

TEST OF ENGLISH FOR INTERNATIONAL COMMUNICATION

Technical Manual



About this Technical Manual

The *TOEIC[®] Technical Manual* was developed to allow TOEIC users to examine the technical characteristics of the TOEIC test and determine its appropriateness for use in specific situations. The *TOEIC Technical Manual* differs from the *TOEIC User Guide*. The *User Guide* describes the uses and administration of the TOEIC test while the *Technical Manual* explains the inner workings of the TOEIC test; for example, test scoring, development of alternate forms, validity, and reliability. The *Technical Manual* was written for readers with at least a basic understanding of statistical terminology. For those who do not have such a background or need a refresher, a glossary of the statistical terms used in the manual is included in Appendix A at the end of the Manual. For additional information about the TOEIC test, readers can refer to the TOEIC internet web site, at www.toeic.com, to the *TOEIC User Guide*, or to TOEIC representatives worldwide.

TABLE OF CONTENTS

	Page
TECHNICAL SUMMARY	SECTION I
Validity.....	I-2
Reliability.....	I-2
Standard Error of Measurement.....	I-3
Standard Error of the Difference	I-3
GENERAL TOEIC INFORMATION.....	SECTION II
Introduction.....	II-1
Content of the TOEIC test.....	II-1
Format of the TOEIC test	II-2
<i>Section I: Listening Comprehension</i>	II-2
<i>Section II: Reading Comprehension</i>	II-3
Reporting of test scores	II-3
Correlation between the Listening Comprehension and Reading Comprehension sections.....	II-4
Scale scores	II-4
Equating of TOEIC test forms	II-4
Score ranges.....	II-5
Speededness.....	II-5
VALIDITY.....	SECTION III
TOEIC test validity	III-1
Construct-related validity	III-1
Relationship with direct speaking measures	III-1
Relationship with direct listening measures.....	III-2
Relationship with direct reading measures.....	III-4
Relationship with a direct writing measure	III-4
Relationship with class levels	III-5
Relationship with classroom performance.....	III-5
Content-related validity	III-6
RELIABILITY	SECTION IV
TOEIC test reliability.....	IV-1
KR-20 Reliability Estimate.....	IV-1
<i>Table 1: Internal Consistency Estimates for TOEIC Scale Scores</i>	IV-2
Pass-Fail Consistency Index	IV-2
<i>Table 2: Pass-Fail (P-F_j) Estimates at Different TOEIC Scale Scores</i>	IV-2
Restriction of Range	IV-3
<i>Figure 1: TOEIC Reading Pre-test and Post-test Comparisons</i>	IV-4
Differences in TOEIC test scores across administrations.....	IV-4
Standard Error of Measurement (SEM).....	IV-4
Conditional Standard Errors of Measurement (CSEM)	IV-5
<i>Table 3: Conditional Standard Errors of Measurement at Different TOEIC Scale Scores</i> ..	IV-6
Standard Error of the Difference (SE _{diff})	IV-6

REFERENCES.....	SECTION V
APPENDIX A:.....	SECTION VI
Glossary of Statistical Terminology	VI-1
APPENDIX B.....	SECTION VII
Local validity studies.....	VII-1
<i>Table 4: Sample Local Validity Study Evidence</i>	VII-2
Smoothing	VII-3
<i>Table 5: Calculating Cutoff Values</i>	VII-4
<i>Figure 3: Percent Qualified--Smoothed vs. Unsmoothed Data</i>	VII-5
Cutoff scores.....	VII-5
<i>Figure 4: TOEIC Cutoff Scores--Smoothed vs. Unsmoothed Data</i>	VII-7
APPENDIX C.....	SECTION VIII
Statistical characteristics of TOEIC test forms.....	VIII-1

I. Technical Summary

The TOEIC® (Test of English for International Communication) test is an English language proficiency test for people whose native language is not English. It measures the English communication skills of people working in an international environment. The test does not require specialized knowledge or vocabulary beyond that of a person who uses English in everyday work activities.

The test is most often taken by employees of multinational corporations or government agencies and by students at English language schools. Companies use the TOEIC test for hiring, job assignment, promotion, and placement into training programs; schools use it to assist students with job placement and to assess improvement in English language proficiency. Stringent measures are taken to ensure that the TOEIC test is culturally unbiased and that items are relevant to test candidates across the world.

The TOEIC test is a paper-and-pencil test that consists of two 100-item multiple-choice sections: Listening Comprehension and Reading Comprehension. Separate scale scores are reported for each section and range from 5 to 495, in increments of 5 points. The Total TOEIC scale score is computed by summing the scale scores from the Listening Comprehension and Reading Comprehension sections. Total scores can range from 10 to 990, in intervals of 5 points. The two sections of the test are subdivided as follows:

<u>Listening Comprehension</u>			
Part	Section	# Items	# Choices
Part I	One Picture	20 items	4
Part II	Question-Response	30 items	3
Part III	Short conversation	30 items	4
Part IV	Short talks	20 items	4
	Total	100 items	
<u>Reading Comprehension</u>			
Part	Section	# Items	# Choices
Part V	Incomplete sentences	40 items	4
Part VI	Error recognition	20 items	4
Part VII	Reading comprehension	40 items	4
	Total	100 items	

TOEIC Total scores should be evaluated in terms of their separate Listening Comprehension and Reading Comprehension scale scores, as some individuals may have identical Total scores, yet have different section scores.

Each TOEIC test form is equated back to two older test forms. Some editions contain sections with scales which do not go up to 495. This is particularly true of the Reading Comprehension section. The lowest Reading Comprehension scale score to date has been 450 (Form 3CIC1, May, 1980). The TOEIC test developers conscientiously attempt to develop each test form so that the Reading Comprehension scale reaches 495 when possible.

Validity

Validity of the TOEIC test has been demonstrated through numerous studies linking TOEIC test scores to performance on other, established measures of English proficiency. Data from some of the studies comparing TOEIC section scores to direct measures of English language skills are shown below.

		Direct Measures			
		Speaking ^a	Listening ^b	Reading ^b	Writing ^c
TOEIC Scores	Listening Comp.	.74	.90	-	-
	Reading Comp.	.68	-	.79	.83
	Total	.74	-	-	-

* All correlations are significant at $p \leq .05$; a. $N=393$, b. $N=99$, c. $N=306$.

Based on data from a diverse sample of 628,789 candidates who took the TOEIC test in 1996 (TOEIC Report on Test-Takers Worldwide), the correlation between the Listening Comprehension and Reading Comprehension sections has been estimated at 0.81. This is consistent with earlier research done in Japan (e.g., Woodford, 1982).

Reliability

Reliability for the TOEIC test is reported as an internal consistency estimate, the KR-20 reliability coefficient (Kuder & Richardson, 1937). The internal consistency estimates for the Japanese secure administration (Woodford, 1982, p. 66) have been as follows:

KR-20 Reliability Coefficients for the TOEIC Test

Listening Comprehension	0.92
Reading Comprehension	0.93
Total Test	0.96

Standard Error of Measurement

The standard error of measurement (SEM) for both the Listening Comprehension and Reading Comprehension sections of the TOEIC test is approximately the same, at ± 25 scale points.

Standard Error of the Difference

The standard error of the difference (SE_{diff}) for both the Listening Comprehension and Reading Comprehension sections of the TOEIC test is approximately the same, at ± 35 scale points.

II. Introduction

The TOEIC® (Test of English for International Communication) test is an English language proficiency test for people whose native language is not English. It is developed and published by The Chauncey Group International, a subsidiary of Educational Testing Service. The TOEIC test measures the everyday English skills of people working in an international environment. Test scores indicate how well people can communicate in English with others in the global workplace. The test does not require specialized knowledge or vocabulary beyond that of a person who uses English in everyday work activities.

The TOEIC test is most often taken by employees of multinational corporations or government agencies and by students at English language schools. Companies use the TOEIC test for hiring, job assignment, promotion, and placement into training programs; schools use it to assist students with job placement and to assess English language proficiency.

Content of the TOEIC test

The TOEIC test was developed to meet the needs of the working world. The test questions are developed from samples of spoken and written language collected from various countries around the world where English is used in the workplace. Test questions incorporate many different settings and situations, such as:

- **General business**—contracts, negotiations, marketing, sales, business planning, conferences
 - **Entertainment**—cinema, theatre, music, art, media
 - **Manufacturing**—plant management, assembly lines, quality control
 - **Finance and budgeting**—banking, investments, taxes, accounting, billing
 - **Dining out**—business and informal lunches, banquets, receptions, restaurant reservations
 - **Corporate development**—research, product development
 - **Offices**—board meetings, committees, letters, memoranda, telephone, fax and e-mail messages, office equipment and furniture, office procedures
 - **Personnel**—recruiting, hiring, retiring, salaries, promotions, job applications and advertisements
 - **Purchasing**—shopping, ordering supplies, shipping, invoices
 - **Technical areas**—electronics, technology, computers, laboratories and related equipment, technical specifications
 - **Travel**—trains, airplanes, taxis, buses, ships, ferries, tickets, schedules, station and airport announcements, car rentals, hotels, reservations, delays and cancellations
 - **Health**—medical insurance, visiting doctors, dentists, clinics, hospitals
 - **Housing/corporate property**—construction, specifications, buying and renting, electric and gas services
-

These settings provide the background for TOEIC test questions—test-takers are not required to know specialized business and technical vocabulary. The TOEIC test is suitable for use in all work environments where English is used by native speakers of other languages.

Every effort is made to ensure that the test is unbiased and culturally relevant to our many test-takers worldwide. The TOEIC test development team is very careful to:

- Avoid language that is specific to U.S. English (e.g., vocabulary, grammatical constructions, idioms)
- Avoid contexts that may be specific to one culture, or that may be foreign to test-takers from some cultures
- Ensure the balanced use of names from different nationalities
- Avoid the use of locations, people, or events that would be known in only certain regions or countries
- Avoid situations that are too specific to one occupational area
- Ensure that different cultures are adequately represented

The finished test undergoes a stringent “fairness review,” in collaboration with outside reviewers, to be certain that all items are appropriate for use on a global basis.

Format of the TOEIC test

The TOEIC test is a two-hour, multiple-choice test that consists of 200 questions divided into two separately timed sections: Section I tests listening comprehension, and Section II tests reading comprehension.

Section I: Listening Comprehension

The first section, Listening Comprehension, consists of 100 questions, subdivided into four parts (Parts I-IV), which are administered by audiocassette over the course of approximately 40 minutes:

- Part I (One Picture) consists of 20 four-choice questions. Candidates view a photograph and listen to four spoken statements in English, after which they select the statement that best describes the photograph.
 - Part II (Question-Response) consists of 30 three-choice questions. Candidates hear a question spoken in English followed by three statements, also spoken in English, and then select the statement that best responds to the question.
-

- Part III (Short Conversations) consists of 30 questions, each based on a short conversation between two people. Candidates listen to each conversation and respond to a written four-choice question about the conversation.
- Part IV (Short Talks) consists of several short, spoken monologues followed by two or more written four-choice questions.

Section II: Reading Comprehension

The second TOEIC section, Reading Comprehension, consists of 100 questions, subdivided into three parts (Part V-VII), which are administered in a total of 75 minutes:

- Part V (Incomplete Sentences) consists of 40 items, each of which is a sentence containing a blank space where a word is missing. The candidate's task is to select one of four choices to best complete the sentence.
- Part VI (Error Recognition) consists of 20 sentences that have four words or phrases underlined, one of which creates an error. The task of the candidate is to select the incorrect portion of the sentence.
- Part VII (Reading Comprehension) consists of 40 questions based on documents typically found in work-related situations (e.g., announcements, advertisements, directions, notices, schedules, signs, etc.). Several four-choice questions are based on each reading selection.

Reporting of test scores

Candidates receive three scores when they take the TOEIC test: (1) a Listening Comprehension scale score, (2) a Reading Comprehension scale score, and (3) a Total test scale score. The scores on each TOEIC section (one scale score for Listening Comprehension and one for Reading Comprehension) are reported by converting the number of questions answered correctly (raw scores) to the scores which are reported (scale scores). Scale scores range from 5 to a maximum of 495 in increments of five points (i.e., scores are reported as 5, 10, 15, . . . , 490, 495); there is no passing or failing score on the TOEIC test. The Total TOEIC scale score is computed by summing the TOEIC scale scores from the two sections and can range from 10 to 990 in intervals of five points (i.e., 10, 15, 20, 25, . . . , 985, 990).

Correlation between the Listening Comprehension and Reading Comprehension sections

The *correlation* between the Listening Comprehension section and the Reading Comprehension section is approximately 0.81, signifying that the two sections exhibit a strong relationship to each other, yet still provide somewhat different information about candidates. This correlation is based on a sample of 682,789 TOEIC candidates from around the world who took the test in 1996 (for more information, see *TOEIC Report on Test-Takers Worldwide*, 1996). This correlation is consistent with earlier research done in Japan (e.g. Woodford, 1982). Some individuals with identical TOEIC Total scores may be better in Listening Comprehension than in Reading Comprehension, while for others the opposite may be true, and for still other candidates, there may be no difference between Listening Comprehension and Reading Comprehension. For this reason, TOEIC Total scores should be evaluated in light of the separate Listening Comprehension and Reading Comprehension scale scores provided.

Scale scores

TOEIC scores are reported as *scale scores* so that scores for all candidates are reported in the same reporting units, or scale values, regardless of which test form is taken. The TOEIC program uses many different test forms, or versions of the TOEIC test, which tend to vary somewhat in overall difficulty level (see Appendix C for statistical characteristics of all TOEIC test forms). Because all TOEIC scores are reported on the same scale, the level of performance needed to obtain a TOEIC Listening Comprehension score of 450, for example, requires equal proficiency in Listening Comprehension regardless of the test form taken. The same is true for a TOEIC Reading Comprehension score value of, say, 395.

Equating of TOEIC test forms

The process by which scaled scores are made equal, sometimes referred to as the *equating* process, makes adjustments to raw scores from test forms of different difficulty levels. In this way, the same level of performance in Listening Comprehension (or Reading Comprehension) is reported as the same scale score regardless of the specific TOEIC test form used.

This process assures that each version of the test is as equal in difficulty level as possible to any other version. Otherwise, a person's score on any particular test form would pertain only to that test form and would not be a valid indicator of their English

ability as compared to other people who took other forms of the test. Equating determines the number of items a candidate needs to answer correctly on any particular test form in order to acquire a particular scale score on that test form. For example, a candidate taking one version of the test may need to answer 88 items correctly to receive a scale score of 450, while a candidate taking another version of the test may need to answer only 85 items correctly to receive the same scale score. This indicates that the first candidate's test form was easier than the second candidate's test form.

Each TOEIC test form is equated back to two older TOEIC test forms by incorporating a block of items from each old form in the new test form. The equaters for each TOEIC test form are chosen by test developers based upon item reliability (r -biserials and delta values) and upon the test specifications. A series of computations are used to equate the test forms. The equating computations are applied to a hypothetical sample, known as the "equalized group." The equalized group contains two Japanese subgroups from the first secure administration of new test forms, those candidates who are affiliated with a company (Affiliates) and those who are not affiliated with a company (Non-Affiliates). The relative number of Affiliates and Non-Affiliates changes with each administration, hence, the groups must be equalized in order to properly equate. TOEIC test scores are reported on a scale which was instituted on the first TOEIC test administration, Form 3BIC, in December, 1979.

Score ranges

TOEIC Total scale scores are the sum of the scale scores achieved on the Listening Comprehension and Reading Comprehension sections. Not every TOEIC form is of equal difficulty, hence, some forms contain sections with scales which do not go up to 495. This is especially true of the Reading Comprehension section. For TOEIC test editions in which there is an insufficient number of difficult Reading Comprehension items, high level candidates will be unable to fully demonstrate their Reading Comprehension proficiency. In such cases, the Reading Comprehension scores for highly proficient candidates are said to be indeterminate in the measurement of reading. This failure to measure very proficient candidates in Reading Comprehension occurs for less than 3/10 of 1% of all TOEIC candidates worldwide, but might occur more frequently in countries where the general level of English proficiency is higher.

If the TOEIC user believes that a candidate is very proficient in English reading comprehension and that a certain TOEIC test form may not adequately measure that candidate's level of reading comprehension, the user should contact their TOEIC program representative for a recommendation of a specific TOEIC test form for use in that particular situation.

Speededness

Speededness is the extent to which candidates have sufficient time to respond to items on a test. It is reported as a speededness index ranging from 0.0 to 1.0. The speededness index is determined by taking the ratio of the variance of the Not Reached distribution to the variance of the scores. If the ratio is 0.15 or less for a certain section of the test, that section is often considered to be unspeeded for the group tested.¹ Another criterion that can be used to judge whether a test is speeded is the percentage of candidates that complete 75 percent of the test items. If the majority of candidates complete at least this percentage of the test, the test is generally considered to be unspeeded.

The TOEIC test is considered to be unspeeded. A quick glance at Appendix C shows that the speededness index for the Reading Comprehension section has never exceeded 0.1. Speededness measures do not apply to the Listening Comprehension section, as this section is administered and paced by audiocassette. All candidates reach the last item of this section.

¹ *ETS Standards for Quality and Fairness*. (1987). Educational Testing Service: Princeton, NJ. Reprinted by permission of Educational Testing Service, the copyright owner.

III. TOEIC test validity

Validity has to do with score use and the interpretation and meaning given to scores. Validity is not a property of the test itself, but has to do with how the scores are used. Test use is validated by the accumulation of evidence that supports a particular use within a certain context.

Construct-related validity

The most common form of test validation is correlation with other, established methods that purport to measure the same construct. If two tests are reported to measure the same ability, then it stands to reason that the scores on these two measures will be highly related to each other. In the case of the TOEIC test, a positive correlation should be obtained between TOEIC scores and scores on other measures of English language proficiency.

The TOEIC test measures two different constructs: listening comprehension and reading comprehension. Each of these should be differentially related to direct measures of speaking, listening, writing, and reading.

It also stands to reason that scores on the TOEIC test should be related to other indicators of English proficiency. Two such possible indicators are class level and performance in an English course. Those students placed at higher levels and those earning higher grades should also receive higher TOEIC scores.

i) Relationship between the TOEIC test and direct speaking measures

The Language Proficiency Interview (LPI) is a well-established direct assessment of oral language proficiency, developed by the Foreign Service Institute of the U.S. Department of State. A criterion-referenced test, the LPI yields a rating which corresponds to a description of actual oral language behavior. While the abilities measured by the TOEIC test (Listening and Reading Comprehension) and by the LPI (active speaking ability) are not completely parallel, there is a likely connection between the ability to understand spoken English and the more complex ability to understand spoken English and then to function in English in response.

Analyses of the relationship between the LPI and TOEIC test were conducted with a total of 393 candidates from France (N=56), Mexico (N=42), Saudi Arabia (N=10) and Japan (N=285). These data were collected in 1987 and 1988 by TOEIC staff in conjunction with corporate clients in these countries. The intercorrelations, or degree of relationship, between the tests are shown in Table 1 below. Performance on the Language Proficiency Interview was strongly and consistently related to specific TOEIC performance. In general, (1) TOEIC Total and Listening Comprehension scores were most closely related to LPI scores; and (2) TOEIC Listening Comprehension was more highly correlated with LPI ratings than was TOEIC Reading Comprehension. There were no significant differences in the strength of these relationships across candidates from the different countries. For this reason, subsequent validation work was done with only Japanese candidates.

Correlations with LPI

Listening Comprehension	.74
Reading Comprehension	.68
Total score	.74

Note: $N = 393$. All correlations are significant, $p \leq .01$.

Several other direct measures of spoken English have also been compared with the TOEIC test. In 1997, Eggly, Musial and Smulowitz² examined the relationship between the SPEAK test and performance on the TOEIC test for a group of 20 medical residents representing twelve different native languages. They found a strong correlation of .78 between the two measures.

In Australia, the Australian Second Language Proficiency Rating (ASLPR) is widely used to measure English proficiency, and includes a rating for spoken English. A study of 38 ESL students studying English in an intensive English program in Australia found that ASLPR scores and TOEIC Listening Comprehension scores correlated quite strongly ($r = .70$). This correlation was significantly stronger than that between ASLPR scores and TOEIC Reading Comprehension.

A second study, this time using the John Test, Part II, with 66 ESL students in Canada found a similar correlation of .69. Once again, this relationship was stronger than that between the John Test and TOEIC Reading Comprehension scores. The John Test is an oral proficiency test used for placement in English-as-a-second language (ESL) courses. It measures speaking and listening skills by asking examinees to look at a series of seven pictures depicting a day in the life of "John" and to either talk about the pictures or answer questions regarding them. Part II tests "connected discourse" by asking the examinees to retell the story of John's day in the past tense.

Although these direct measures require a response in spoken English, while the TOEIC test requires a candidate to answer questions printed in English in the test booklet, all of these tests and the TOEIC Listening Comprehension section measure the underlying ability to comprehend spoken English. The fact that correlations proved to be consistently high between these tests and TOEIC Listening Comprehension strongly suggests that all of these tests are, in fact, effectively measuring the common ability to understand and use spoken English.

ii) Relationship between the TOEIC test and direct listening measures

In one of the original studies to investigate the relationship between TOEIC scores and direct listening measures, a sample of 99 Japanese TOEIC candidates took a listening test that involved listening and responding to 25 taped English stimuli consisting of 15 short statements or questions and ten dialogues. For each of the 25 exercises there were three questions to be answered by the candidates. The questions were asked in Japanese by a Japanese examiner, and the candidates were encouraged to answer in Japanese. The two

² See Eggly, S., Musial, J., & Smulowitz, J. (1997). The relationship between English language proficiency and success as a medical resident. The ESP Journal.

listening measures correlated very highly ($r = .90$), indicating that the Listening Section of the TOEIC test is indeed an accurate indicator of a candidate's ability to comprehend spoken English.

Other standardized tests of listening comprehension can also be examined. Because these tests purport to measure the same construct as the Listening Comprehension section of the TOEIC test, we would expect to find strong correlations between each of these tests and TOEIC Listening Comprehension scores. The table below summarizes the results of several separate studies that examined the relationship between standardized tests of listening comprehension and the TOEIC Listening Comprehension section. In addition, one non-standardized listening measure is included in the table. The results clearly show that the TOEIC test is strongly related to other measures of listening comprehension. In all cases, the relationship between these measures and the TOEIC Listening Comprehension section was stronger than that with the TOEIC Reading Comprehension section, providing further evidence of the validity of the TOEIC Listening Comprehension test as a direct measure of listening skills.

Test	Correlation*	Sample	Year
ASLPR, Listening	.73	38 ESL students in Australia	1999
TOEFL Listening Comprehension	.84	116 business school students in France	1996
TOEFL Listening Comprehension	.88	103 ESL students in Canada and USA	1999
In-house Listening placement test	.92	26 ESL students in USA	1998
CASAS Listening Comprehension	.85	31 students in California Community Colleges	1998
Michigan Listening Comprehension Test	.76	185 ESL students in Canada and USA	1999
Canadian Language Benchmarks Assessment, Listening	.67	30 ESL students in Canada	1998

* All correlations are significant at the 0.01 level (two-tailed).

iii) Relationship between the TOEIC test and direct reading measures

The same sample of 99 Japanese TOEIC candidates also took a direct reading test in which a series of different reading tasks were presented to the candidates. Candidates were provided ample time to read each selection and were then asked questions, in Japanese, about the content of the selection. The candidates answered the questions, orally, in Japanese. A strong correlation ($r = .79$) was observed between the direct reading measure and the Reading Comprehension section of the TOEIC test. This suggests that the TOEIC Reading section provides a good indication of the candidate's ability to read English with understanding.

Later research with three standardized tests of Reading Comprehension also demonstrated that the TOEIC test provides an accurate measure of an examinee's comprehension of written English. The CASAS test is currently used for placement and admission purposes in the California Community College system, the TOEFL test measures English used in academic programs in North American colleges and universities, and the Canadian Language Benchmarks Assessment provides an indication of the English proficiency of newcomers to Canada. Each test has a Reading Comprehension section. The results of four studies involving these tests are shown in the table below.

Test	Correlation*	Sample	Year
CASAS Reading Comprehension	.73	111 students in California Community Colleges	1998
TOEFL Reading Comprehension	.76	116 business school students in France	1996
TOEFL Reading Comprehension	.83	103 ESL students in Canada and USA	1999
CLBA Reading Comprehension	.87	120 ESL students in Canada	1998

* All correlations are significant at the 0.01 level (two-tailed).

iv) Relationship between the TOEIC test and a direct writing measure ($r = .83$)

The direct writing measure consisted of three parts: dehydrated sentences, writing a business letter, and writing the translation of ten Japanese sentences. Three hundred six Japanese candidates participated in the direct writing measures. The direct writing measures correlated strongly ($r = .83$) with the TOEIC Reading Comprehension score.

This high correlation suggests that the TOEIC reading score is a good indication of the candidate's ability to write in English.

v) Relationship between the TOEIC test and class levels

During 1998 and 1999, we conducted a large-scale validation study in Canada and the United States. The primary purpose of this study was to support the use of the TOEIC test as a placement and evaluation tool for language programs in community colleges, four-year institutions, and intensive English programs.

As part of that study, we examined the relationship between TOEIC scores and class levels to determine how accurate the test was as a placement instrument. In this study, no students were placed into levels based on their TOEIC scores. Rather, the TOEIC test was taken as an exit test or as an independent measure of proficiency. Therefore, a sizeable relationship between class levels and TOEIC scores would indicate that the TOEIC test could be used as an accurate placement instrument for students in English programs in North America.

We found that as levels increased, there was a corresponding increase in TOEIC Listening, Reading, and Total scores. The correlations between class levels and TOEIC scores, shown below, were moderate and were similar to those found between class levels and placement tests commonly used in language programs in North America. This is a clear indication that TOEIC test scores reflect the same type of English language proficiency that is assessed by the placement instruments used at the schools participating in this study. The full results of this study can be found in '*A Comparison of Standardized ESL Tests*'.

Correlation with class level*	
Listening Comprehension	.61
Reading Comprehension	.59
Total	.63

N = 1109. * All correlations significant at the 0.01 level (two-tailed).

vi) Relationship between TOEIC scores and classroom performance

In another part of the validation study, we examined the relationship between TOEIC scores and student grades. The comparison of test scores with classroom grades was conducted within class level, rather than across levels, because a student with a specific grade in a higher level class would be expected to perform better on a standardized measure of English proficiency than would a student with the same grade in a lower level class.

As classroom grades increased, TOEIC Listening Comprehension, Reading Comprehension, and Total test scores also tended to increase. The strongest relationships between grades and TOEIC scores, and those most consistent with expectations, were found at the intermediate level and above. At these levels, TOEIC Listening Comprehension scores were most strongly correlated with classroom grades for Listening and also exhibited a moderate correlation with Speaking grades. In general, these correlations were stronger than those between TOEIC Listening Comprehension and grades for other subjects, providing further evidence of the validity of the TOEIC Listening Comprehension section as a direct measure of listening skills and an indirect measure of speaking skills.

The relationship between TOEIC Reading Comprehension scores and classroom grades was somewhat less uniform. At the intermediate and advanced level, TOEIC Reading

Comprehension scores were most highly correlated with grades for grammar, reading, and writing so that as grades increased within each class level, so too did Reading Comprehension scores. All three of these areas are related in that they require the same basic skills and so should also be related to TOEIC Reading Comprehension scores. However, at other levels the grades for listening and/or speaking were also correlated with TOEIC Reading Comprehension scores. Grades in these two areas are partly dependent on a student's grammar and reading skills, particularly at higher class levels, and so would likely be related to scores on a standardized test of reading comprehension. The full results of this study can be found in '*A Comparison of Standardized ESL Tests*'.

	Grade	Listening	Reading	Total		Grade	Listening	Reading	Total
Advanced Beginner	Grammar	.03	.33*	.18	Advanced Intermediate	Grammar	.48	.38	.45
	Listening	.19	.40*	.31*		Listening	.70*	.59*	.74*
	Reading	.10	.32*	.21		Reading	.49*	.37*	.50*
	Speaking	.20	.32*	.28*		Speaking	.42*	.05	.29
	Writing	.17	.29*	.24*		Writing	.26	.25	.29
Intermediate	Grammar	.35*	.57*	.52*	Advanced	Grammar	.11	.74*	.57
	Listening	.40*	.21	.36*		Listening	.75*	.63*	.72*
	Reading	.28*	.51*	.45*		Reading	.66*	.82*	.78*
	Speaking	.35*	.48*	.47*		Speaking	.68*	.69*	.73*
	Writing	.26*	.54*	.45*		Writing	.50*	.65*	.61*

* $p < .01$

Content-related validity

Major evidence for the content-related validity of the TOEIC test comes from a series of needs analyses conducted by Educational Testing Service® (e.g., Tannenbaum & Rosenfeld, 1995; Woodford, 1982), during which many international companies were asked about the English language skills needed by employees who use English for international communication. These needs analysis studies identified certain aspects of English usage that are commonly required in many different countries by multinational companies. TOEIC test specifications are designed to measure performance in terms of these requirements, which are now reflected in the kinds of test questions, sections, and subsections included in the TOEIC test, as well as in the context and setting of test questions.

Since validity has to do with the interpretations given to TOEIC scores, many companies and local users conduct local validity studies and collect their own evidence to support the TOEIC test's usefulness or validity for their particular purpose. Some studies link TOEIC scores to speaking measures, such as the Listening Proficiency Interview (LPI), while other local studies demonstrate that candidates considered to be competent English users obtain higher TOEIC scale scores than candidates considered not to be as competent in their English language proficiency (Educational Development Center, 1992). See Appendix B for information on how to conduct a local validity study.

IV. TOEIC test reliability

Reliability is an indicator of the extent to which test scores will be consistent across different conditions of administration and/or administration of alternate forms of the test.³ Such conditions can include factors relating to the candidate, the kind of test given, situational factors external to the test, and/or the way in which a particular test is scored. Reliability is reported as an index ranging from .00 (complete absence of reliability) to 1.00 (perfect reliability). The closer this index is to 1.00, the closer the test scores can be said to be free from errors of measurement.

Errors of measurement occur when a candidate performs differently on one occasion or test form than on another for reasons that may or may not be related to the purpose of the test. A person may try harder, be more (or less) tired or anxious compared to some other occasion, have greater familiarity with the content of questions on one test form than on another test form, or simply guess more questions correctly on one occasion than on another. For these and other reasons, a person's score will not be perfectly consistent from one occasion to the next or from one test form to the next form. These types of reasons for inconsistency are referred to as errors of measurement.

KR-20 Reliability Estimate

Reliability of the TOEIC test is reported as an internal consistency measure (r_{it}) using the KR-20 reliability coefficient (Kuder & Richardson, 1937). The KR-20 reliability index assesses the consistency of candidate responses to all of the items in each section and measures homogeneity of the test content. In the past, this internal consistency estimate for the total test score (content consistency) for the secure Japanese sample has been approximately 0.95 for the Total score and between 0.91 and 0.93 for the Listening Comprehension and Reading Comprehension sections, respectively (see Table 2).⁴ Note that an internal consistency reliability coefficient is estimated based on two elements: (1) the specific test form administered and (2) the population tested with that test form.

³ *ETS Standards for Quality and Fairness*. (1987). Educational Testing Service: Princeton, NJ. Reprinted by permission of Educational Testing Service, the copyright owner.

⁴ The TOEIC test is administered to a standardization sample in Japan when the test is first given. It is also administered in Korea at the same time. The first administration is referred to in this technical manual as the *secure administration*.

Table 1
Internal Consistency Reliability Estimates for TOEIC Scale Scores

Test Name	Number of Items	Reliability (KR-20)
Listening Comprehension	100	0.91 to 0.93
Reading Comprehension	100	0.92 to 0.93
Total Score	200	0.95 to 0.96

The internal consistency score reliabilities for the seven parts of the TOEIC test have ranged from a low of 0.67 to a high of 0.87; these subpart scores are not reported to candidates because they are not sufficiently reliable for use in making decisions about candidates' English language abilities. In compliance with testing standards, scores with reliabilities over 0.90 are considered to be adequate for reporting and usage (cf. *Standards for Educational and Psychological Testing* 1985).

Pass-Fail Consistency Index

Another reliability estimate for clients who use the TOEIC test to make decisions at different TOEIC score points is a pass-fail consistency index, or Subkoviak index (Breyer & Lewis, 1994; Subkoviak, 1976). This index tells users the proportion of individuals who would be classified as at or above a specific TOEIC scale score if the candidates were retested with the same TOEIC form and all other factors were considered to remain the same. The size of the Subkoviak index is affected by the reliability of the test scores and the distance above (or below) the mean test score in a normal distribution. Listening Comprehension mean scale scores have ranged from 282 to 311, and Reading Comprehension mean scale scores have ranged from 254 to 270 for the Japanese population. Table 3 shows the possible different pass-fail consistency estimates ($P-F_T$) along different score points where the scale score mean (and standard deviation) were set to be equal to 300 (and 86) for Listening Comprehension and 260 (and 87) for Reading Comprehension. In both cases, the internal consistency reliability (r_{ii}) was set to 0.92.

These pass-fail consistency estimates will vary depending on the specific TOEIC form used and the group to whom the test was administered. For groups in which there is a great deal of homogeneity (for example, when candidates are pre-selected to meet some company criteria), reliability estimates will be lower.

The pass-fail reliability can also vary depending on where the cut score is in relation to the mean and standard deviation of the group tested. Table 3 shows that the closer the cut score is to the mean, the lower the pass-fail reliability.

Table 2
Pass-Fail Consistency ($P-F_T$) Estimates at Different TOEIC Scale Scores

$r_{tt} = 0.92$ Listening Comprehension		$r_{tt} = 0.92$ Reading Comprehension	
$Mean_{LC} = 300$ $SD_{LC} = 86$		$Mean_{RC} = 300$ $SD_{RC} = 86$	
Scale	P-F _r	Scale	P-F _r
450	.97	450	.99
400	.94	400	.97
375	.91	375	.95
350	.89	350	.93
325	.88	325	.90
300	.87	300	.88
275	.88	275	.87
250	.89	250	.87
200	.94	200	.90
150	.97	150	.94
100	.99	100	.98
50	.99	50	.99

Restriction of Range

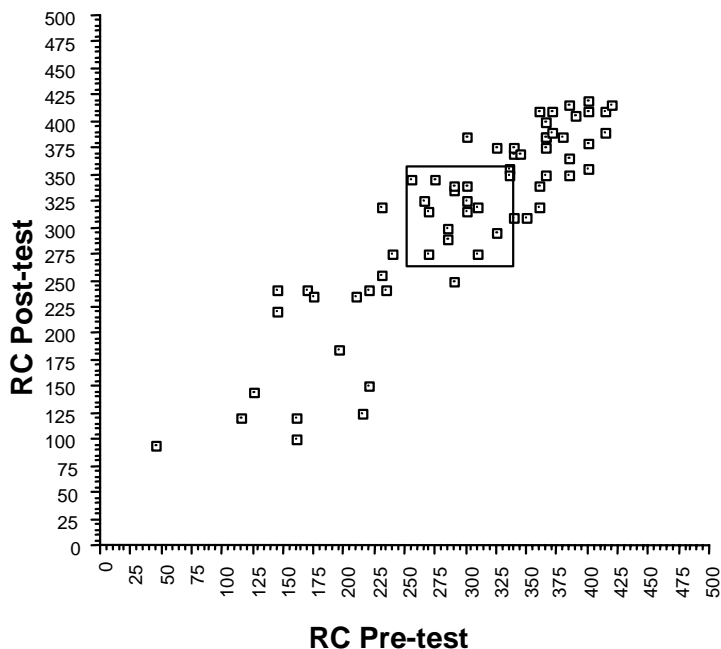
Restriction of range is a case in which the variance of scores in an analysis sample is lower than the variance of scores in the population from which the sample was selected.⁵ Test form reliability estimates are a function of variance, or spread of candidates' test performances in comparison to each other. If you have a sample of candidates who are very similar to each other, the reliability of the test within that specific homogeneous group will be quite low. Test reliability is technically defined as the proportion of true score variance out of Total score variance and is presented as a correlation ranging between .00 and 1.00. If there is no (or very little) variation among candidates' test scores then, by definition, there can be no accurate estimate of reliability.

Suppose in Figure 1 that a company gave the TOEIC Reading Comprehension test once as a pre-test and once as a post-test, using a different form each time and that the scores on the pre-test and the post-test were plotted. Suppose that another company did the same thing, but all the scores of the second company's employees fell within the square indicated in Figure 1. The first company would observe a rather high alternate form reliability, but the second company would not. In fact, the second company might incorrectly conclude that the

⁵ *ETS Standards for Quality and Fairness*. (1987). Educational Testing Service: Princeton, NJ. Reprinted by permission of Educational Testing Service, the copyright owner.

was not reliable. If the second company were to test a wider sample of employees, they might observe that the TOEIC test was as reliable as found by the first company.

Figure 1
TOEIC Reading: Pre-test and Post-test Comparisons



Differences in TOEIC test scores across administrations

If a second form of the TOEIC test were administered close to the first TOEIC form, a candidate would probably obtain slightly different scores on the different form. No test measures performance with perfect accuracy or consistency. If several different forms of the test were taken close together in time, a candidate would probably obtain a number of different scores that cluster around a typical, or average, value. Some scores would be higher and some would be lower than the average score. The difference between the average test score and the actual test score on a particular form of the test is called the *standard error of measurement*.

Standard Error of Measurement (SEM)

The question of importance for an individual candidate is: “How much will my TOEIC Listening Comprehension or Reading Comprehension score fluctuate due to chance factors unrelated to the TOEIC test?” Typically, a candidate will add 1 standard error of measurement to his or her observed scale score to determine how much it might fluctuate

due to factors not related to the test content. In other words, how much better might that candidate have scored if he or she had, for example, paid more attention during the test?

The ***standard error of measurement (SEM)*** for the Listening Comprehension section scale score is about 25 scale score points (1 SEM=25 scale points) and is approximately the same for the Reading Comprehension section. This means that TOEIC scores can be expected to fluctuate within a band ± 25 scale points around the reported TOEIC Listening Comprehension or Reading Comprehension scores. This band captures the region in which a candidate's TOEIC Listening Comprehension ***true score*** or Reading Comprehension true score would most likely fall 67% of the time. You can increase this region to capture 95 percent of all the times a person might retake the TOEIC test -- with all else being considered equal -- by doubling the width of the band to ± 49 points above and below the observed scale score (1.96 x SEM). Thus, if a candidate received a Listening Comprehension or Reading Comprehension scale score of 325, the true score is most likely (with 95% confidence) to be between 276 and 374 (or within approximately 100 scale score points) due to factors related to differences in tested content on each form. To expand that even further to 99% of the cases, a score band of ± 63.5 points would be used (2.54 x SEM). This variation in test performance is due to the fact that different TOEIC forms use different test items, some which might be more or less difficult to the candidate.

A 95 percent confidence band of 100 TOEIC scale points might seem like a wide margin for error, but the user is reminded that TOEIC scores are reported in 5-point increments. A 95 percent confidence band therefore covers only 20 possible scores.

Conditional Standard Errors of Measurement (CSEM)

Standard errors of measurement will differ depending on the scores to which they are connected or on which they are conditioned. Errors of measurement connected to or conditioned on specific scores are referred to as ***conditional standard errors of measurement (CSEM)***. The CSEM at various TOEIC scale scores is provided in Table 4, based upon a method discussed by Lord (1984, Method IV). For TOEIC test clients who use cutoff scores, it is important to know how much fluctuation may occur at specific cutoff scores due to factors unrelated to English language proficiency. Note that the CSEMs presented in Table 4 vary by about ± 5 scale score points at the top of the TOEIC scale, about ± 2 scale score points in the middle of the scale, and about ± 10 scale points at the bottom of the scale. These CSEMs are from one specific TOEIC form (Form 3PIC2); other TOEIC forms would have slightly different conditional standard errors of measurement but would follow the same general pattern.

Table 3
Conditional Standard Errors of Measurement at Different TOEIC Scale Scores
 (TOEIC Form 3PIC2, May, 1993)

Listening Comprehension		Reading Comprehension	
<i>Scale</i>	<i>CSEM</i>	<i>Scale</i>	<i>CSEM</i>
490	17	480	2
450	21	450	12
400	25	400	19
350	27	350	23
300	28	300	25
250	29	250	26
200	28	200	26
150	26	150	26
50	19	50	21
5	13	5	16

Standard Error of the Difference (SE_{diff})

This question is different: “How big of a difference do I need to get between two Listening scores or between two Reading scores before I can say I have a real difference?” Here the questioner is asking about two independent tests, given either at two different times or given to two different people. The questioner may be describing a situation in which a test was given, followed by a training program and then a follow-up test, or the questioner may need to place students into levels based on their TOEIC scores. The questioner wants to know how big the difference would have to be in Listening Comprehension or Reading Comprehension scores before it could be determined that the training had some positive effect or that the scores of the two people were actually different. This difference between scores is the ***standard error of the difference*** (SE_{diff}). The standard error of the difference for both the Listening Comprehension and Reading Comprehension sections is approximately 35 scale points.

If a person begins training with a Listening score of 300 and, following training, scores 340 on a different test form, has that candidate really improved in Listening? To determine whether this is a real increase in the TOEIC test score the candidate would construct a band of $\pm 1 SE_{diff}$, or ± 35 points, around the obtained score. In this case, the candidate has truly improved because the post-training score falls outside the SE_{diff} (i.e., 265-335). If the candidate wishes to be more conservative in determining whether his or her score has increased, he or she should construct a band of ± 69 points around the TOEIC score ($1.96 \times SE_{diff}$). Scores that fall outside this band would show a true improvement 95% of the time. Using this band, we can say with 95% confidence that the candidate has not improved (i.e., 340 falls within the 231-369 band).

If fluctuation of test scores is the point of the question, then the SEM should be used, not the standard error of the difference. All TOEIC users should calculate score bands around candidate observed scores using the SEM.

If the question is meant to be about (independent) score differences, then the SE_{diff} should be used. The SE_{diff} can also be helpful in choosing between two individuals on the same test or in telling the difference between one person's listening and reading scores.

The SE_{diff} is useful when comparing two candidates with the intent to select the higher-scoring candidate because of a *different* test score. For example, unless the Listening Comprehension score of the first person is at least one SE_{diff} higher than the Listening Comprehension score of the second person, there is no basis (in terms of TOEIC scores) for selecting the higher-scoring person over the lower-scoring person because the two scores are not really *different*; the observed score differences are likely due to chance factors. This comparison could be made by constructing a score band of $\pm 1 SE_{diff}$ (to be 68% confident) or of $\pm 1.96 SE_{diff}$ (to be 95% confident) around one of the scores. If the score of the 2nd person falls within this band then there is no real difference between those two individuals in terms of Listening Comprehension as measured by TOEIC. However, if the score of the second person falls outside this band, then we can say with 68% (or 95%) accuracy that the second person is better in Listening Comprehension than the first person.

V. References

- American Psychological Association. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.
- Breyer, F. J., & Lewis, C. (1994). Pass-fail reliability for tests with cut scores: A simplified method. Research Reports (94-39). Princeton, NJ: Educational Testing Service.
- Educational Development Center. (1992). Final Report: English Language Survey of Unocal Offshore Personnel. Unpublished Manuscript. Bangkok, Thailand: Unocal Thailand, Ltd.
- Educational Testing Service. (1987). ETS Standards for Quality and Fairness. Princeton, NJ: Educational Testing Service.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.
- Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.
- Lord, F. M. (1984). Standard errors of measurement at different ability levels. Journal of Educational Measurement, 21, 239-243.
- Subkoviak, M. J. (1976). Estimating the reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276.
- Wilson, K. M. (1989). Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experience in the TOEIC test context. TOEIC Research Report, No. 1. Princeton, NJ: Educational Testing Service.
- Woodford, P. E. (1982). The Test Of English for International Communication (TOEIC). In C. Moffit (Ed.), English for International Communication. New York: Pergamon Press.

Appendix A

Glossary of Statistical Terminology

VI. Glossary of Statistical Terminology

Alternate Form, or Form: An edition of a test that is written to meet the same specifications and is comparable in most respects to another edition of the test except that some or all of the questions are different.*

Analysis Sample: The group of people on whose performance a statistic or set of statistics has been calculated.*

Conversion Parameters: Quantitative rules for expressing scores on one test form in terms of scores on an alternate form. See *Alternate Form, Equating*.*

Correlation: A statistical index that expresses the degree of correspondence, or relationship, between two sets of scores. The values of a *correlation coefficient* can range from perfect agreement (+1.00) to no agreement between the two scores (0.00) to perfect disagreement (-1.00). Most correlation coefficients for test scores fall between 0.00 and +1.00.

Conditional Standard Error of Measurement (CSEM): Errors of measurement connected to or conditioned on specific scores. See *Standard Errors of Measurement*.

Delta: A normalized transformation of the P+ value to an interval-based difficulty scale that ranges from 1 (indicating extremely easy) to 25 (indicating extremely difficult). See *Percent Pass (P+)*.

Equated Deltas: Delta values which are transformed to put the average item difficulties on the same scale regardless of the English proficiency of the test-taking population.

Equating: A statistical process used to convert scores on two or more alternate forms of a test to a common scale so that the scores may be used interchangeably.*

Item: A test question.*

KR-20 Reliability Formula: The Kuder-Richardson formulas are used for estimating the reliability of a test from statistical data on the individual items in the test (Formula #20), or from the mean score, the standard deviation, and the number of items in the test (Formula #21). The Kuder-Richardson formula #20 is generally used for estimating test reliability.

Parameter: (1) The value of some variable for a population as distinguished from an estimate of the value based on a sample drawn from the population. (2) In item response theory, one of the characteristics of an item, such as its difficulty. See also *Conversion Parameter*.*

* Terms marked with an asterisk have been reprinted from the *ETS Standards for Quality and Fairness* (1987) by permission of Educational Testing Service, the copyright owner.

* Terms marked with an asterisk have been reprinted from the *ETS Standards for Quality and Fairness* (1987) by permission of Educational Testing Service, the copyright owner.

Percent pass (P+): The percentage of candidates answering an item correctly.

Mean: The sum of a set of scores divided by the number of scores; the average of the scores.

Median: The middle score in a distribution of scores, or 50th percentile. Half of a group of scores fall above the median and half fall below it.

Raw Score: The number of items answered correctly on a test.

R-biserial: Also known as *r* or *r bis*. A correlation coefficient relating performance on a test question and performance on the total test. It is an index of discrimination measuring the extent to which candidates who score high on the total test tend to get the question correct and those who score low tend to get the question incorrect.

Reliability: An indicator of the extent to which test scores will be consistent across different conditions of administration and/or administration of alternate forms of the test. See also *Test-Retest Reliability*.*

Restricion of Range: A case in which the variance of the scores in an analysis sample is lower than the variance of scores in the population from which the sample was selected. See *Analysis Sample, Variance*.*

Scale Scores: Also referred to as *Scaled Scores*. The scores on each TOEIC test section which are reported. They are obtained by converting the number of questions answered correctly, *raw scores*, using conversion parameters. Scale scores range from a minimum of 5 to a maximum of 495 in increments of five points. Candidates receive three scale scores when they take the TOEIC test: a Listening Comprehension scale score, a Reading Comprehension scale score, and a Total test scale score. See *Parameter* and *Conversion Parameter*.

Secure Administration: The first administration of a test form. The scale for any test form is based upon the results of its secure administration. Secure administrations for the TOEIC test are conducted in Japan and/or Korea.

Skewness: The tendency of a distribution to depart from symmetry or balance around the mean. If the scores tend to cluster at the lower end of the distribution, the distribution is said to be positively skewed; if they tend to cluster at the upper end of the distribution, the distribution is said to be negatively skewed.

Smoothing: A statistical procedure often used when data does not increase or decrease steadily where it would be expected or desired. Data is adjusted to bring it closer to what would have been observed if the sample was larger.

Speededness: The extent to which test takers have time to respond to items on a test. One indicator of speededness is the percent of test takers who answer all of the items in the test. See *Speededness Index*.*

Speededness Index: Determined by taking the ratio of the variance of the Not Reached distribution to the variance of the scores. If the ratio is 0.15 or less for a certain section of the test, that section is often considered to be unspeeeded for the group tested.

Standard Deviation: A statistic characterizing the magnitude of the differences among a set of measurements. Specifically, it is the square root of the average squared difference between each measurement and the mean of the measurements. See *Variance*. The standard deviation is the square root of the variance.*

Standard Error of Measurement (SEM): The difference between the average test score and the actual test score on any particular form of the test. Standard errors of measurement will differ depending on the scores to which they are connected or on which they are conditioned (also known as CSEM, or conditional standard error of measurement).

Standard Error of the Difference (SE_{diff}): The difference between two scores affected by the SEM of each of the scores. The SE_{diff} between two scores is larger than the SEM of any one score. The formula for the SE_{diff} between two scores is:

$$SE_{diff} = \sqrt{SEM_{LC1}^2 + SEM_{LC2}^2}$$

which can be simplified to $SE_{diff} = 1.414 \times SEM$.

Subkoviak index: A single administration estimate that represents the proportion of candidates who can be classified as having similar ability level on two administrations of a test. The magnitude of the Subkoviak index is affected by the reliability of the test scores and the distance of the cutoff score from the mean test score in a unimodal distribution. The Subkoviak index ranges from a low of 0.50 to a high of 1.00. More inconsistent classifications will result the closer the cutoff score is to the mean.

Test-Retest Reliability: An estimate of reliability based on the correlation between scores on two administrations of the same test form to the same group of people.*

True Score: The average score which would result if the TOEIC test were taken an infinite number of times by a single person and all factors were kept equal and no measurement errors were present at the time of testing or scoring.

* Terms marked with an asterisk have been reprinted from the *ETS Standards for Quality and Fairness* (1987) by permission of Educational Testing Service, the copyright owner.

* Terms marked with an asterisk have been reprinted from the *ETS Standards for Quality and Fairness* (1987) by permission of Educational Testing Service, the copyright owner.

Variance: A statistic characterizing the magnitude of the differences among a set of measurements. Specifically, it is the average squared difference between each measurement and the mean of the measurements.*

Validity: The measure of how well the test results fit the original purpose of the test. Validity is not a property of the test itself, but rather, pertains to how the obtained scores are used. Validity refers to the extent to which a test measures that which it was designed to measure.

Appendix B

Conducting a Local Validity Study

VII. Local validity studies

There are typically three steps involved in conducting local validity studies. First, two groups of people are identified, in which the members of one group are considered to be “qualified” in their English proficiency and the members of the other group are considered to be “not qualified.” Next, the TOEIC test is administered in a group setting to both groups. Finally, the scores of the members of each group are listed in separate columns (see Table 4) and the proportion of persons considered qualified is calculated for each score level.

To initiate the first step in this local validity process, independent judgements need to be made about each candidate’s English proficiency in order to appropriately assign each candidate to either the “qualified” or “not-qualified” group. This step is very important and needs to be accomplished first without any knowledge of TOEIC scores. A supervisor or an English language instructor who has had an opportunity to observe the person’s English language capabilities can make this judgement.

In making the judgement, it is important to establish a reasonable standard for the level of English proficiency individuals must demonstrate in order to be classified as “qualified.” This standard is typically a written statement that describes the minimal level of English that is required within the context in which it is to be used. The specification of this context will provide the basis for connecting the English proficiency needed with TOEIC test scores. Local TOEIC representatives can help define this standard of minimal English proficiency for TOEIC clients through survey methodologies and can provide development tools for this purpose. Once this minimal standard of English proficiency is defined locally, individual candidates can be classified as either meeting, or exceeding, the minimal English proficiency standard (Qualified), or not meeting the minimal English proficiency standard (Not Qualified). Finally, an appropriate TOEIC score can be determined to reflect this distinction.

If the proportion of people considered qualified in English proficiency by independent local judges increases as TOEIC section scores increase, the data provide evidence that TOEIC is useful in categorizing candidates’ English proficiency in the local setting.

An example of local validity study evidence is provided in Table 4 and is discussed below. Note that the data are not based on any real company study nor are they based on actual validity evidence. The information presented in Table 4 is provided only to show how such evidence can be presented and analyzed. Notice that Table 4 contains only Listening Comprehension score data; a similar table would be required for the Reading Comprehension section of the TOEIC test.

Table 4
Sample Local Validity Study Evidence

Listening Scores	Number of Candidates			Percent Qualified
	Qualified	Not Qualified	Total	
475-495	5	0	5	100
450-470	3	1	4	75
425-445	6	2	8	75
400-420	18	1	19	95
375-395	17	3	20	85
350-370	15	10	25	60
325-345	20	9	29	69
300-320	7	8	15	47
275-295	6	17	23	26
250-270	2	9	11	18
225-245	6	8	14	43
200-220	2	4	6	33
175-195	2	12	14	14
150-170	0	7	7	0
125-145	0	3	3	0
100-120	0	1	1	0
5-95	0	0	0	0

Other formal methods exist for determining the appropriate TOEIC score that is best for a specific use or situation (see, for example, Livingston & Zieky, 1982). The method described above is meant to serve only as an example of one professionally acceptable way to determine an appropriate TOEIC score to suit the local need. The TOEIC representatives can best advise clients on how to conduct such a study. In addition, the Chauncey Group has a staff that is experienced in consultative services on an international basis and is available to provide further advice on validation or other technical aspects of the TOEIC test. **Users should not arbitrarily pick a TOEIC score as a cutoff.** A formal process for the selection of a TOEIC cutoff score is highly recommended.

Smoothing

When the percentage of candidates at each score level or score band who are “qualified” is computed, the percentage may not increase steadily from one level to the next, as might be expected or desired. Notice how the sample numbers provided in the *Percent Qualified* column of Table 4 do not increase steadily as the TOEIC scale increases. This may be due to the small number of cases at each score level. One way to handle the bouncing-around effect observed in the *Percent Qualified* column is to group TOEIC scores together in broader intervals than are seen here. This has the effect of adjusting the percentages to bring them closer to what would have been observed if more candidates had been assessed. The general term used for adjustments in such a situation is smoothing.

Table 5 provides an example of how to smooth validity evidence from such a study using the example data provided from Table 4. In Table 5, the observed percentage qualified at each score level was replaced with the average of the percentages from that level and the two adjacent score levels. This method is also known as moving averages. It is calculated by summing the number qualified at a specific score level with the number qualified in the level above and below it (see the *Qualified* column in Table 5). This sum is then divided by the total number of candidates in the three levels. The calculations based on these sample numbers are shown in Table 5 in the *Smoothed Percent Qualified* column. Notice that the moving average cannot be computed at the very lowest score interval nor can it be computed at the very highest score interval because, for the lowest score, there is no *Percent Qualified* score below it and, for the highest score, no score above it.

In addition to calculating the *Smoothed Percent Qualified* (as shown in the fourth column of Table 5), these data can be presented graphically for visual inspection. Using a simple graphing computer program, the user should first calculate the midpoint for each TOEIC score interval. For example, in Table 5, the midpoint for the interval “250-270” is computed by simply summing 250 with 270 and dividing that interval by 2 for a midpoint score of 260. These TOEIC midpoint scores can then be graphed against the smoothed percent qualified. The smoothed line in Figure 3 -- the *Smoothed Percent Qualified by TOEIC Score Interval Midpoint* -- can be used as evidence of the local validity for the TOEIC test in that higher TOEIC scores are obtained by more candidates considered to be qualified in their English proficiency (i.e., as TOEIC scores go up, a greater percentage of candidates are considered qualified). Notice that if a TOEIC user decides not to group scores into intervals, as in Table 5, but, instead, prefers to keep them separate, no midpoint calculations are necessary.

Table 5
Calculating Cutoff Values

Listening Scores	Number of Candidates		"Smoothed" Percent Qualified
	Qualified	Total	
475-495	5	5	---
450-470	3	4	$\frac{5+3+6}{5+4+8} = 82\%$
425-445	6	8	$\frac{3+6+18}{4+8+19} = 87\%$
400-420	18	19	$\frac{6+18+17}{8+19+20} = 87\%$
375-395	17	20	$\frac{18+17+15}{19+20+25} = 78\%$
350-370	15	25	$\frac{17+15+20}{20+25+29} = 70\%$
325-345	20	29	$\frac{15+20+7}{25+29+15} = 61\%$
300-320	7	15	$\frac{20+7+6}{29+15+23} = 49\%$
275-295	6	23	$\frac{7+6+2}{15+23+11} = 31\%$
250-270	2	11	$\frac{6+2+6}{23+11+14} = 29\%$
225-245	6	14	$\frac{2+6+2}{11+14+6} = 32\%$
200-220	2	6	$\frac{6+2+2}{14+6+14} = 25\%$
175-195	2	14	$\frac{2+2+0}{6+14+7} = 15\%$
150-170	0	7	$\frac{2+0+0}{14+7+3} = 8\%$
125-145	0	3	$\frac{0+0+0}{7+3+1} = 0\%$
100-120	0	1	$\frac{0+0+0}{3+1+0} = 0\%$
5-95	0	0	---

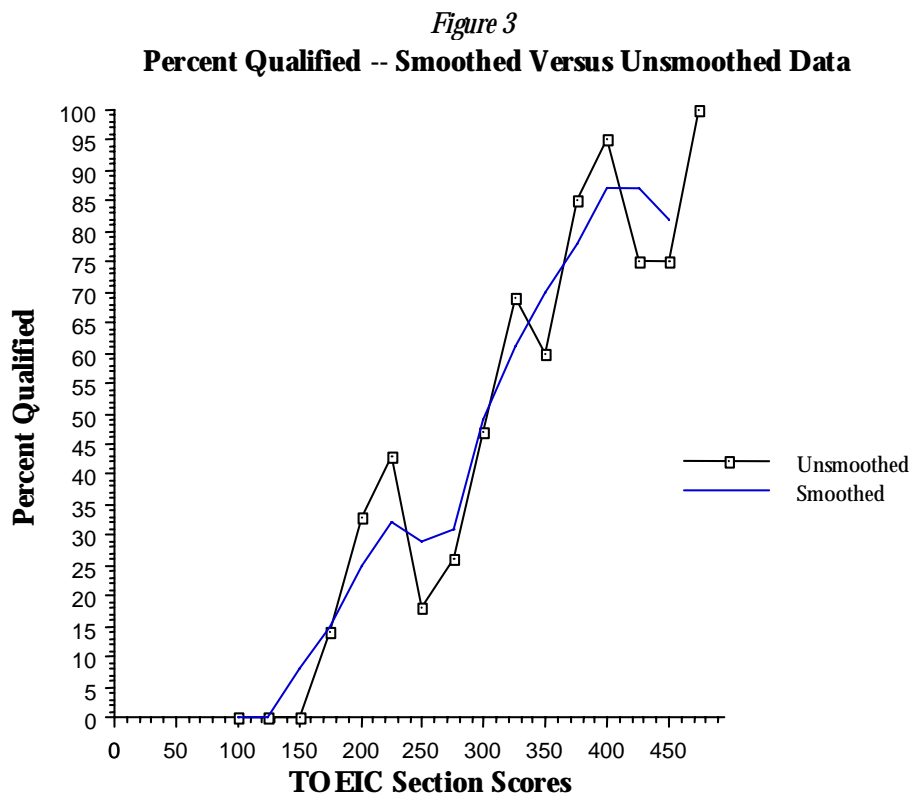


Figure 3 shows how the percentages provided in Tables 4 and 5 compare to each other when presented graphically. Note that many graphic programs available for the personal computer provide both smoothing and graphing capabilities; users who do not wish to calculate these numbers by hand may not have to do so if a graphing or charting program with a smoothing feature is available.

Cutoff scores

Data from local validity studies can be used to help determine a cutoff score for local use with TOEIC Listening Comprehension and Reading Comprehension sections. This section describes how to determine the appropriate TOEIC cutoff point for a specific use, such as the minimum level of English proficiency required for a work assignment. Typically, the score at which 50 percent of the candidates are said to be "qualified" and 50 percent are said to be "not qualified" can be used as the dividing line between "qualified" and "not qualified" in TOEIC English proficiency. At any lower TOEIC test score level, a candidate is more likely to be judged unqualified than qualified (i.e., the percentage of qualified candidates is less than 50 percent), while the reverse is true at any higher test score level. Of course, a TOEIC user (i.e., a company) could decide to select a different cutoff score than the TOEIC score that occurs at the 50 percent qualified-unqualified level and might want to assess the effect of such a cutoff score for a set period of time before making a commitment to a single TOEIC cutoff score level.

The rationale for setting the passing score at that TOEIC test score which corresponds to a 50 percent chance of being judged as "qualified" is based on the assumption that the two

types of possible wrong decisions about a candidate are equally serious. These two possible wrong decisions can be best expressed as follows: first, a test user could consider someone qualified who should not be selected as “qualified”; second, a candidate could be rejected as “not qualified” who should have been selected as “qualified”. Choosing that TOEIC score which is at the 50 percent level of those candidates judged “qualified” assumes that the two types of wrong decisions are equally important. Sometimes this may not be true and this is an issue that TOEIC users must decide for themselves. But, here is one way (of many) to face that issue.

In practice, it is easy to determine the percentage of people qualified by asking decision makers the following question:

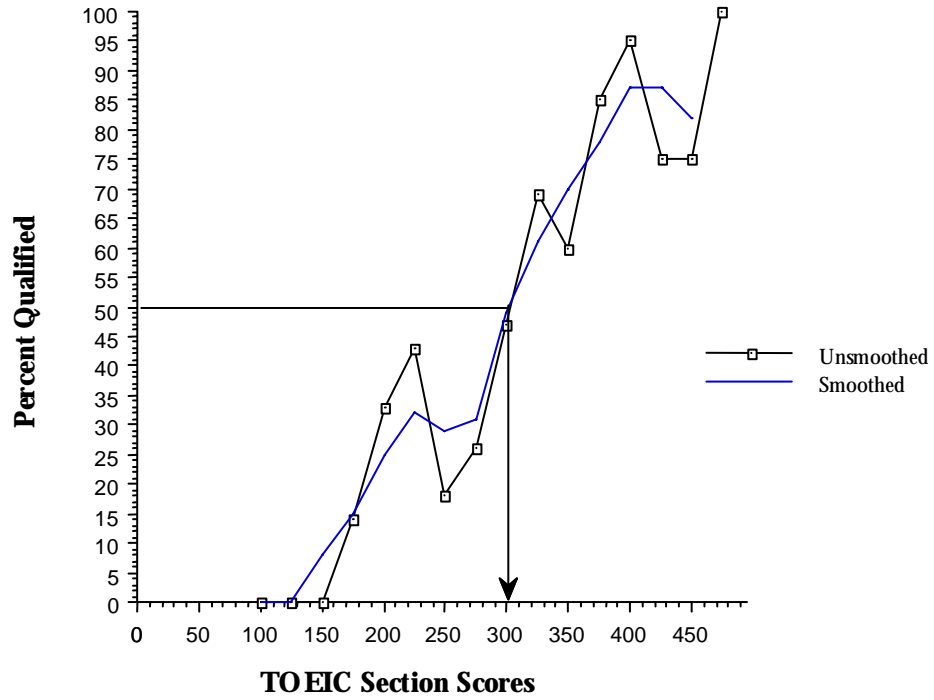
Suppose a company had a group of 100 people and you knew that 50 were qualified and 50 were not qualified in their English proficiency. If you had to pass all 100 or fail all 100, which would you do?

If the answer is “pass them,” then the percent-qualified decision point is 50 percent. If the answer to the above question is “fail them,” then ask the same question for a group of 60 qualified persons and 40 unqualified persons. If the answer is to pass them, then ask the same question for a group of 55 qualified persons and 45 unqualified persons. Keep adjusting the percent qualified in this way until you cannot decide whether to pass or to fail the group of 100 people. The score that corresponds to this percent qualified will be the percent-qualified decision point for your company.

Figure 4 shows how a corresponding TOEIC scale cutoff score may be determined graphically, provided the 50 percent qualified to not-qualified decision point is used. This is accomplished by first drawing a straight line from the 50 percent number on the vertical axis of the graph over to the line of the smoothed percent qualified data and then, from this intersection point, dropping another straight line down to the horizontal axis with the TOEIC scores on it. The TOEIC cutoff score would be represented by that score with the arrow in Figure 4. Note again that the data represented in Figure 4 are hypothetical and do not represent any real study. They are provided only as an example that TOEIC score users may wish to follow.

A description of the TOEIC Benchmarking Analysis (local validity study) is available in a brochure published by TOEIC Service International, entitled ‘*TOEIC Benchmarking Analysis*’.

Figure 4
Smoothed Versus Unsmoothed Data



Appendix C

Statistical History of TOEIC Test Forms

VIII. Statistical characteristics of TOEIC test forms

Appendix C presents information describing the internal statistical characteristics for most TOEIC test forms as of the publication of this Technical Manual. All TOEIC test forms are developed according to the same content and statistical specifications. Each TOEIC form, however, differs slightly in terms of its statistical characteristics, given the population of candidates taking the examination. This statistical information is based on the secure administration of the examination in Japan.⁶ Thus, the statistical characteristics of the test will vary slightly for specific groups, and it is good practice to collect local normative information. The staff of The Chauncey Group is willing and available, for a consultation and service fee, to help set up monitoring programs, providing tables similar to those presented in Appendix C, for a company's test data.

A concise summary of the statistical characteristics of the TOEIC forms administered in Japan is presented in this Appendix. The TOEIC test was originally scaled in Japan and the statistical information presented here is based upon candidates who initially took the TOEIC examination in Japan under secure testing conditions. Other areas of the world will show different statistical characteristics, dependent on the general level of English proficiency of the population. For this reason, it is important for companies to gather their own data. The data from the TOEIC test-taking population in Japan is relevant only within the Japanese population-specific context. Please note that, as of the secure administration of test form 3UIC7 (April 1998), statistics are run on first-time testers only, rather than on the total group of testers, as was previously the practice. First-time testers provide more unbiased results, as they have never previously been exposed to the TOEIC test or to any of the test items.

Raw score information for the Japanese and Korean total groups (first-time groups, starting with form 3UIC7) includes the following: the number of items in the test, the number of items scored for reporting, obtained minimum and maximum raw scores, and raw score mean, standard deviation, and median. Scale score information for the Japanese and Korean total groups (first-time groups, starting with form 3UIC7) includes scale score mean and standard deviation, minimum and maximum possible scores, as well as minimum and maximum obtained scores.

Item and test statistics listed at the bottom of the table are based on the Japanese test analysis sample for each form. Item statistics include the mean P+ (the average proportion of items correct), mean and standard deviation of the observed deltas (a normalized transformation of the P+ value to an interval-based difficulty scale that ranges from 1 [extremely easy] to 25 [extremely difficult or hard]), mean and standard

deviation of the equated deltas (equated deltas put the average item difficulties on the same scale regardless of the English proficiency of the test-taking population), and mean and standard deviation of the biserial correlations (biserial correlations show how test items on

⁶ Please note that the TOEIC test forms presented in Appendix C which are marked with an asterisk (*), such as 3RIC5 and 3TIC6 through 3TIC8, are secure forms which were administered only in Korea, rather than in Japan. A comparison of the data presented in the Appendix for the Korea and Japan secure administrations prove the Japanese and Korean populations to be slightly different in some ways, reflecting such population-specific characteristics as are mentioned above.

average can discriminate between those who are proficient in English and those who are not). Test statistics include reliability, standard errors of measurement for the raw score and scale score, special score data, and selected speededness data.

Raw score data and unadjusted (observed) item and test statistics depend on both the ability level of the group and the difficulty level of the test form and therefore cannot be directly compared across forms. Scale score data indicate ability level only (with differences in difficulty level among test forms having been taken into account) and are useful for comparing the performance of different persons or groups. Equated deltas indicate difficulty level only (with differences in ability level among groups having been taken into account) and are useful in comparing the difficulty levels of different forms of the test.
