Error Recognition Tests as a Predictor of EFL Learners' Writing Ability

Adel Dastgoshadeh

English Language Department, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran

E-mail: Englishadel@yahoo.com

Parviz Birjandi

English Department, Science and Research Branch, Islamic Azad University, Tehran, Iran

Kaveh Jalilzadeh

English Department, Science and Research Branch, Islamic Azad University, Tehran, Iran

E-mail: kaveh_j2004@yahoo.com

Abstract

It is not certain whether multiple-choice tests have essentially the same predictive validity for candidates in different academic disciplines, where writing requirements may vary. Still, at all levels of education and ability, there appears to be a close relationship between performance on multiple-choice and essay tests of writing ability. And yet each type of measure contributes unique information to the overall assessment. In this study the relationship between Iranian EFL students' performance on an error recognition test and their writing ability was investigated. Using appropriate statistical tests such as Pearson correlation coefficient formula and Matched t-test, the data collected from the participants who were selected randomly and voluntarily cooperated during the different phases of the study were analyzed. The results of the study showed that there is no statistically significant relationship between test takers' performance on the error recognition test and their writing ability. The finding of the study can be justified on the ground that error recognition tests gauge construct-irrelevant factors which might not be ever-present factors influencing test takers' writing ability.

1. Introduction

Writing process, as commonly conceived, is a highly sophisticated skill combining a number of diverse elements, only some of *which* are strictly linguistic.

The assessment of writing ability has recently received much attention from educators, legislators, and measurement experts, especially because the writing of students in all disciplines and at all educational levels seems, on the whole, less proficient than the writing produced by students five or ten years ago. The GRE Research Committee has expressed interest in the psychometric and practical issues that pertain to the assessment of writing ability. Specifically, recent scholarship and information from established programs are used to investigate the nature and limitations of essay and multiple-choice tests of writing ability. The statistical relationship of performances on these types of tests, the performance of population subgroups on each kind of task, the possible need of different disciplines for different tests of composition skill, and the cost and usefulness of various strategies for evaluating writing ability need to be investigated. The literature indicates that essay tests are often considered more valid than multiple-choice tests as measures of writing ability. Certainly they are favored by English teachers. But although essay tests may sample a wider range of composition skills, the variance in essay test scores can reflect such irrelevant factors as speed and fluency under time pressure or even penmanship. Also, essay test scores are typically far less reliable than multiple-choice test scores. When essay test scores are made more reliable through multiple assessments, or when statistical corrections for unreliability are applied, performance on multiple-choice and essay measures can correlate very highly. The multiple-choice measures, though, tend to predict the performance of minority candidates on essay tests. It is not certain whether multiple-choice tests have essentially the same predictive validity for candidates in different academic disciplines, where writing requirements may vary. Still, at all levels of education and ability, there appears to be a close relationship between performance on multiple-choice and essay tests of writing ability. And yet each type of measure contributes unique information to the overall assessment. The best measures of writing ability have both essay and multiple-choice sections, but this design can be prohibitively expensive.

One type of multiple choice tests is error recognition tests which are frequently used with the purpose of testing learners' witting ability. This testing method appears at least in two forms on multiple-choice tests: stemless error recognition items, that is, the test takers are supposed to read four independent choices of an item and choose the accurate one, this item type is referred to as editing items and stemed error recognition items, that is, the item contains either one inaccurate word or phrase or followed by a no-error option.

1.1 Statement of the Problem

Writing skill is tested through different test methods including error recognition, sentence completion, self-writing, intensive performance, and essay or composition. These different test types appear on both low-skates and high-skates tests. In most language testing programs, error recognition tests are used frequently, especially with elementary learners to evaluate their writing ability. What is interesting to note is that writing is a process of production rather than a matter of recognition or selection. In this study the researchers investigates the relationship between, if any, Iranian EFL students' performance on an error recognition test and their writing ability. Here the word 'writing" is highlighted as it gauges students' success in producing the language.

1.2 Significance of the Study

Writing skill has been tested through the application of different types of tests, that is, objective and subjective ones. One of the commonest types of objective test items which appear on many language tests is error recognition items. Error recognition items, in turn, appear in two common forms: error recognition items with stem and error recognition items without stems which are called editing items. But it would seem obvious that the most direct way of measuring students' writing ability would be to have them write instead of asking them to recognize the erroneous sentence in editing items or the erroneous word or phrase in erroneous items with stem. But as it was mentioned in introduction section, even essay type items have got their own disadvantages. So according to Harris (1994) an ideal practice is undoubtedly to measure writing skill with a combination of the two types of tests, and it is recommended that this procedure be followed whenever conditions permit. Such a combination will probably produce somewhat more valid results than would either of the two types of measures used by itself. The purpose of the present study is investigating the relationship, if any, between Iranian EFL students' performance on error recognition items and their writing ability.

2. Review of literature

Error recognition is a specific item type on which the examinee is required to indicate which of several underlined parts of a sentence is unacceptable for formal written English, or to indicate that the sentence contains no "error".(Harris, 1994). To take an example, he provides the following one:

The position <u>taken</u> in his <u>most recent</u> speeches <u>seem</u> to indicate willingness to compromise.   <u>No error</u>.
                        A   B   C   D

This item type puts the examinee in the position of a reader who must make judgments about the acceptability of a piece of writing and identify any point of weakness. As such it has sometimes been objected to on the grounds that the examinee (1) does not have to prove ability to correct the errors he finds, and (2) may possibly choose the correct answer for the wrong reason. Despite these theoretical objections, however, experience would seem to indicate that error-recognition items function as well as those which require the examinee to select from among several alternative methods of expression. To ensure comprehensive coverage, the test writer would probably do well to combine an error-recognition subtest with a sentence-completion or sentence-correction subtest in his measure of writing ability.

Sentence completion

The examinee is required to select the best way of completing a sentence in terms of grammar, diction, tone, and sense. This item type provides an excellent method of measuring a wide range of problems relating to the effective use of written English. It is probably the most commonly used of the multiple-choice techniques for testing sensitivity to appropriate style in writing.

Sentence correction

The examinee is required to select the best revision of an underlined portion of a sentence. If the sentence is acceptable as it stands, the examinee selects choice A, which is always identical to the underlined portion of the sentence. It would seem obvious that the most direct way of measuring students' writing ability would be to have them write. Yet, as all language teachers are surely aware, there has in the past fifty years been much criticism of the conventional language tests on the part of educational-measurement specialists. And this criticism has, in turn, brought forth a very spirited defense of the essay examination by many teachers and educationists.

Harris (1994) states that those who have championed the essay or composition have generally included the following points in their defense:

1). Composition tests require students to organize their own answers, expressed in their own words. Thus composition tests measure certain writing abilities (e.g., ability to organize, relate, and weigh materials) more effectively than do objective tests.

2). Composition tests motivate students to improve their writing; conversely, if examinations do not require writing, many students will neglect the development of this skill.

3). Composition tests are much easier and quicker to prepare than objective tests, an important advantage to the busy classroom teacher.

The critics of composition testing have usually answered along the following lines:

1). Composition tests are unreliable measures because (1) students perform differently on different topics and on different occasions; and (2) the scoring of compositions is by nature highly subjective.

2). In writing compositions, students can cover up weaknesses by avoiding problems (e.g., the use of certain grammatical patterns and lexical items) they find difficult. Such evasion is impossible with well-prepared objective tests.

3). Composition tests require much more scoring time than objective tests; for this reason, compositions add greatly to the expense and administrative problems of large-scale testing.

It is unfortunate that, in this long-standing debate, many people have adopted fixed attitudes on one side or the other and have failed to keep abreast of, or have ignored, a number of new studies which definitely provide a basis for reconciliation. The current "moderate position in regard to testing writing ability, based on recent findings, may be summarized as follows:

1). Well-constructed objective tests of the language skills have been found to correlate quite highly with general writing ability, as determined by the rating of actual samples of free writing. Thus in situations where the scoring of compositions would be unfeasible (as in some large-scale testing operations), objective tests can be used alone as fairly good predictors of general writing skill.

2). At the same time, it is now clear that there are ways to administer and score composition tests so that they, too, may be used by themselves as reliable instruments. Put briefly, high reliability can be obtained by taking several samples of writing from each student and having each sample read by several trained readers. Thus the classroom teacher who lacks the experience and/or the time to construct objective tests of writing ability, or who feels strongly about the pedagogical value of testing writing through writing, can use compositions with a reasonable degree of confidence.

3). Inasmuch as both objective tests and composition tests have their own special strengths, the ideal practice is undoubtedly to measure writing skill with a combination of the two types of tests, and it is recommended that this procedure be followed whenever conditions permit. Such a combination will probably produce somewhat more valid results than would either of the two types of measures used by itself. (Weigle, 2002).

Various types of objective tests that might be used to measure writing ability have been developed . The kinds of grammatical problems that are tested in objective writing-ability tests differ markedly from the problems included in structure tests for foreign students. In the latter tests we are concerned with measuring control of the basic grammatical patterns of the language. In our structure items, therefore, the contrast is between English and non-English, and we would assume that a native speaker would, except through carelessness, score 100 percent correct. In our measures of writing ability, -on the other hand, we are testing sensitivity to the grammatical patterns appropriate to the written, as contrasted with the spoken, form of the language, and we would suppose that many native speakers would fail to make some of the distinctions. Examples of the kinds of formal grammatical matters that we might include in our tests of writing ability-but not in structure tests-are the following:

> *Subject-verb agreement*
>
> *Structural parallelism*
>
> *Case of pronouns*
>
> *Comparison of adjectives*
>
> *Formation of adverbs*
>
> *Formation of irregular verbs*

If some of these problems seem too reminiscent of "traditional" or "prescriptive" grammar, it must be understood that our goal is to determine how acceptable the foreign student's written English will be to the native speakers of English who will read his compositions, letters of application, business communications, and the like. Whether for good or ill, most of these readers will be applying very conservative criteria. Therefore, if our advanced-level test is to have relevance and validity, it should undoubtedly contain the kinds of formal grammatical points by which the student will subsequently be judged in real-life situations. What is vitally important is that such a test be clearly identified, by its title, directions, and problem contexts, as a measure of writing skill and not be confused with tests of basic structural control.

Weir (1990) speculates that in addition to control of grammatical forms and syntactic patterns, effective written expression depends on the writer's lexical resources. The vocabulary tests will certainly have some relevance in

measuring the foreign student's writing ability. Such tests, however, are usually concerned only with general meanings, and good writing requires considerable precision in the use of lexical items. Specifically, the writer must be fully aware of the "social status" and connotative meanings of the items he uses and must be able to combine them felicitously into phrases and longer units. Thus, for instance, although it is true that the words *companion, comrade, chum,* and *crony* all carry the same general "meaning" and may all be listed as synonyms in the student's pocket bilingual dictionary, he will certainly run into trouble if he attempts to use them interchangeably. Comprehensive tests of writing skill therefore require attention to appropriate style and diction.

Henning (1987) believes that to determine the best way of assessing writing in a language curriculum requires considerations on all facets related to assessment. A lot of research has been done and a lot of literature has been written on specialists' or teachers' views towards writing assessment. Yet, it should be stressed here that an understanding towards students' perceptions will also be able to inform assessment development. Assessment practices may not be as objective as they claim to be because there are a lot of subjective elements. As described by Bachman (1990), test developers and writers make subjective decisions on design of assessment procedures, production of test items and interpretation of testing results; test takers make subjective judgment on strategies and approaches in tests. A way to achieve greater objectivity and fairness in an assessment situation seems to be minimizing the gap between assessors' or teachers' knowledge on theories and aims, and assesses or students' perceptions towards the assessment.

In choosing or designing writing test, the logical place to begin is by considering what we plan to use the test for. In other words, why are we interested in testing writing ability - what is our purpose. Bachman and Palmer (1996) discuss two main purposes for language tests, of which we can consider writing tests to be a subset. The primary purpose is to make inferences about language ability, and the secondary purpose is to make decisions based on those inferences. That is, since we cannot directly observe a person's language ability we use his or her responses to test items as data from which we make inferences about the ability that underlies the test performance. These inferences are then used as data for making a variety of decisions at an individual, classroom, or program level.

For example, let us consider three types of inferences that we can make on the basis of a language test: proficiency, diagnosis, and achievement. Leaving aside for the moment a precise definition of language proficiency, we use inferences about general language proficiency to make decisions such as admission to academic program, placement into different levels of language program, exemption from certain course work, or selection for a particular job. Inference about diagnosis - that is, the strengths and weaknesses of individual students - are used primarily by teachers to tailor their instruction to meet their students' needs. Inferences about achievement - or the degree to which individuals or groups of students have met specific instructional goals - are used to make decisions about grading and promotion on the individual level, and about modification of instruction on the classroom level. Inferences about achievement are also used on a program-wide or even state or national levels to make decisions about curriculum and funding for programs.

It can be can concluded that different testing methods are used to gauge learners' writing ability and these preferences are mostly regardless of the functions of a gives testing method, for example in order to test how proficiently learners can write in the target language, the test takers judge their performance on multiple-choice tests including error recognition items, although writing is a productive skill. In the present study the researchers attempt to investigate the extent to which language learners' performance on error recognition tests correlates significantly with their writing ability.

2.1 Research Question and hypothesis

Q. Is there any relationship between Iranian EFL students' performance on error recognition tests and their writing ability?

NH. There is no relationship between Iranian EFL students' performance on error recognition tests and their writing ability.

3. Methodology

The purpose of this study was to investigate the nature of relationship between Iranian EFL students' performance on error recognition tests and their writing ability. The methodology section describes the participants of the study, the instruments, and the procedure through which different phases of the study were conducted.

3.1 Participants

The total number of the participants in this study was 125; all of them were Iranian students of EFL at B.A. level. The participants were selected from two universities in Sanandaj: Kurdistan University and Islamic Azad University, Sanandaj Branch. The participants included both male and female students who took part voluntarily and cooperatively in different phases of the study. Homogeneity of these participants was determined through the use of a proficiency test. In the first phase of the study, to determine the error recognition test's reliability and validity, 90 out of 128 students were selected as homogeneous sample. Therefore, in the final administration that was conducted to answer the research question, 34 students were selected.

3.2 Instrumentations

The instruments used to gather the data for this study were:

1). A well-known proficiency test, i.e., TOEFL containing 150 items and consisting of four parts was given to determine the homogeneity of the participants. And also it served as the valid test against which the newly developed test was validated.

2). A new error recognition test including 46 items developed by the researcher was given to the participants.

3). Eight writing samples written by the participants in a writing course during a semester were collected and scored.

3.3 Procedure

The first step:

The second section of a TOEFL test, that is, structure and written expressions was selected and given to the participants and then the newly developed error recognition test consisting of 45 items was also given to them for the purpose of validating it. Four items were discarded as they were too easy or difficult. The newly developed test enjoyed acceptable validity index, that is, 0.79. The reliability of the new test was calculated using Spearman.-Brown prophecy Formula and the test had the reliability index of 0.70.

The second step:

Participants had a writing course at university and they were required to write eight writing samples during a semester and all of them were corrected and scored by the instructor who was the researcher of this study. The mean of these eight grades was considered as participants' level of proficiency in writing skill. The error recognition test developed by the researcher was given to the participants. After collecting the data the performance of participants on the error recognition test was compared and correlated with their mean scores in writing ability.

3.4 Design of the study

Although the participants were selected randomly and factors aside from participants' performance on an error recognition test and their writing ability were not controlled, a causal relationship between participants' performance on an error recognition test and their writing ability could not be drawn; instead, the domain of our claims has to be limited and cause-and-effect statements are avoided. In this case, because no treatment was given to the participants, a controlled design called Ex post facto was used.

3.5 Data analysis

To test the hypothesis of the present study certain statistical techniques were implemented. First descriptive statistics for the scores obtained from participants' performances on both error recognition test and writing course were compared. This was followed by calculating the degree of correlation between these two sets of score using Pearson product correlation coefficient.

4. Results and discussions

The data gathered from the participants involved in the study are analyzed through some statistical tests and also the results are interpreted and discussed. The first analysis dealt with determining the descriptive statistics of the scores from the administration of error recognition test and writing ability. Since the two tests were different from each other in terms of the number of items on each, consequently different sets of raw scores were obtained. So, the raw scores were changed into standardized scores and that is why the mean scores are below 1.

The second analysis deals with determining the degree of correlation between test takers' performance on error recognition and writing test. For this purpose, the appropriate statistical test, i.e., Pearson correlation coefficient was implemented. As the results show, the correlation coefficient is -.058 and p-value is .745. Thus, it can be concluded that the correlation coefficient is not significant, that is, there is no correlation between performance of the test takers on the two different tests, that is, writing and error recognition tests. So, the null hypothesis that there is no relationship between Iranian EFL students' performance on error recognition tests and their writing ability is accepted.

The t observed value for the relationship between test takers' performance on the two tests , 0.066, at 33 degree of freedom is not greater than the critical value of t at the .05 level of significance, i.e., 0.94 , therefore it can be concluded that there is no relationship between test takers' performance on error recognition test and writing test. So, the null hypothesis that there is no relationship between Iranian EFL students' performance on error recognition tests and their writing ability can not be rejected.

5. Conclusion

The findings of the present study indicated that error recognition tests including editing tests can not be a reliable measure to be used in assessing test takers' writing skill as there was no relationship between test takers' performance on these tests and their writing ability. The finding can be justified on the ground that performance on an error recognition test is not a matter of production as the main task of the test takers on such tests is to

decipher the information according to the writer's style of thinking and therefore analyze the organizational structure of the items in congruence with it. We can conclude that in an error recognition test the language has been put by some one else and it is not under the control of the reader because the reader's performance is affected by so many construct-relevant and irrelevant factors. But in writing, the language is selected attentively and appropriately by the writer himself and also it is in congruence with the way he likes to follow. In addition, writing is productive skill which means that the writer has to produce a message communicatively considering almost all extralinguistic factors influencing the process of writing.

References

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: OUP.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: OUP.

Backman, L. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Brown, D. H. (2004). *Language assessment principles and classroom practice*. San Francisco: Pearson Education.

Brown, J. D. (1996). *Testing in language program*. UPPER Sadder River, NJ: Prentice-Hall.

Donald E. P. et al. (1999). Validating the revised test of spoken English against a criterion of communicative success. *Language Testing,* 16(4), 399–425. doi:10.1177/026553229901600401, http://dx.doi.org/10.1177/026553229901600401

Farhady, H., Jafarpour, A. & Birjandi, P. (1996). *Testing language skills: From theory to practice*. Tehran: SAMT.

Harris, P. D. (1994). *Testing English as a second language*. NY: McGraw Book Company.

Heaton, J.B. (1988). *Writing English language tests*. London: Longman.

Henning, G. (1987). *A guide to language testing: development, evaluation and research*. New York: Newbury House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kane. A. C. (2009). *Assessing student writing: A study group approach*. Unpublished Doctoral dissertation. Walden University.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Michigan: Blackwell Publishing.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-237. doi:10.1111/j.1745-3984.1984.tb01030.x, http://dx.doi.org/10.1111/j.1745-3984.1984.tb01030.x

Purpura, J.E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C.J. (1990). *Communicative Language Testing*. Hemel Hempstead: Prentice-Hall.

Table 1. Summary of descriptive statistics of the two sets of scores on error recognition and writing tests

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Writing	.0615	34	.95135	.16316
	Err.recog	.0456	34	.98018	.16810

Table 2. The results of Pearson Correlation coefficient

Paired Samples Correlations				
		N	Correlation	Sig.
Pair 1	Writing & Err.recog	34	-.058	.745

Table 3. Summary of paired sample t-test

Paired Samples Test									
		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Writing - Err.recog	.01588	1.40485	.24093	-.47429	.50606	.066	33	.948