322-44.
Weigle, S. 1994, Effects of Training on Raters of ESL Compositions. **Language Testing**, 11, 197-223.
Weir, C. 1990. **Communicative Langeage Testing**. New York: Prentice Hall.
Winer, J. 1971. **Statistical Principles in Experimental Design**. New York: McGraw-Hill.
Zughoul, M. and Kambal, O. 1983. Objective Evaluation of EFL Composition. **IRAL**, 11, 87-103.

Appendix (1)

The M-c Coze test used to assess EFL learners' writing ability

Part 1

Directions: From the four given alternatives, choose the one which you think best fits the context by copying the letter you choose in the included answering sheet.

Example: John [(a) is (b) am (c) are (d) has] a good student.

(a) is the answer, so in the answering sheet the letter (a) of the corresponding question number should be circled.

Auctions are public sales of goods, conducted by an officially approved auctioneer. He **[1. (a) asked (b) was asking (c) asks (d) ask]** the crowd assembled in the auction-room **[2. (a) to made (b) to make (c) to makes (d) to should made]** offers, or **[3. (a) bids (b) beds (c) bads (d) buds]**, for the **[4. (a) vary (b) various (c) variety (d) varies]** items on sale. He encourages buyers to bid higher figures **[5. (a) . (b) ; (c) , (d) :]** and **[6. (a) finally (b) final (c) finalize (d) finalizing]** names the highest bidder **[7. (a) when (b) as (c) then (d) after]** the buyer of the goods. This is **[8. (a) calls (b) calling (c) call (d) called]** 'knocking down' the goods, for the **[9. (a) bidding (b) bid (c) bids (d) to bid]** ends when the auctioneer bangs a small hammer on a table at which he stands. This is often set on a raised platform called a **[10. (a) rostrum (b) rostrua (c) rostrui (d) rostrue]**.

The ancient Romans probably **[11. (a) has invented (b) have invented (c) invents (d) invented]** sales by auction, **[12. (a) or (b) and (c) either (d) after]** the English word **[13. (a) come (b) comes (c) have come (d) have came]** from the Latin auction **[14. (a) , (b) ; (c) - (d) .]** meaning 'increase'. The Romans usually sold in this way the **[15. (a) spell (b) hard (c) spoils (d) instruments]** taken in war **[16. (a) , (b) ; (c) - (d) :]** these sales **[17. (a) were called (b) has called (c) had called (d) were call]** sub hasta, meaning 'under the spear', a spear **[18. (a) be stuck (b) been stucking (c) be stucking (d) being stuck]** in the ground as a signal for the crowd **[19. (a) to collect (b) together (c) to pick (d) to gather]**. In England in the eighteenth and nineteenth centuries goods were often sold 'by the candle' **[20. (a) : (b) , (c) ; (d) -]** a short candle was lit by the auctioneer, and bids could be made while **[21. (a) it (b) him (c) its (d) them]** stayed alight.

Practically all goods whose qualities vary are sold by auction. Among these are coffee, hides, skins, wool, tea, cocoa, furs, spices, fruit and vegetables. Auction **[22. (a) sold (b) sales (c) selling (d) sell]** are also **[23. (a) usual (b) seldom (c) mostly (d) largely]** for land and property, antique furniture, pictures, **[24. (a) rare (b) recent (c) good (d) inexpensive]** books,

old china and similar works of art. The auction-rooms at Christie's and Sotheby's in London and New York **[25. (a) are (b) is (c) have (d) has]** world-famous.

An auction is usually advertised beforehand with full particulars of the articles to be sold **[26. (a) while (b) furthermore (c) therefore (d) and]** where and when they can be viewed by prospective buyers. If the **[27. (a) collection (b) approval (c) articles (d) advertisement]** cannot give full details, catalogues **[28. (a) has be (b) are (c) has been (d) will has been]** printed, and each group of goods to be sold together, called a 'lot' **[29. (a) ; (b) , (c) : (d) -]** is usually given a number. The auctioneer need not begin with Lot 1 and **[30. (a) continues (b) continue (c) continued (d) have continued]** in numerical order **[31. (a) ; (b) : (c) , (d) -]** he may wait until he registers the fact that certain dealers are in the room and then produce the lots **[32. (a) them (b) they (c) their (d) there]** are likely to be interested in. The auctioneer's services are paid for in the form of a **[33. (a) percentage (b) rising (c) exceeding (d) raising]** of the price the goods are sold for. The auctioneer **[34. (a) therefore (b) but (c) furthermore (d) before]** has a direct interest in pushing up the bidding as high as possible.

The auctioneer must know **[35. (a) fairly (b) farely (c) freely (d) fearly]** accurately the **[36. (a) old (b) temporary (c) current (d) remote]** market values of the goods he is selling, and he should be acquainted with regular buyers of such goods. He will not waste time by starting the bidding too low. He will also play on the **[37. (a) rivalries (b) fight (c) quarrel (d) belief]** among **[38. (a) his (b) their (c) its (d) her]** buyers and succeed in getting a high price by encouraging two business competitors to bid against each other. It is largely on his advice that a seller will fix a 'reserve' price, that is, a price below which the goods cannot be sold. Even the best auctioneers, however, **[39. (a) finds (b) find (c) have find (d) finding]** it difficult to stop a 'knock-out', whereby dealers **[40. (a) ielegally (b) illegally (c) unlegally (d) ilegally]** arrange beforehand not to bid against each other, **[41. (a) or (b) but (c) therefore (d) and]** nominate one of **[42. (a) themself (b) themselves (c) theirselves (d) themsleves]** as the only bidder in the hope of buying goods at **[43. (a) extremely (b) expectedly (c) perfectly (d) interestingly]** low price. If such a 'knock-out' comes off, the **[44. (a) real (b) late (c) outstanding (d) suspicious]** auction sale takes place **[45. (a) privately (b) particularly (c) precisely (d) erroneously]** afterwards among the dealers.

Part II

Directions: Now after reading the text, answer the following questions by choosing the best alternative which you think best answers the question by copying the letter you choose in the included answering sheet.

46. The best topic sentence for paragraph (1) is
    (a) In auctions, articles are sold by an auctioneer to persons making the highest offers.
    (b) Auctions are held everywhere.
    (c) Auctions are rarely held on Sundays.
    (d) There are some differences between auctions held on Saturdays and those conducted on Fridays.

47. The best topic sentence for paragraph (3) is
    (a) The articles of any auction are usually kept in storage before the time of the auction.
    (b) Buyers are usually aware of the articles, the place, and time of an auction before it is held.

(c) People who participate in an auction have to pay in cash for whatever they buy.

(d) Some people attend auctions as spectators.

48. The best summary sentence for paragraph (3) is

(a) If advertisement does not help in getting prospective buyers, colored catalogues have to be printed.

(b) An auctioneer has to advertise the time and place of the auction.

(c) Since buyers are usually interested in some goods rather than others, the auctioneer does not have to continue in numerical order

(d) In order to get more money, the auctioneer tries to push the price of an article as high as possible.

49. The best summary sentence for paragraph (1) is

(a) An auctioneer is the one who runs an auction.

(b) People have to come early to the auction sale to buy what they need.

(c) Auctions are usually a good chance for merchants to meet each other.

(d) The auctioneer has to stand on a rostrum.

50. The whole text could be an answer to

(a) What is the role of the auctioneer?

(b) What is an auction sale?

(c) Are auctions an old practice in human life?

(d) What are the qualifications of a good auctioneer?

Appendix (2)

A modified marking scale adopted from Zughoul and Kambal (1983:97) for intermediate EFL learners.

Name _____

1. **Structure**	Grammatical accuracy (S-V agreement, tense, word order, function words etc.), sentence complexity and variety of constructions	
	23-25	Excellent
	18-22	Good
	14-17	Fair
	9-13	Poor
	0-8	Very Poor
2. **Vocabulary**	Appropriate choice of lexical items, range, directness and register	
	18-20	Excellent
	15-17	Good
	12-14	Fair
	7-11	Poor
	0-6	Very Poor
3. **Organization**	Logical theme development, coherence and clear statement of ideas	
	18-20	Excellent
	15-17	Good
	12-14	Fair
	7-11	Poor
	0-6	Very Poor
4. **Mechanics**	Spelling, punctuation, capitalization, paragraphing and hand-writing	
	14-15	Excellent
	12-13	Good
	9-11	Fair
	5-8	Poor
	0-4	Very Poor

Total _____ Marker _____

Appendix (3)

Questionnaire developed by Oller and Conrad (1971), used by Lado (1986), and modified by Al-Fallay (1997) to assess attitude of subjects towards the Cloze testing procedure as a measure of overall language proficiency. This questionnaire is the form used by Al-Fallay (1997): However, it was also modified to assess the attitude of subjects towards the suitability of suing the M-c cloze test to assess their writing ability.

1. The test is balanced in its difficulty level.

strongly disagree	disagree	do not know	agree	strongly agree
1	2	3	4	5

2. The test is fair. It is not related to a specific group background. It does not contain difficult or specialized vocabulary.

strongly disagree	disagree	do not know	agree	strongly agree
1	2	3	4	5

3. If your answer to statement (2) is strongly disagree, disagree, or do not know, select one reason from the following alternatives:

(1) The test is designed to screen prospective university students.

(2) The test contains difficult or specialized vocabulary unknown to most subjects.

(3) Subjects are not familiar with the test's topic which does not consider their culture and environment.

(4) Subjects are not familiar with the test's format.

4. The test is representative. It can be used with the population of new EFL students.

strongly disagree	disagree	do not know	agree	strongly agree
1	2	3	4	5

5. The test is carefully designed to be a complete assessment tool to measure the reading comprehension, writing, knowledge of English grammar.

strongly disagree	disagree	do not know	agree	strongly agree
1	2	3	4	5

6. I suggest that the test be used as part of a placement test for prospective EFL students who will be placed in different levels based on their scores in this test.

strongly disagree	disagree	do not know	agree	strongly agree
1	2	3	4	5

7. The test measures English writing ability of new EFL students.

strongly disagree	disagree	do not know	agree	strongly agree
1	2	3	4	5

8. If your answer to statement (7) is strongly disagree, disagree, or do not know, select one reason from the following alternatives:

   (1) The M-c Cloze test only assesses reading comprehension.

   (2) The M-c Cloze test only assesses reading comprehension and grammar.

   (3) The M-c Cloze test only assesses reading comprehension, grammar, and writing. It does not assess either speaking or listening.

   (4) The M-C cloze test does not assess any of the EFL skills and elements.

❋ ❋ ❋

Appendix (4)

The Arabic version of appendix (3) questionnaire

استبانة

الرجاء وضع دائرة حول الاجابة التي تعتقد أنها مناسبة أو تصف ما تعتقد أنه صحيح:

١ – أعتقد أن هذا الاختبار هو تقييم فعال ومتوازن في مستوى صعوبته، ويحدد بصدق المهارات المستهدفة في اللغة الإنجليزية.

١	٢	٣	٤	٥
غير موافق جداً	غير موافق	لا أعلم	موافق	موافق جداً

٢ – أعتقد أن هذا الاختبار عادل، لا يفرق بين خلفية متعلم وآخر، ولا يحوي مفردات صعبة أو متخصصة، ومصمم لا يكون أداة متوازنة للغة الانجليزية.

١	٢	٣	٤	٥
غير موافق جداً	غير موافق	لا أعلم	موافق	موافق جداً

٣ – إذا كان جوابك لا أعلم أو غير موافق على البند (٢) أعلاه، فما هو السبب؟

(أ) الاختبار مصمم للمتعلمين الذين يريدون دخول الجامعة دون غيرهم.

(ب) يحوي الاختبار مفردات صعبة أو متخصصة لا يعرفها أغلب المتقدمين.

(ج) مواضيع الاختبار غريبة على أغلب المتقدمين، ولا تأخذ بالاعتبار البيئة المحلية.

(د) صياغة الاختبار وبشكله غير معروف لمعظم المتقدمين.

٤ – أعتقد ان هذا الاختبار يمثل بشكل متوازن جمهور المتعلمين المستجدين للغة الانجليزية، ولا يشكل في طريقة كتابته وقواعد النحو ومفرداته تحيزاً لبعضهم على الآخر.

١	٢	٣	٤	٥
غير موافق جداً	غير موافق	لا أعلم	موافق	موافق جداً

٥ – أعتقد أن هذا الاختبار مصمم بعناية لأن يكون تقييماً كاملاً لمهارات القراءة والكتابة والقواعد والمحادثة ولا ينقصه شيء من هذه المهارات.

١	٢	٣	٤	٥
غير موافق جداً	غير موافق	لا أعلم	موافق	موافق جداً

٦ – أقترح أن يستخدم هذا الاختبار لتحديد مستوى المتعلمين المستجدين الراغبين في دخول برامج اللغة الانجليزية بغية وضعهم في المستويات التي تناسب كفائتهم اللغوية في اللغة الانجليزية، وأن يتم قبولي وقبولهم بناء على نتائجه.

١	٢	٣	٤	٥
غير موافق جداً	غير موافق	لا أعلم	موافق	موافق جداً

٧ – أعتقد أن هذا الاختبار يقيس بشكل ملائم المقدرة اللغوية لمهارة الكتابة لمتعلمي اللغة الانجليزية كلغة أجنبية.

١	٢	٣	٤	٥
غير موافق جداً	غير موافق	لا أعلم	موافق	موافق جداً

٨ – إذا كانت اجابتك غير موافق أو لا أعلم على البند (٧) أعلاه، فماذا تعتقد أن الاختبار يقيس.

(أ) القراءة فقط.

(ب) القراءة والقواعد فقط.

(ج) القراءة والقواعد والكتابة فقط، ولا يقيس المحادثة والاستماع.

(د) لا شيء من هذه المهارات.