TOEIC®

Research Summaries

An Introduction to TOEIC: The Initial Validity Study

by Protase E. Woodford

ETS® Educational
Testing Service

Reprinted from "The Test of English for International Communication" by Protase Woodford. In English for International Communication, Christopher Brumfit (Editor), Pergamon Press: Oxford and New York, 1982.

This work was originally delivered as a paper at the English Speaking Union Conference on English in International Communication, November 1980.

Introduction

Educational Testing Service in Princeton, New Jersey was founded in 1947. Prior to 1947, three major institutions: The College Entrance Examination Board, the Carnegie Foundation for the Advancement of Teaching and the American Council on Education had each, as non-profit educational agencies, carried out testing programs for schools, colleges and universities in the United States. In 1947 the three organizations founded a new, independent, nonprofit organization dedicated to the development of programs and research in educational measurement and assessment. The new organization — Educational Testing Service (ETS) — has become the largest and most well-known educational measurement organization in the United States and perhaps in the world. Currently, there are over 2,000 full-time employees in Princeton and ten regional offices from Puerto Rico to California.

Test of English for International Communication (TOEIC) — PROTASE E. WOODFORD, Educational Testing Service, Princeton

ETS experience in foreign language testing dates from the beginning of the organization. The College Board, one of the three parent organizations, had administered foreign language tests since the turn of the century. It is interesting to note that the foreign language tests changed very little between 1900 and 1960. Foreign language teaching stressed the development of reading skills only and the memorization of grammatical rules. Foreign language tests were limited to testing reading ability and the knowledge of grammatical rules.

In the 1960s, the emphasis in foreign language teaching in the United States changed from reading and writing to understanding and speaking. The testing industry, always a few steps behind, began to find ways to measure these new skills.

For many years foreign language teachers taught something they called "language" but often had little to do with the ability to communicate with a speaker of the language being taught. We cannot blame the teachers because they themselves often lacked those same skills. Furthermore, it was rare that a student would ever actually find himself in a real life situation that required him to communicate in another language. And so language instruction and language testing existed in a vacuum, unaffected by the real world. Those few people who really had to learn French went to Paris for a few years, those who needed German, to Berlin, Russian, to the Soviet Union. I expect that a similar situation existed in other countries. But as you will see, the way in which we test can inform the manner in which we teach.

Consider a language as the language teachers and testers do, not as a single unit but in its four component parts or skills: Listening or understanding, speaking, reading and writing. Two of the skills are receptive or passive — reading and listening — and two are active, productive or creative — speaking and writing. The testing of passive skills has always been much more common than the testing of productive or active skills for a variety of reasons. One of the most important reasons is that passive or receptive skills can be measured by objective, machine-scorable tests. Productive skills testing always requires human judging and thus becomes more expensive, more time-consuming and less reliable. The Test of English as a Foreign Language (TOEFL®) is administered to tens of thousands of students on the same day around the world. The books are sent to Princeton, New Jersey and the scores or marks are sent to the students within a few weeks. This is possible because the test is machine scorable. If a speaking test or a writing test were added and had to be rated by a person, it would most likely take months rather than weeks to report the scores and be much more expensive. The test, therefore, is limited to testing passive or receptive skills.

There are two words basic to the vocabulary of the test makers — reliability and validity. A test must be both valid and reliable if it is to serve any useful purpose. A reliable test is one that gives consistent results. If I take a mathematics test now and score 200 and take another version of the same test tomorrow without having learned any more mathematics, I should get 200 again. When we measure a kilometer of distance, our kilometer always is 1000 meters long. If my kilometer is 800 meters and yours is 1100 meters, we do not have reliable measurement. Reliability then is the consistency of the measure. A test can be reliable, however, and not necessarily be valid. A test is valid if it tests what it is supposed to test. This seems obvious but many, many tests fail completely to test what they say they test. Nowhere is this more true than in the testing of foreign languages. You might know all the rules of Arabic grammar, the history of the Middle East and thus receive very high marks on an Arabic test. But if the test is supposed to indicate how well you speak Arabic and doesn't require you to speak Arabic, it is of doubtful validity. You might be permanently unable to speak and still do very well on such a test. You also would do well on the test if it were given to you a second or third time. The test would be a very reliable but invalid test of your ability to speak Arabic.

In addition to the questions of validity and reliability, there is the issue of interpretability. What does a score or mark mean? Those who have taken the TOEFL know that 650 or 700 is a "good" score and "400" is "not so good." But what do we mean by "good" or "not good"? For tests like the TOEFL, the Graduate Management Admissions Test (GMAT®), and others, a score or mark is a way to compare performance against that of a standard or reference population. The tests used for university admission have a scale from 200 to 800, a mean or average of 500. The standard deviation is 100. That means that about 68% of all scores fall between 400 and 600. Scores above 600 are "very good" and scores below 400 are "not good." The scores themselves, however, provide only one piece of information about an examinee. Universities that accept the examinee will follow their history. If most of the students who score over 550 and who have good secondary school records succeed at the university and those below 500 do not, then the university may set 550 as the minimum acceptable score. The kind of validity involved with university admissions tests is "predictive" validity. Do the tests predict student performance in the future at a university? Foreign language tests require a different kind of validity. We call it "concurrent" validity. If a language test is supposed to measure whether a person can read Japanese or not, then the person who scores high on the test should be able to pick up the Japanese newspaper and tell us what the lead article says. The low scorer should not be able to do it.

To say that someone is in the top 10% of the group that took a test is not very informative, if we don't know what that high scorer is capable of doing. We need a description of the tasks that can be accomplished by examinees at different score levels. This, to my knowledge, has rarely been done before.

TOEIC® is a multiple-choice test of English for nonnative speakers of English. It consists of two sections: Listening Comprehension and Reading. There are one hundred questions in each section. In the Listening Section, the examinee is required to listen to a variety of recorded English stimulus material and answer questions printed in English in the test book that are based on the recorded stimuli. In the Reading Section, the examinee is required to read a variety of passages of varying subject matter and of different lengths and levels of difficulty, and respond to questions based upon the content of the passages.

Separate scaled scores are provided for each section of TOEIC. The part score scales range from 5 to 495. The total score is obtained by adding the two part scores thus providing for a total score scale ranging from 10 to 990.

What was proposed was a departure from the traditional academic testing of English grammar rules and literature. The challenge, as we saw it, was to develop highly valid and reliable measures of real-life reading and listening skills and, to the extent possible, indirect measures of speaking and writing. In addition, and this is the most exciting part of the project, we were to develop a procedure for score interpretation that would allow score recipients to actually see the kind of English that the examinees at different score levels could read and to see also, typical samples of the examinees' writing efforts, in English, for different score levels. Score recipients would also be provided with samples of English speech by examinees at different levels. For many years programs such as TOEFL have attempted to develop such a system for score interpretation. To develop such a system with examinees from hundreds of different language backgrounds is a monumental job. Separate materials would have to be created for speakers of Spanish, Chinese, Japanese, French, etc. because we know that the kinds of native language interference are different for each group. Now, however, with a monolingual examinee population, the development of interpretive materials with performance samples became feasible.

TOEIC was designed to meet the need for a measure of English language skills outside of the traditional academic context. What was sought was the development of highly valid and reliable direct measures of real-life reading and listening skills. It was hoped that TOEIC would also provide, indirectly, an indication of an examinee's ability to write and speak English as well.

The final test specifications called for 100 items or questions in listening comprehension and 100 in reading comprehension as follows:

LISTENING COMPREHENSION		
Part I	One picture, four spoken sentences	20 items
Part II	Spoken utterances, three spoken responses	30 items
Part III	Short conversation, four printed answers	30 items
Part IV	Short talks, four printed questions and answers	20 items
	TOTAL	**100 items**

READING COMPREHENSION		
Part V	Incomplete sentences	40 items
Part VI	Error recognition — underlines	20 items
Part VII	Reading comprehension — passages	40 items
	TOTAL	**100 items**

The first form of TOEIC was administered in Japan on December 2, 1979. 2,710 examinees were included in the sample upon which the analysis is based.

On the basis of the preliminary item analysis two items from Part IV and one item each from Parts VI and VII were deleted from the final scoring of the test. Therefore, the raw scores for each section of the test were based on 98 items, 98 for listening comprehension and 98 for reading, or a total of 196 scorable items out of the 200 in the full examination. The scores that were reported were not raw scores, they were converted scores.

The raw scores on every form of TOEIC will be converted to the common scale established at the first administration. For each of the two sections, the scale was set to range from 5 to 495. The Total Score is the sum of the converted scores for the two sections. Thus, the range of possible Total Scores is 10 to 990. A statistical procedure called "score equating" will be used to determine the appropriate conversion formula for each new form so that a given converted score (e.g. 640) will represent the same level of ability regardless of the form taken or the ability level of the group with whom it was taken.

It is important to note that the total score is not *directly* related to the total number of correct answers. Subsequent forms of the TOEIC have been equated through section scores and not from total score to total score.

The examinee group who sat for the TOEIC in December 1979 contained an unexpectedly large proportion of persons who were *not* affiliated with companies. For this reason separate raw score statistics were obtained for each group.

TABLE A: RAW SCORE DATA
(Based on 98 items listening, 98 items reading)

	Company affiliated	Unaffiliated
Number of persons	828	1,884
Listening comprehension		
Mean score	56.98	62.55
Standard deviation	14.42	13.75
Reading comprehension		
Mean score	64.97	69.75
Standard deviation	15.69	13.72
Correlation between listening comp., reading comp.	0.80	0.74

It should also be noted that because an examinee's listening comprehension and reading comprehension scores could be compared to each other, the section scores were scaled in such a way that the means and the standard deviations for the two sections are equal. An important result of this procedure is that the two sections have equal weight or importance in the total score.

Is the Test Appropriate for the Examinees?

Middle difficulty is defined as the midpoint between the expected chance score — the score that would be expected if every item were answered at random — and the maximum possible score. Middle difficulty for the TOEIC listening comprehension section, consisting of 30 three-choice and 68 four-choice items would be a score of 62.5.

Therefore, the listening comprehension section for the "affiliated" group was somewhat harder than middle difficulty since the mean score for that group was 56.98. However, the listening section was almost exactly at middle difficulty for the "unaffiliated" group whose mean score was 62.55.

The reading comprehension section was easier than middle difficulty (61.25) for both the "affiliated" and the "unaffiliated" groups whose mean scores were 64.97 and 69.75 respectively.

Table A shows that the raw score for listening comprehension section was 21 to 98 (of 98). The range for the reading section was 17 to 96 (of 98).

Of the 20 items in Part I — one picture four spoken choices, the mean score for the 2,710 examinees in the sample was 16.44 or a mean of 82% of the total possible score.

Part II (Question and three spoken responses) was slightly harder than middle difficulty (20) for three-choice items. Parts III and IV (short conversations and short talks) were the two most difficult parts of the test.

The most difficult part in the reading section was Part VI (error recognition), which was slightly harder than middle difficulty. Parts V and VII (incomplete sentences and reading comprehension passages) were relatively easy.

The parts of the test arranged in order of increasing difficulty are I (easiest), VII, V, VI, II, IV, III (hardest).

▶▶▶▶ Reliabilities

The reliability of the listening comprehension section was 0.916 and the standard error of measurement in scaled score units was 25.95.

For the reading section, the reliability was 0.930 and the standard error was 23.38.

Total test reliability was estimated at 0.956 and the standard error was 34.93.

These reliabilities are well within the generally accepted limits for measurement of individual achievement.

▶▶▶▶ Correlation Between the Two Sections — Listening and Reading

The correlation between the sections was 0.769 for the analysis sample. This would indicate that each score provides somewhat different information about the examinee and justifies reporting separate scores.

Is the test too long or too short for the time available to the examinee?

Because the listening section of the test is timed and paced by the tape recording, it is assumed that all the examinees finish the section. Eighty-seven per cent of the examinees in the sample completed the reading comprehension section of the test and 99.5% completed three-quarters of the test.

It is also interesting to note that the average number of questions not answered was less than one for listening comprehension and less than two for reading comprehension.

These data indicate that speed was not an important factor for either section of the test.

How difficult was the test for the Population?

The average percent correct for the items in the listening comprehension section was 62%.

The average percent correct for the items in the reading section was 70%.

How well do the test questions discriminate?

The criterion used for each item is the section of the test in which the item appears. The mean biserial correlation for the listening section was 0.45.

The mean biserial correlation for the reading section was 0.49.

▶▶▶▶ The Scale

The TOEIC scale has a range from 5 to 495 for each section. For the Listening Comprehension section the observed range — the scores actually obtained by the examinees — went from a low of 40 to a high of 495. The mean scaled score was 290.

The observed range of scaled scores for the reading section was from a low of 5 to a high of 455. The mean scaled score was 288. (No *real* score of 288 exists since all scores are reported in multiples of 5. A 288 score would be reported as 290.)

As shown in Table A (see page 6), most of the scores on the listening section fall between 200 and 370. Approximately 68% of the scores fall within that range.

For the reading section most scores fall between 210 and 385. Approximately 70% of the scores fall within that range.

The total score for TOEIC is the sum of the two section scores as was mentioned earlier. The mean total scaled score was 578. Most total scores fall between 400 and 745. Sixty-eight percent — approximately — of the examinees' scores were within that range.

It is quite gratifying to note that the scale functions as intended. Almost all points on the scale are utilized for both sections of the test as well as for the total score.

▶▶▶▶▶▶▶ TOEIC Validity Study

Subsequent to the first administration of the Test of English for International Communication on December 2, 1979, a series of validation exercises were carried out in Japan to determine the degree to which performance on the TOEIC corresponded to performance on more direct measures of each of four language skills: Listening Comprehension, Reading, Writing, and Speaking. In addition, a version of the Test of English as a Foreign Language (TOEFL) was administered to a sample of TOEIC examinees in order to determine the relationship of performance on one measure to performance on the other.

When score distributions were obtained for the first administration of TOEIC, 500 examinees were selected to take TOEFL. The 500 were selected on the basis of their scores on the TOEIC. One hundred examinees were selected at each of five approximate score levels: 950, 765, 580, 315, 45. A smaller group of 20 examinees from each group of 100 was selected. To these examinees a series of direct measures of language ability were administered.

The Direct Measures were as follows:

Listening Comprehension

Twenty-five taped English stimuli consisting of 15 short statements or questions and 10 dialogues were played to the examinees. For each of the twenty-five exercises there were three questions to be answered by the examinees. The questions were asked in Japanese by a Japanese examiner, and the examinees were encouraged to answer in Japanese. There was a total of 75 scorable items on the Direct Measure of Listening Comprehension.

Reading Comprehension

Reading tasks in English of many kinds were presented to the examinees. Some exercises consisted of a single English word as it might appear on a label or a sign. Other exercises required the examinee to read a table of contents in a catalog in order to find a particular product; or to understand a piece of advertising copy. Examinees were provided ample time to read each selection. When an examinee had completed reading the selection, an examiner would ask questions, in Japanese, about the content of the selection. There were 30 content questions in the Direct Measure of Reading. The examinees answered the questions, orally, in Japanese.

Writing

There were three parts, each with a different kind of exercise included in the Direct Measure of Writing. The first part consisted of 10 "dehydrated sentences." The "dehydrated sentences" were sentence elements from which the examinee was to produce a coherent English sentence making any necessary changes or additions, for example: employees/receive/raise/next year → The employees will receive a raise next year. Fifteen minutes were allowed for Part I.

In the second part, the examinee was required to write a 25-40 word letter to a manufacturer complaining about the manufacturer's delay in shipping an order to him or to her. Examinees were given twenty minutes in which to write the letter. In the third part of the Direct Measure of Writing, the examinee was asked to write the English translation of ten Japanese sentences. The sentences were chosen because they contained specific structural and lexical problems.

Possible scores on Part I ranged from 0 to 50. Possible scores on Part II ranged from 0 to 14. Possible scores on Part III ranged from 0 to 75.

A composite direct measure of English language writing skill was created from the three direct measure exercises. This was done to create a single score that would reflect the various components of writing skill in a reliable way. Scores from the three exercises were made comparable by the process of standardization. Each person's score on each exercise was transformed by subtracting the group's mean score on the exercise from the person's score, and dividing by the group's standard deviation. This score was multiplied by 10 and added to 50. In this way each exercise was placed on a scale with a mean of 50 and a standard deviation of 10. Based on an analysis of the writing tasks, it was decided to differentially weight the three tasks. The following table gives the means, standard deviations, and weights for the three tasks:

TASK	MEAN	STANDARD DEVIATION	WEIGHT
Dehydrated sentences	37.824	7.243	0.3
Business letter	5.859	3.211	0.5
Sentence translation	64.033	9.406	0.2

The range for the composite score for the Direct Measure of Writing was 12-70.

Speaking

The Direct Measure of Speaking Ability was the Language Proficiency Interview (LPI) used by the U.S. Department of State, the Peace Corps and various state and local government agencies. The LPI is a face-to-face interview procedure. Examinees are rated on a 0-5 scale with plus values for all ratings from 0 through 4. For this study interviewers in Japan who were native speakers of English and who were experienced linguists or language instructors were trained by ETS staff to conduct the LPI. Each interview was recorded and the recording sent to ETS in Princeton where it was rated by an experienced ETS rater.

Results

The examinees who underwent the direct measures were divided into five groups for purposes of analysis. Examinees were grouped according to their part scores on the TOEIC. For both Listening and Reading, Group I had TOEIC part scores below or equal to 100; Group II had TOEIC part scores between 100 and 205; Group III had TOEIC part scores between 205 and 300; Group IV had scores between 305 and 400; and Group V had scores at 405 or above.

Listening

Ninety-nine examinees were included in the sample that took the Direct Listening exercises. The total possible score was 75. The mean scores for each group are as follows:

Group I	(TOEIC Listening part scores equal to or less than 100)	15.4 (of 75)
Group II	(TOEIC Listening part scores between 105-200)	23.4 (of 75)
Group III	(TOEIC Listening part scores between 205-300)	45.0 (of 75)
Group IV	(TOEIC Listening part scores between 305-400)	56.4 (of 75)
Group V	(TOEIC Listening part scores above 400)	65.6 (of 75)

The two listening measures correlate very highly — 0.90. The TOEIC Listening Section has a multiple choice format. The examinee must *read* the answer choices in English. For that reason, the TOEIC Listening Section is not a "pure" test of listening ability. The results of the study indicate that the Listening Section of TOEIC is indeed an accurate indicator of an examinee's ability to comprehend spoken English.

Reading Comprehension

A total of 99 examinees were administered the Direct Reading Measures. The total possible score was 30. The mean scores for each group are as follows:

Group I	(TOEIC Reading part scores equal to or less than 100)	10.0 (of 30)
Group II	(TOEIC Reading part scores between 105-200)	17.6 (of 30)
Group III	(TOEIC Reading part scores between 205-300)	22.6 (of 30)
Group IV	(TOEIC Reading part scores between 305-400)	24.1 (of 30)
Group V	(TOEIC Reading part scores above 400)	26.8 (of 30)

Examinees were required to read English language material and answer questions in Japanese posed by Japanese examiners. The TOEIC Reading Test is in a multiple choice format. The examinee reads a question in English based on the content of the selection and chooses the one of four printed English options that he or she considers to be the best answer to the question.

The high degree of similarity of performance by the examinees on both the TOEIC Reading section and the Direct Measure of Reading suggest that the TOEIC Reading Test provides a good indication of the examinee's ability to read English with understanding. The correlation between the two reading measures is 0.79.

Speaking

The level of agreement between TOEIC Listening part scores and the ratings for the Language Proficiency Interview is very high. There were 100 examinees included in the sample to whom the interview (LPI) was administered. The highest possible rating that could be achieved was a 5.0. The mean LPI ratings for each group are as follows:

Group I	(TOEIC Listening part scores equal to or less than 100)	0+ (0.76 of 5.0)
Group II	(TOEIC Listening part scores between 105-200)	1 (1.16 of 5.0)
Group III	(TOEIC Listening part scores between 205-300)	2 (1.99 of 5.0)
Group IV	(TOEIC Listening part scores between 305-400)	2+ (2.66 of 5.0)
Group V	(TOEIC Listening part scores above 400)	3+ (3.53 of 5.0)

A "plus" is recorded as 0.7.

The correlation between the TOEIC listening part score and the direct Language Proficiency Interview is 0.83. This high degree of correlation would seem to indicate that the TOEIC part score is a good predictor of the candidates' abilities to speak English even though the objective measure tests a receptive oral skill while the direct speaking measure tests a productive oral skill.

Writing

Three hundred six examinees were included in the sample that took the direct writing exercises. The total possible weighted composite score was 70. The mean scores for each group are as follows:

Group I	(TOEIC Reading part scores equal to or less than 100)	23.8 (of 70)
Group II	(TOEIC Reading part scores between 105-200)	33.8 (of 70)
Group III	(TOEIC Reading part scores between 205-300)	46.0 (of 70)
Group IV	(TOEIC Reading part scores between 305-400)	50.9 (of 70)
Group V	(TOEIC Reading part scores above 400)	57.2 (of 70)

The direct writing measures correlated 0.83 with the TOEIC reading part score. This high correlation suggests that the TOEIC reading score is a good indication of the examinee's ability to write in English. It should be noted that the reading section of TOEIC contains questions that relate both to reading and to writing. Since these two component parts correlate highly with each other, as well as with the direct measures in writing and reading, a separate part score for writing is not necessary.

TOEFL

A total of 187 examiners were administered both TOEFL and TOEIC. Average TOEFL listening scores are presented for the five groups based on TOEIC Listening Comprehension score.

Group I	(TOEIC Listening part scores equal to or less than 100)	40.0
Group II	(TOEIC Listening part scores between 105-200)	41.9
Group III	(TOEIC Listening part scores between 205-300)	48.9
Group IV	(TOEIC Listening part scores between 305-400	54.7
Group V	(TOEIC Listening part scores above 400)	59.6

Likewise, average TOEFL Reading Comprehension scores are presented for the five groups based on TOEIC Reading Comprehension score.

Group I	(TOEIC Reading part scores equal to or less than 100)	34.7
Group II	(TOEIC Reading part scores between 105-200)	43.1
Group III	(TOEIC Reading part scores between 205-300)	48.0
Group IV	(TOEIC Reading part scores between 305-400)	54.3
Group V	(TOEIC Reading part scores above 400)	60.2

Conclusion

It can be concluded from an analysis of the data that TOEIC provides a good indication of candidates' language abilities in English. The Listening Comprehension part score of TOEIC correlates very highly with other measures of both listening and speaking. The TOEIC Reading part score correlates highly with other measures of candidates' abilities in both reading and writing. Although the mean scores of all the direct measures show a consistent relationship with the appropriate TOEIC part scores, it should be remembered that a standard error of measurement is inherent in all the measures in this study. Therefore, not all candidates scoring high or low on one measure will necessarily score equally high or low on another measure.

TOEIC®

Test of English for International Communication

85071-09780 • S24M20 • 275689 • Printed in U.S.A.