Development and Use of a Foreign Language Proficiency Test

Tetsuro Chihara　　Tamara Swenson　　Steve Cornwell

外国語能力テストの開発と使用

智　原　哲　郎　　タマラ　スェンソン　　スティーヴ　コーンウェル

Abstract

This paper describes the development of the 2000 version of the Osaka Jogakuin College English Proficiency Test. A re-evaluation of the 1987 version of the test was considered necessary because of changes in the student population and an increase in the number of students taking the test. The authors undertook this task by first evaluating the 1987 version and determining which areas needed changes or improvement. The authors then wrote new material. New cloze reading passages were pilot tested with the students in the third year program, and versions were then made for pre-testing. Pre-testing of new reading and listening materials was then conducted with first-and second-year students. Following pre-testing, final selection of the material to include in the 2000 version was made. Two version of the revised test, and a short version, are currently in use. The authors plan to evaluate them following the first year's administration.

Key words: language testing, placement tests, test development, analysis

(Received September 13, 2000)

抄　　録

本研究は、大阪女学院短期大学英語能力テストの2000年度版の開発について論じたものである。学生の学力分布の変化並びにテスト受験者数の増加に伴い、1987年に作成されたテストの見直しが必要とされた。著者らは、まず1987年度版テストを評価し、テストのどの部分を変更または改良すべきかを検討した後、問題作成に取りかかった。クローズ方式による読解テストは専攻科生に予備テストを行った後、プリテスト用のものを作成した。最終的に、本学一、二年生に聴解力テストおよび読解力テストのプリテストを行った。テスト問題項目の選別を経て2000年版テストが完成され、現在の使用に至っている。このテストの初年度施行後の評価も計画されている。

キーワード：言語テスト、配置テスト、テスト開発、分析

(2000年9月13日　受理)

Introduction

The function of language testing can be stated in two ways. One is to measure learners' attainment of language skills, while the other is to evaluate the effectiveness of teaching materials/methods in a particular language program. The most frequent of the two is to measure attainment, which can be classified into several categories according to a test's use : 1) screening tests measure a learner's readiness to learn in a particular language learning setting, such as entrance examinations ; 2) placement tests separate learners into the appropriate level of language course ; 3) diagnostic tests provide information on learners' strengths and weaknesses in language skills ; 4) achievement tests measure learners' language achievement in a particular course of study ; 5) proficiency tests measure learners' general language ability without consideration of a specific course of study. It should be noted, however, that these purpose categories are not mutually exclusive. One test can be used for more than one category. TOEFL (Test of English as a Foreign Language), for example, can be a screening test and/or proficiency test. *TOEFL Tips*[1] states as follows:

> The Test of English as a Foreign Language (TOEFL) is designed to evaluate the English proficiency of people whose native language is not English. TOEFL scores are required for admission purposes by more than 2,400 colleges and universities in the United States, Canada, and eighty other countries. Because the TOEFL test is independent of any curriculum or teaching method, the proficiency level of any test taker can be compared with that of any other student or group of students regardless of academic background or English training. (Educational Testing Service, 1999, p. 5)

TOEFL can also be a placement test when it is used to determine the appropriate level or course for students to begin their study. Therefore, it is of no use to say that a particular test is only for one specific task. It is educators that decide what a test is for.

Although some published tests such as TOEFL and TOEIC are available for the purposes mentioned above, they are not always valid in foreign language learning settings. TOEFL has been validated by being administered more than 11 million times in over 180 countries since 1964 (Educational Testing Service, 1999, p. 5). However, in a testing situation where students' English language abilities are not normally distributed, the TOEFL results may not distribute students widely enough for placement or screening purposes.

Another point to be noted is that published tests have a problem of practicality. They usually require a considerable length of time to administer, three to four hours, and it takes from four to six weeks to receive the results. In addition, the administration fees can be cost prohibitive. Therefore, it may be troublesome for institutions to use such tests for the schools' intended purposes, particularly when they want to use them more than once a year.

Developing an in-house language test has some benefits. Once constructed and vali-

dated, in-house tests can be used for the purpose that the institution wants. As far as practicality is concerned, they can be administered any time necessary, in a shorter time, and results can be obtained much faster. After initial development costs, the in-house test costs little to administer.

This paper reports the process of development of the 2000 version of the Osaka Jogakuin Junior College English Proficiency Test and discusses its use.

Origins of the 2000 version of the OJJC English Proficiency Test

Background

In 1987, OJJC established a content-based language curriculum that aimed to improve students' proficiency in English. Prior to the construction of the curriculum, the faculty was charged with developing goals to be attained in English education. According to a volume about OJJC's history published in Japanese, one of these was to develop a test which measured the English proficiency and helped place students in proper levels of proficiency (*Souritsu 30 Shunen Kinen Iinkai*, 1998, p. 29). In order to realize this, a test development working group was established in 1986, and it started writing a test of English for multiple purposes to coincide with the implementation of the new curriculum for the 1987 academic year.

Test Format

After a sequence of procedures on the writing, administration, and analysis of test items, the OJJCEPT (Osaka Jogakuin Jr. College English Proficiency Test) was created in 1987. The test, which had two alternate forms and one shorter version for placement purposes with incoming students, consisted of three sub-tests: listening, structure, and reading (see Table 1). In this, the listening section had 15 statements, 15 short dialogues, and a cloze-style dictation. The structure section had 40 multiple choice questions, and the reading section consisted of one multiple-choice cloze passage and a C-test.[2]

Table 1 : OJJC English Proficiency Test Format

Section	Part	Type	Number of Questions	Points
Listening	Part 1	Statements (Multiple-choice)	15	30
	Part 2	Dialogues (MC)	15	30
	Part 3	Dictation	20	40
Structure		Sentences (MC)	40	40
Reading	Part 1	1 cloze passage (MC)	20	20
	Part 2	2 modified cloze passages	40	40
			TOTAL	200

What helped increase the OJJCEPT content validity as a measure of English language proficiency was the inclusion of a dictation task in the listening section and a cloze procedure in the reading section. Dictation is a very reliable and valid testing procedure to measure foreign language proficiency (Oller, 1979). Oller argues:

> The traditional dictation . . . is an interesting example of a pragmatic language testing procedure. . . . a simple traditional dictation meets the naturalness requirements for pragmatic languages tests. First, such a task requires the processing of temporally constrained sequences of material in the language and second, the task of dividing up the stream of speech and writing down what is heard requires understanding the meaning of the material–i.e., relating the linguistic context to the extralinguistic context. (1979, p. 39)

The cloze test, which requires test takers to utilize their pragmatic expectancy grammar[3] (Oller, 1979), has been in use more than 35 years and have been supported by a number of research studies as a reliable and valid instrument for measuring foreign language proficiency (Chavez-Oller, Chihara, Weaver, Oller, 1985 ; Chihara, Oller, Weaver, Chavez-Oller, 1977 ; Oller, 1979 ; Backman, 1990 ; Oller, Chihara, Chavez-Oller, Yu, Greenberg, Hurtado de Vivas, 1993 ; Brown, 1994).

OJJCEPT was administered two times a year[4] by using the two forms alternately. The short form was used with in-coming students. The scores of OJJCEPT were taken into consideration for information on placement in English classes, eligibility for participation in practice teaching, screening candidates for overseas programs, and observing students' progress in the target language over the two years of the curriculum. A correlation of .87 was found between OJJCEPT and TOEFL (OJJC *Kyomuka*, personal communication, April 1993). An OJJCEPT score of 100 corresponds to TOEFL score of 450–470, OJJCEPT 120 to TOEFL 470–500, and OJJCEPT 140 to TOEFL 500–550.

When the new curriculum started in 1987, the enrollment of new students was 242. The number became 258 in the following year, 346 in 1990, and 408 in 1995. As of the 1999 academic year, the number was 377. As the pool of entrants widened, lower English proficiency classes had more students. The distribution of students' proficiency deviated widely and hence skewed from the mean. Therefore, it was suspected that some of the items of OJJCEPT no longer discriminated well. In addition, as no re-evaluation of the test had been done since its inception, it was felt that the test needed to be evaluated and revised as necessary. Finally, there was some concern that the frequency of test administration might allow for a practice effect and caused some students to skip some administrations because of the perception that the test had no purpose.

In order to provide both educators and students with more valid information on students' English language proficiency, the authors were asked to revise the 1987 version of the test.

Revising the Test

Evaluation of the 1987 Version Test Items

The first stage in the creation of a new version of the OJJCEPT was the evaluation of the existing test. Item analysis was conducted to determine which items in each section of the test discriminated among students and which had the best item facility. The Item Discrimination and Item Facility scores were used to determine which items to retain. In addition, for the listening statements, and dialogues, and for the structure section, the top scoring items were selected as anchor items. (This is discussed in more detail below.)

Pre-testing for New Test Items

Design and Procedures

The authors decided that the revision of test should be minimized because the item analysis showed most items were working. The areas that they determined needed revision included: updating the contents of dictation and cloze passages, eliminating the C-test sub-test, and adding a reading comprehension sub-test in its place. Because the C-test procedure requires test-takers to do a similar task to that of a cloze test and the marking of the C-test was taking more time than was available, the authors decided to eliminate the C-test.

Cloze tests: It should be stated that when the cloze tests were constructed, a series of steps were required. First, five cloze test passages of approximately 250 words each were written by the authors. In each test, blanks were created by deleting every 7th word. When a word to be deleted was a proper noun or either too easy or too difficult to fill in, the following word was deleted. The first and last sentences were left intact. A total of 26 to 27 words were eliminated from each passage. Twelve students studying in the graduate course (third-year program) at OJJC were asked to complete the cloze passages. The responses from this piloting were listed and ordered from high to low frequency for each blank. The top three incorrect words were selected as distractors, and thus the cloze tests for pre-testing were constructed.

Reading comprehension test questions: Eight reading passages of approxmimately 200 words each, and five to six questions with four to six distractors for each question were written by the authors. All passages were similar in length and reading difficulty.

Dictation tests: Three passages of approximately 200 words were written by the authors and six segments of text were eliminated from each passage.

As the final form for pre-testing, eight reading passages, five cloze passages, and three dictation passages along with questions were written. This was considered by authors to be the minimum number needed in order to select at least six reading passages, four cloze passages and two dictation passages to be used in the 2000 version of the test. Table 2 shows the format for pre-testing.

Table 2 : Pre-Test Format

Test	Type	Number of Questions
Reading Comprehension	8 passages	5 to 6 each
Cloze	5 cloze passages	26 to 27 each
Dictation	3 passages	6 each

Pre-testing

Subjects: A total of 146 students (94 first-year students and 52 second-year students) enrolled at OJJC served as subjects in pre-testing. All were volunteers and they came from all levels of English ability. In order to provide broader range of their English proficiency levels, the numbers of the volunteers from each level was balanced.

Test administration: The three tests, reading comprehension test, cloze test, and dictation test, were administered on different days so that the students would concentrate on taking one test at a time. Some students took all of the tests and some others took one or two. The subjects were allowed as much time as needed to answer all the questions.

Analysis of Pre-test Results

After pre-testing was completed, the data was analyzed. Table 3 gives mean percentages of correct answers and their respective standard deviations. The data show that all the tests were challenging for the students. On average, they answered questions correctly about 36% of the time for the reading comprehension test passages, 38% for the cloze test passages, and 47% for the dictation test passages respectively.

Table 3 : Mean Percent Correct-Scores and Standard Deviations on Pre-tests.

Test	*N*	*M*	*SD*
Reading Comprehension Test	105	36.06	5.51
Cloze Test	88	37.69	10.58
Dictation Test	82	46.48	8.59

Test item analyses were conducted on the reading comprehension and the cloze tests. Item facility (IF) and item discrimination (ID) were checked. The IF index gives percentages of correct answers, and the ID index is the Point Biserial Correlation between the item score and the total score on the test. As a rule of thumb, a range of IF with a. 30 and. 70% is usually recommended, while items with IDs far below .30 should be discarded (Brown, 1996, p. 70). For example, IF and ID indices for one of the cloze passages are shown in Table 4. Item 2 (ID = .01), Item 5 (ID = .15), Item 18 (IF = 87.50), Item 21 (ID = .11), Item 24 (IF = 6.82), and Item 26 (ID = . 16) were found to be weak items and thus discarded to construct a total of 20 items. Furthermore, weak distractors in the 20 items were revised.

Table 4 : Item Facility (IF) and Item Discrimination (ID) for Cloze Test Passage 2

Item	IF	ID	Item	IF	ID
1	12.50	.30	14	17.05	.45
2	25.00	.01	15	71.59	.46
3	56.82	.29	16	64.77	.36
4	65.91	.26	17	31.82	.26
5	64.77	.15	18	87.50	.34
6	21.59	.37	19	14.77	.34
7	28.41	.56	20	47.73	.35
8	51.44	.25	21	46.59	.11
9	37.50	.59	22	30.68	.36
10	25.00	.51	23	14.77	.32
11	64.77	.35	24	6.82	.78
12	55.68	.41	25	44.32	.24
13	63.64	.28	26	27.27	.16

As an external validity check, Pearson product-moment correlations were computed for the three tests and the various parts on the two forms of 1987 version (see Table 5). The analyses indicated that there was a moderate relationship between them. Correlations between the cloze tests and the dictation tests and their corresponding parts of the 1987 version of the test ranged from .57 to .66. However, the reading comprehension test passages correlated lowest with the reading section of the 1987 version of the test (.41 for Form A and .52 for Form B). This may stem from the low average score by students on the pre-testing passages.

Table 5 : Matrix of Product-moment correlations between the pre-tests and the OJJCEPT

	Number of examinees	correlation coefficient
Dictation Test x PT Form A-Listening	74	.65
Dictation Test x PT Form B-Listening	70	.66
Reading Comprehension Test x PT Form A-Reading	95	.41
Reading Comprehension Test x PT Form B-Reading	90	.52
Cloze Test x PT Form A-Reading	79	.65
Cloze Test x PT Form B-Reading	75	.57
(Reading Comprehension Test+Cloze Test) x PT Form A-Reading	54	.65
(Reading Comprehension Test+Cloze Test) x PT Form B-Reading	52	.58
PT From A x PT Form B	473	.87

Designing the 2000 Version

What emerged from the period of evaluation, design, pre-testing, and re-evaluation was the contstruction of a new test, the 2000 version. The 2000 version has two alternate forms

and a short version. Additionally, to address the concern that students were skipping some administrations, the forms of the test were given different names to make their purposes clear: First Year Placement Test, Proficiency Test, Second Year Placement Test, and Graduation Test. This was intended to increase the face validity of each test and hence to motivate students to take each one seriously. The Proficiency Test administered at the end of the first term of the first year and Graduation Test administered at the end of the second term of the second year are identical so that students' progress over the two years can be observed. Additionally, it was decided not to administer the test in any form to students in the middle of their second year as these results were not used by the school.

Finally, the authors decided that all three forms of the test should include some of the same items, anchor items, in the listening section statements and dialogues and in the structure section. The items with the best Item Discrimination scores on the analysis of the 1987 version of the exam were selected to serve as anchors. Three listening section statements, three dialogues, and eight structure questions were selected to serve as anchor items in all versions of the test. Because the authors felt that repeating cloze, dictation, or reading passages on all the three forms of the test would be detrimental to their overall value, no anchor items were included for these.

Test Format

The 2000 version of the test was finalized for administration. Table 6 shows the test format. The test consists of three sub-tests: Listening, Structure, and Reading. The total score is 200. All directions are given in Japanese.

Table 6 : English Proficiency Test Format

Section	Part	Type	Number of Questions	Points
Listening	Part 1	Statements (Multiple-choice)	15	30
	Part 2	Dialogues (MC)	15	30
	Part 3	Dictation	20	40
Structure		Sentences (MC)	40	40
Reading	Part 1	3 passages (MC)	20	20
	Part 2	2 cloze passages (MC)	40	40
			TOTAL	200

Listening Section

The listening section consists of three parts: Part A (statements), Part B (dialogues), and Part C (dictation). In Part A, examinees hear a sequence of sentences and then choose the

most appropriate answer to the statement from four answer choices written in the test book.

Example: Upon being questioned by his father, George Washington admitted that he cut down the cherry tree.

(Written in the test books)

a) George Washington was angry.

b) George Washington was honest.

c) George Washington was confused.

d) George Washington was shy.

In Part B, examinees hear short conversations between two people. Each conversation is followed by a question from a third speaker. Examinees choose the most appropriate answer to the question from four answer choices written in the test book.

Example:

(Students hear two speakers)

Man: Thanks for agreeing to donate some of your work.

Woman: Well, it's the least I could do when you said Bob had given some poems, and Joan had given some music.

Man: Do you ever work in color?

Woman: No, I only shoot black and white.

(Third voice)

Q: What did the man receive from the woman?

(Written in the test books)

a. Some poems

b. Some music

c. Some paintings

d. Some photos

Part C is a cloze-type dictation[5]. Examinees see a passage in which six parts are deleted. The passage is read once, and then each deleted part is read two times with pauses. Examinees write down what they hear for the missing parts on the answer sheet.

Example:

Like the United States, Japan is a highly industrialized and urbanized society. However, the Japanese view of aging is markedly different. In contrast to seniors in America, the elderly in Japan occupy a position of honor. In part, this respect for age [1] not readily adaptable to American culture. . . . To encourage the active involvement of all older citizens in social activities, the government subsidizes Elder Clubs and sports programs. Through these programs the elderly supply each other [6].

(repeated twice)

1. ... respect for age is based on ancient social and religious traditions not readily

.

.

.

6. ... supply each other with mutual support and gain a sense of self-pride.

Students write the underlined words in the space provided on the answer sheet. Question 1, for example, would be divided into three segments, each worth one point, as follows:

is based on / ancient social / and religious traditions

Structure Section

The structure section aims to check examinees' formal use of English grammar. Examinees choose the most appropriate answer from four answer choices to complete a sentence.

Example: I had difficulty_____in English.

a) making myself understood

b) making myself understand

c) to make myself understood

d) to make myself understand

Reading Section

The reading section consists of two types. One requires that examinees read three passages of approximately 200 words each and answer 3 to 4 questions by choosing the most appropriate answer from four answer choices.

Example:

The growth in the human population began slowly but exploded rapidly. At the beginning of . . . However, in the late 1700s, the population began to surge because of a dramatic drop in the death rate. This can be traced to construction of sewage systems, the discovery of vaccinations against diseases, and improvements in hygiene. In addition, increased agricultural output meant most people ate enough and, therefore, lived longer and healthier lives. In other words, the population explosion was caused by improvements in public facilities, health care, and food production. As these combined to reduce death rates, the lack of a corresponding drop in the birth rates led to rapid population growth.

Questions

.

.

.

3. What is not **one** of the reason for the change in the populations beginning in the 1700s?

a. High birth rate.
b. Drop in death rate.
c. Discovery of new medicines.
d. Increasing cleanliness.

4. What is the best title of this passage?
a. The Population Explosion
b. The Drop in Death Rates
c. Population Growth and Life Expectancy
d. Changes in Population

The other type of readings are cloze-type reading comprehension passages. Twenty words from a passage of approximately 250 words are deleted, according to the procedure outlined above. Examinees must fill in each blank by selecting one of the four choices given.

Example:

Curricula in many countries are built slowly over a period of many years. However, when the situation in a ___(1)___ changes, the curriculum may need to be revised to reflect the new educational ___(2)___. This is what the country of Zimbabwe ___(3)___ after its independence in 1980 . . . .

1.	a) year	2.	a) needs	3.	a) soon
	b) period		b) duties		b) knew
	c) school		c) restructure		c) sought
	d) country		d) curriculum		d) discovered

Discussion and Conclusion

We were quite pleased during the process of evaluation to realize how effective the 1987 version of the test was. Though a few individual items had become less useful to OJJC, the overall quality of the exam was high.

Additionally, the process of evaluation and construction of the 2000 version of the placement test gave many benefits. OJJC can administer the test when it wants, the results are more timely, and the costs are far lower than any published test. Furthermore, extra administrations can be scheduled as necessary to accomplish our goal of near 100 percent participation. More importantly, by having students take identical forms after one term of study and prior to graduation, we have a clearer idea of the students' actual increase in English proficiency. By reducing the number of administrations of the same form, and keeping them widely separated, we hope to decrease any practice effect.

Regarding the future, the need to evaluate and analyze the test on a regular basis is clear. Though we may not be able to complete yearly evaluations, because of the time involved, regularly scheduled evaluations should be undertaken to ensure that items maintain a high

degree of validity.

A step in this direction will be the planned evaluation of the first years' administration of the 2000 version of the placement test. This will allow us a clearer understanding of how the test is working with our current students.

A concern we have is that, while this test is appears to satisfy most of our needs, it still does not include oral or communicative components. Both of these have been suggested as necessary for a complete picture of students' proficiency. However, their inclusion was beyond what we were both directed and able to do during the 12 months of test development.

Notes

1. A guidebook published by ETS (Educational Testing Service) to provide information on how to take TOEFL.
2. The test, developed by Raatz and Klein-Braley in 1982, is a modification of the cloze procedure retaining the underlying theory of general language proficiency. Chihara, Cline and Sakurai (1996) showed that C-tests were representative samples of language as far as the ratio of content and structural words in the text were concerned, and that they were validated with TOEFL.
3. A theory developed by Oller in the 1970s. "*A pragmatic expectancy grammar* is defined as a psychologically real system that sequentially orders linguistic elements in time and in relation to extralinguistic contexts in meaningful ways" (Oller, 1979, p. 34).
4. The 1987 version of the test was administered in September and February until 1997, and was given in July and December due to a change to a trimester system in the academic calendar in 1998. The 2000 version was administered in April, 2000 to in-coming students and at the end of the spring term in July, 2000. Administrations are scheduled to occur in April (to incoming students), the end of the first term (to first-year students), the end of the first year of study (to first-year students), and prior to graduation (to second-year students).
5. For scoring dictation, the deleted parts are divided into meaningful segments. Each of the segments is considered one item. A segment without error, including spelling errors, is allowed one point.

References

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

Brown, H.D. (1994). *Principles of language learning and teaching*. Third edition. Englewood Cliffs, NJ : Prentice-Hall.

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

Chavez-Oller, M.A., Chihara, T., Weaver, K., & Oller, J.W. Jr. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning*, *35*, 181–206.

Chihara, T., Oller, J. W. Jr., Weaver, K., & Chavez-Oller, M.A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning*, *27*, 63–73.

Chihara, T., Cline, W., & Sakurai, T. (1996). If the cloze test is a question, is the C-test the answer? In R. Grotjahn (Ed.), Der *C-Test : Theoretische grundlagen und praktische anwendungen band3* (pp.183–195). Bochum, Germany : Universitätsverlag Dr.N.Brockmeyer.

Educational Testing Service. (1999). *TOEFL Tips : Preparing students for the computer-based test*. Princeton, NJ : Author.

Oller, J.W. Jr. (1979). *Language tests at school : A pragmatic approach*. London : Longman.

Oller, J. W. Jr., Chihara, T., Chaves-Oller, M.A., Yu, G.K.S., Greenberg, L., & Hurtado de Vivas, R. (1993).

The impact of discourse constraints on processing and learning." In J.W. Oller Jr. (Ed.), *Methods that work : Ideas for literacy and language teachers* (pp.206–29). Boston: Heinle & Heinle.

Ratz, U. & Klein-Braley, C. (1982). The C-test—a modification of the cloze procedure. In T.Culhane, C. Klein-Braley & D.K. Stevenson (Eds.), *Practice and problems in language testing 7* (pp. 113–138). Colchester : University of Essex.

Souritsu 30 Shuunen kinen iinkai, (Ed.)(1998). *"Nani ga dekite, nani ga dekiteinaika" : Souritsu 30 shuunen kinen "jikokentoushi"*. Osaka: Osaka Jogakuin Junior College.