3. Ontario Test of English as a Second Language

Bibliography

Allison, D. & Webber, R. (1984). What place for performative tests? **ELT Journal, 38(3),** 199-205.

Aron, H. (1986). The influence of background knowledge on memory for reading passages by native and nonnative readers. **TESOL Quarterly, 20(1),** 136-140.

Bachman, L. F. (1990). **Fundamental considerations in language testing.** Oxford: Oxford University Press.

Bachman, L. F. & Clark, J. L. D. (1987). The measurement of foreign/second language proficiency. **ANNALS, 490,** 20-33.

Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. **Modern Language Journal, 70(4),** 380-390.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. **Language Testing, 12(1),** 1-14.

Carrell, P. L. (1984). Schema theory and ESL reading: classroom implications and applications. **Modern Language Journal, 68(4),** 332-342.

Carrell, P. L. (1987). Content and formal schemata in ESL reading. **TESOL Quarterly, 21(3),** 461-481.

Carrell, P. L. & Eisterhold, J. (1983). Schema theory and ESL reading pedagogy. **TESOL Quarterly, 17(4),** 553-573.

Clapham, C. (1996). **The development of IELTS: a study of the effect of background knowledge on reading comprehension.** Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Farhady, H. (1979). The disjunctive fallacy between discrete-point and integrative tests. **TESOL Quarterly, 13(3),** 347-357.

Farhady, H. (1983a). On the plausibility of the unitary language proficiency factor. In J. W. Oller (Ed.), **Issues in language testing research** (pp. 11-28). Mass.: Newbury House.

Farhady, H. (1998). **Sociopolitical aspects of ethics in language testing.** Paper presented at the Summer Institute, Carleton University, CA.

Fox, J. (Ed.). (1996). **The Carleton Academic English Language (CAEL) Assessment: Test Manual.** Ottawa: Carleton University.

Jennings, M., Fox, J. & Graves, B. (1998). **Validating a topic-based test of language proficiency.** Paper presented at the Language Testing Research Colloquium, Monterey, CA.

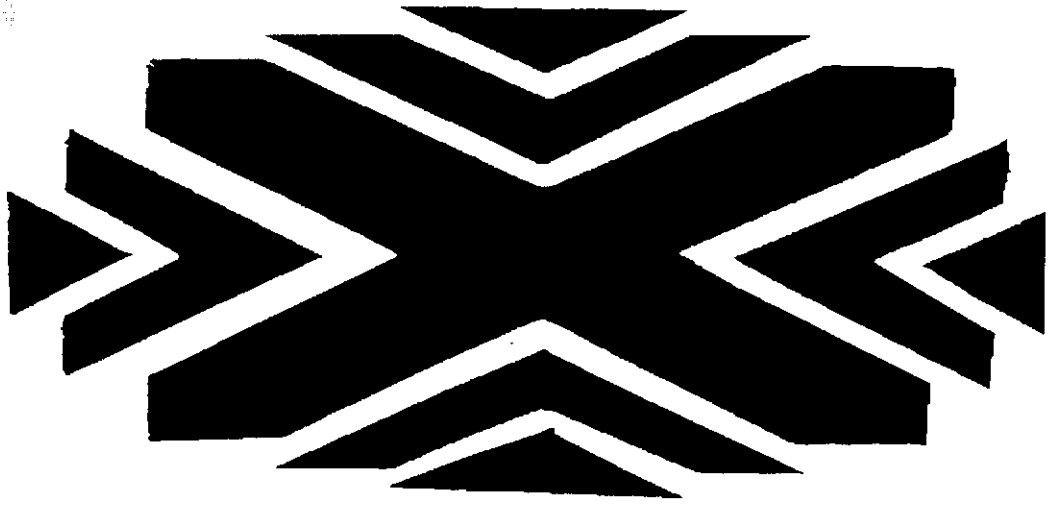
Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a test. **TESOL Quarterly, 15(2).** 169-181.

Johnson, P. (1982). Effects on reading comprehension of building background knowledge. **TESOL Quarterly, 16(4),** 503-516.

Millman, J. & Greene, J. (1993). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), **Educational measurement.** New York: American Council on Education.

Shohamy, E. (1993). The exercise of power and control in the rhetorics of testing. **Carleton Papers in Applied Language Studies, x,** 48-62.

Upshur, J. & Turner, C. (1995). Constructing rating scales for second language tests. **ELT Journal, 49(1),** 3-12.



to Allison and Webber (1984), it is now believed that individuals' performances should be tested directly using realistic tasks to check what they can do with the language. Therefore, being able to get a passing score on a test which does not involve the subjects in conducting realistic activities may not guarantee an individual's future success in an academic program (Fox, et al., 1993).

What might be the strength of the TBLP test is that it takes into account the assumption that a test developer should aim at finding out what language learners can do rather than what they know using authentic materials.

The test might be criticized on the grounds that is has taken advantage of MC format, however. The literature available in this respect shows that in order to score tests as precisely as possible, the best format can be the multiple-choice in which, if the items are constructed properly, only one answer will be correct. Moreover, MC format seems to be highly correlated with other formats.

On the other hand, subjective scoring of tests is a threat to test reliability. In other words, it might not be very easy to score the individuals' performances subjectively even when carefully-prepared rating scales are available. Therefore, making tests in MC form might alleviate the problem of rating leading to more reliable tests in comparison to those subjectively-scored tests.

Still another advantage of the test is its practicality because administering and scoring multiple-choice tests are much easier than those other test formats.

What all of the above-mentioned points indicate is that since the newly-constructed TBLP test seems to possess validity, reliability, and practicality to some extent, it may be considered as a suitable alternative for measuring the language proficiency of EFL students on the basis of which sound decisions can be made.

Notes:

1. Canadian Academic English Assessment
2. International English Language Testing System

Table 4: Correlation Matrix 1

	TBLP(LC)	TBLP(RC)	TBLP(WR)
TBLP(LC)	1.00		
TBLP(RC)	.36	1.00	
TBLP(WR)	.28	.52	1.00
TBLP(total)	.68	.80	.81
TOEFL(total)	.55	.61	.70
TOEFL(LC)	.59	.29	.47
TOEFL(ST)	.37	.47	.54
TOEFL(VOC)	.29	.54	.54
TOEFL(RC)	.20	.49	.42

Table 5: Correlation Matrix 2

	TOEFL(LC)	TOEFL(ST)	TOEFL(VOC)	TOEFL(RC)
TOEFL(LC)	1.00			
TOEFL(ST)	.42	1.00		
TOEFL(VOC)	.27	.30	1.00	
TOEFL(RC)	.21	.26	.48	1.00
TOEFL(total)	.74	.74	.66	.64
TBLP(total)	.59	.60	.60	.48

As the results of these analyses show, it seems that the subtests of the TBLP test correlate, at a reasonable degree, with those of TOEFL. On the other hand, the correlation index of the two tests (TOEFL and TBLP) seems rather high. However, a high correlation between two tests does not necessarily mean that they are measuring the same ability (Farhady, 1979, 1983). Therefore, in order to reveal the nature and the number of the underlying traits which might be measured by these tests, different factor analyses were conducted on subjects' scores. The results are presented in table 6.

Table 6: Varimax Rotated Factor Matrix of Study Measures

Subtest	Factor 1	Factor 2
TOEFL(LC)	**	.86
TOEFL(ST)	.40	.59
TOEFL(VOC)	.78	**
TOEFL(RC)	.79	**
TBLP(LC)	**	.82
TBLP(RC)	.75	**
TBLP(WR)	.67	.43
TOEFL(total)	.68	.66
TBLP(total)	.67	.66

**Loadings less than .30 are deleted.

As the results of the factor analysis show, nearly all the subtests of both tests, except the listening sections, are heavily loaded on factor 1. This may mean that whatever is tested by the TOEFL can also be tested by TBLP test. On the other hand, the listening sections of both tests are heavily loaded on factor 2. This can also mean that listening ability can be measured by both TOEFL and TBLP test.

Conclusion

The results of the study seem to support the hypothesis formulated in this study. That is, TBLP test which is in a multiple-choice form is as valid and reliable as the TOEFL.

Pedagogical Implications

As mentioned before, testing involves the process of decision making. And in decision making, the future lives of so many people are at stake. It follows that interpretation of test scores should be made with caution. That is, the results should be obtained through a valid and reliable instrument. This reveals the crucial role of the instrument used.

In order to make sound decisions, according

was because of the nature of the tasks. Since most Iranian students are not familiar with this type of test, lengthy instructions are needed to have them understand what they are supposed to do. Based on the performances of the participants on the pilot administration of the test, the characteristics of the individual items (item facility, item discrimination, and choice distribution) were determined. According to the information gained through item analysis, some items were revised again. In order to ensure test separability, the reliability of the test was also calculated using KR-21 formula at this stage. The results showed a reliability index of .81. The last step was the final administration of the test. At this stage, the new test was administered to a total number of 127 female subjects. Based on the subjects' performance at the pilot administration of the test, the subjects, at this phase, were given one hour to complete the test. The original TOEFL was given to the same subjects the next session. For the TOEFL an hour and a half was allowed.

Results

Several statistical analyses were conducted in order to answer the research question. A brief summary of the findings is presented below.

First, the descriptive statistics for the subsections of TOEFL and TBLP test are given in tables 1 & 2.

Table 1: Descriptive Statistics for the TOEFL

Variable	Mean	SD	Total points
TOEFL(total)	73.58	14.38	150
TOEFL(LC)	18.38	6.32	50
TOEFL(ST)	19.74	5.51	40
TOEFL(VOC)	17.09	4.03	30
TOEFL(RC)	18.38	4.50	30

Table 2: Descriptive Statistics for the Topic-Based Test

Variable	Mean	SD	Total points
TBLP(total)	34.40	7.05	60
TBLP(LC)	10.78	2.82	20
TBLP(RC)	12.91	2.88	20
TBLP(WR)	10.72	3.43	20

Second, in order to estimate and compare the reliability indexes of both tests along with their subsections KR-21 formula was utilized. The findings are presented in table 3.

Table 3: Reliability indexes

Variable	Reliability	Variable	Reliat
TOEFL(total)	.82	TBLP(total)	.71
TOEFL(LC)	.72	TBLP(LC)	.39
TOEFL(ST)	.68	TBLP(RC)	.47
TOEFL(VOC)	.56	TBLP(WR)	.60
TOEFL(RC)	.66		

Third, in order to examine the degree of go-togetherness of the two measures, several correlational analyses were conducted between the two tests and their subtests. First, the correlation coefficient of the two tests was computed. The result showed a correlational index of .81. Then, different correlational analyses were conducted in order to see the degree of go-togetherness of (a) each test with its subsections, (b) each test with the subsections of the other test, (c) the subsections of the TBLP test with those of the TOEFL, and (d) subsections of each test. The results of these analyses are presented in two correlation matrixes presented in tables 4 & 5.

study was collected through two types of tests: an original TOEFL and a topic-based test developed exclusively for the purpose of this research. Each test is briefly explained below.

First, an original TOEFL was utilized as a criterion against which the newly-developed test was validated. The test consisted of listening comprehension, structure and written expression, and vocabulary and reading comprehension sections with 50, 40, and 60 MC items respectively. For each item, one point was assigned. Therefore, the test had a total of 150 points.

Second, a multiple-choice topic-based language proficiency test (TBLP) was developed the topic of which was "air pollution". This topic was selected on the reasoning that some experts do not favor topic-specific tests because of the threat they may impose on test validity due to some subjects' familiarity with the topic. The newly-developed test consists of three sections: listening comprehension, reading comprehension, and writing, each comprising 20 multiple-choice items. For each item, one point was assigned. Therefore, the test had a total of 60 points. Each section starts with the necessary directions on how to deal with different subsections of the test, and, where needed, additional directions were given for each subsection, and examples were provided. In this test, the subjects mainly deal with completing diagrams, inserting missing information, paraphrasing, etc. As an example, in the first activity of the listening section of the test, testees are asked to listen to a passage and fill in a table while listening.

Example

You hear: As a result of the smog episode which occurred in Los Angeles in 1975, 36 people died.

You see:

You see:

	Year	City	Rate of deaths	Rate of illnesses
	1	Los Angeles	2	Unknown

1. a. 1975 b. 1917 c. 1957 d. 1970
2. a. 63 b. 13 c. 36 d. 60

Procedure. To accomplish the purpose of the study, the following procedures were followed:

First, since the listening and reading comprehension passages of the new test had to be comparable in difficulty to those of the TOEFL, the readability indexes of reading passages of the TOEFL test were calculated using the Fog index of readability. Then, it was decided to select those texts which possessed the readability index within the range of one standard deviation above or below the mean readability of the passages in the TOEFL. The chosen paragraphs were taken from the Encyclopedia Americana. The next step was to tape-record the listening passages. This was done by a near-native speaker of the English language at natural rate of speech. Then, multiple-choice items were constructed following Millman and Greene's (1993) *Rules for Writing Multiple-Choice Test Items*. When the test was prepared, it went under repeated revisions, and the final version was used for the pilot administration phase of the study. Although test instructions were provided for each section and subsection of the test, they were also given, before the test began, orally in both languages -- English and Farsi. This

on areas outside their academic field.

The importance of background knowledge in reading comprehension has been addressed by many researchers (Johnson, 1981, 1982; Carrell & Eisterhold, 1983; Aron, 1986; Carrell, 1984, 1987) who stress the salience of content schemata. If students read a passage about a subject with which they are familiar, it seems logical that they will comprehend more than when they read about an unfamiliar subject. But this view may not be true in the sense that providing testees with ample listening and reading tasks on the basis of a particular topic will shed light on the issue in question (Fox, 1996). Clapham (1996) also refers to the importance of background knowledge and maintains that the student's comprehension is affected by his or her ability to draw on specialized knowledge bases and advises against subject-specific testing.

Another problem which may be present in performance tests relates to their scoring. As Upshur and Turner (1995) maintain, scoring these tests are more demanding than scoring DP tests:

Rating language performance is more damanding than scoring discrete-point tests. Lower reliability and validity for ratings are to be expected, ... (Upshur and Turner, 1995, p. 3)

This was confirmed by research conducted on the effect of rater variables on assessment. It was observed that there were significant differences in ratings awarded which reveal different perceptions of what constitutes good performances (Brown, 1995). Bachman and Savignon (1986) also believe that commonly employed rating scales present major problems of reliability and validity.

A third problem with performance tests, which concerns scorers and administrators, is practicality. That is, administering, scoring, and, more importantly, interpreting results are not easy tasks when these types of tests are utilized. This is because most performance tests do not have MC format.

In order to alleviate some of the above-mentioned problems, however, the topic chosen for the new test was a rather general one. As for the second and third problems, all the tasks and activities were constructed in the form of multiple-choice. This was due to the fact that using multiple-choice format can be quite successful because of objectivity in scoring such tests.

Method

Subjects. The subjects of this study were 171 Iranian English as a Foreign Language (EFL) learners selected on the basis of cluster sampling. The sample consisted of female university seniors who were majoring in English Translation at Islamic Azad university.

Of the total number of subjects, 44 took part in the pilot administration phase of the study. Because of the carelessness of few subjects, however, the item analysis procedure was based on the performance of 40 subjects and the rest was excluded from the data. Furthermore, of the 127 subjects participating at the final administration phase of the test, the results of 7 were not included in the data analysis because of leaving one of the tests incomplete.

Instrumentation. The needed data for this

Introduction

Tests are used for a variety of purposes -- to measure students' knowledge in relation to future tasks which they are expected to perform, to place students in appropriate levels, to grant certificates, to determine whether students can continue in future studies, etc. Thus, tests seem to be powerful because many people including language testers, teachers parents, administrators, governments, public and so many other individuals and organizations are affected by test scores (Farhady, 1998).

According to Bachman (1990) "the fundamental use of testing in an educational program is to provide information for making decisions, that is, for evaluation" (p. 54). In order to make sound decisions, it is now believed that individuals' performances should be tested directly using realistic tasks in order to see what they can do with the language (Allison and Webber, 1984). That is why there has been a great tendency on the part of test developers to create tests which assess communicative performance. As a matter of fact, in recent years, much effort has been directed towards developing tests that not only measure a broad range of language abilities but are also **authentic** in that they require testees to interact with and process both the explicit linguistic information and the illocutionary force of the test material. Terms such as "functional" and "communicative" have also been used by researchers to refer to the same notion. That is, all of these terms -- functional, communicative, authentic -- refer to the extent to which the tasks required on a given test are similar to some standard or real-life language use (Bachman and Clark, 1987).

A closely-related issue, in this regard concerns thematic or topic-based tests in which the whole test is built around a single topic. According to Jennings et al. (1998), a number of language proficiency tests which are used in university admission procedures for ESL students (for example, CAEL Assessment[1], IELTS[2], OTESL[3]) have adopted a thematic or topic-based approach in order to reflect language in use. Unlike tests like TOEFL in which testees face independent discrete-point items, topic-based tests use a more integrated approach.

A good rationale for constructing topic based tests is that in real life situations, one of which is the university classroom, the professor usually centers his/her lecture on just one topic and provides the students with ample information and readings on that particular theme or topic. In other words, topic-based tests attempt to provide the context for language use which resembles the language demands in academic settings. Thus, if a test constitutes the elements of real-life situation, it can be said to be a more valid test in comparison to those which do not take advantage of authentic activities.

While topic-based tests are considered to be focusing on real-life situations, choosing a proper topic is not an easy task. That is, one threat to the validity of these types of tests is that test bias may occur as a result of testee's familiarity with the test topic.

This problem is addressed by Alderson and Urquhart (1988) who state that academic background can have an effect on reading comprehension, and that particular groups of students may be disadvantaged by being tested

THE CONSTRUCTION AND VALIDATION OF A MULTIPLE-CHOICE TOPIC-BASED ENGLISH LANGUAGE PROFICIENCY TEST

BY
HOSSEIN FARHADY Ph.D.TEFL
&
BITA SABETI DIANAT
IRAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Abstract

This study was an attempt to construct a multiple-choice topic-based language proficiency test and validate it against traditional language proficiency tests like TOEFL. The subjects participating in the study were 171 Iranian female university seniors majoring in English Translation. The instruments utilized were an original TOEFL and the newly-developed topic-based test (TBLP). The newly-developed test was, first, pretested with 44 subjects and necessary modifications were made. Then, it was administered along with the TOEFL to 127 subjects in two consecutive sessions. The results of the analyses revealed that the new test seemed to be as valid and reliable as the TOEFL. Therefore, it was concluded that the present test, which followed recent approaches to performance testing, might be used as a good indicator of a learner's language proficiency.