

*Establishing the Validity of
TOEIC Bridge™ Test Scores
for Students in Colombia,
Chile, and Ecuador*

Sandip Sinharay

Ying Feng

Luis Saldivia

Donald E. Powers

Anthony Ginuta

Annabelle Simpson

Vincent Weng

November 2008

ETS RR-08-58



**Establishing the Validity of the TOEIC Bridge™ Test Scores
for Students in Columbia, Chile, and Ecuador**

Sandip Sinharay
ETS, Princeton, NJ

Ying Feng
CTB, McGraw Hill, New York

Luis Saldivia, Donald E. Powers, Anthony Ginuta,
Annabelle Simpson, and Vincent Weng
ETS, Princeton, NJ

November 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Copyright © 2008 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, GRE, LISTENING. LEARNING. LEADING., and TOEIC are registered trademarks of Educational Testing Service (ETS).

TEST OF ENGLISH FOR INTERNATIONAL COMMUNICATIONS and TOEIC BRIDGE are trademarks of ETS.



Abstract

The validity of TOEIC Bridge™ scores as a measure of English language skill was examined from the standpoint of a unified concept of test validity. In this study, more than 6,000 test takers in 3 Latin American countries (Chile, Colombia, and Ecuador) took 1 form of the TOEIC Bridge test, and their scores were compared to additional information about the students (teacher judgments, self-assessments, and performance on other academic achievement measures). The evidence collected was generally quite consistent with the interpretation of TOEIC Bridge scores as indicators of the English language competencies for the students examined.

Key words: English language skills, test validity, convergent and discriminant evidence

Acknowledgments

The authors wish to thank Gayle Bravo for her help with the data set; Dan Eignor, Shelby Haberman, Brent Bridgeman, John Young, and Rick Morgan for their helpful comments; and to Mark Syp for the editorial help.

Table of Contents

	Page
Test Administrations.....	2
Summary Statistics.....	3
Correlations and Factor Analysis.....	4
Comparison of TOEIC Bridge Scores to Scores on Other Tests of English Proficiency	6
Colombia.....	6
Ecuador	11
Chile.....	12
Comparing TOEIC Bridge Scores to the Self-Assessment Reports of Students	12
Ecuador and Chile.....	15
Colombia.....	18
Discussion of the Self-Assessment Results	18
Comparing TOEIC Bridge Scores to Teacher Ratings.....	18
Discussion/Implications.....	21
References.....	24
Notes	26
Appendixes	
A – Summary Statistics for the Combined Population in the Different Parts of the Test	27
B – Listening Comprehension Can-Do Questionnaire	28
C – Reading Can-Do Questionnaire	30

List of Tables

	Page
Table 1. Summary Statistics of the Scores.....	3
Table 2. Summary Statistics of the Colombian Examinee Scores in the Different Parts of the Test.....	4
Table 3. Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Ecuador.....	11
Table 4. Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Vocational Institute 1	13
Table 5. Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Vocational Institute 2	14
Table 6. Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Chile Ministry	14
Table 7. Summary of the Correlations Between the Can-Do Scores and the TOEIC Bridge Scores for Ecuador and Chile	16
Table 8. Means (and Standard Deviations) for Teacher Ratings of Can-Do Tasks and Correlations With TOEIC Bridge Scores	20
Table 9. Correlations among Student Self Ratings, Teacher Ratings, and TOEIC Bridge Scores for the Examinees in Chile.....	21

List of Figures

	Page
Figure 1. Comparison of the Colombian State Examination scores and TOEIC Bridge test scores for about 10% of the Colombian students who took the TOEIC Bridge test.....	8
Figure 2. Comparison of average school-level Colombian State Examination scores and average school-level TOEIC Bridge scores.....	9
Figure 3. Comparison of average school-level Colombian State Examination (CSE) scores and average school-level TOEIC Bridge (TB) scores.	10
Figure 4. Average TOEIC Bridge Listening scores and corresponding 95% confidence intervals for the different answer choices to the statement, “Using English, I can understand headline news broadcasts on the radio,” for Chile Ministry.	16

The TOEIC Bridge™ test was developed by ETS to measure emerging English-language competencies (for more information, see <http://www.ets.org/toEICbridge>). Designed to assess the skills of beginning and intermediate learners of English, the test is administered to more than 180,000 examinees a year in several countries. Because the test is meant to help teachers and students assess, practice, and improve listening and reading comprehension of English, it features questions that are easier to answer than those in the Test of English for International Communications™ (TOEIC®). The TOEIC Bridge test also provides somewhat detailed feedback to examinees on their test performance.

The TOEIC Bridge test is composed of two sections: listening comprehension and reading. Each section is separately timed and has 50 multiple-choice questions. The sections are set up as follows:

Section I. Listening Comprehension	25 minutes (approximately)
Part 1. Photographs	15 items (four-choice)
Part 2. Question-Response	20 items (three-choice)
Part 3. Short Conversations/Talks	15 items (four-choice)
Section II. Reading	35 minutes
Part 4. Incomplete Sentences	30 items (four-choice)
Part 5. Reading Comprehension	20 items (four-choice)

Evidence that the TOEIC Bridge measures English language proficiency currently comes from two sources: (a) the very high correlation of TOEIC Bridge scores (high .80s) with other local tests of English language proficiency, and with the longer, more difficult, TOEIC test and (b) the careful design and assembly of the test by language experts to include a range of important English language tasks. A few validity studies, however, have been conducted for the TOEIC Bridge test (ETS, 2007).

One such study (Tannenbaum & Wylie, 2008) found the TOEIC Bridge test to be appropriate for English language learners at levels that are generally equivalent to basic user levels A1 and A2 for the Common European Framework (CEF; e.g., Tannenbaum & Wylie, 2005). Some test takers at the top of the score scale also fall into the independent user level B1.¹

This report details a study that aims to provide evidence of the validity of TOEIC Bridge scores as a measure of English language proficiency for South American examinees. In this

study, the validity of the TOEIC Bridge test is examined from a unified concept of test validity. The focus in construct validation historically has been the evaluation of theoretically expected patterns of relationships among item scores and between test scores and other measures (Messick, 1995). Support for test meaning is a “mosaic of convergent and discriminant findings,”(Messick, 1995, p. 744) that evolves from several sources of evidence; this mosaic is one of six distinguishable *aspects* of construct validity. Messick (1995) cautioned that test validity cannot *rely* on any one form of evidence, nor does it *require* any one form, “*granted that there is defensible convergent and discriminant evidence supporting score meaning*” (p. 744; emphasis added). The primary focus of this report is on convergent and discriminant evidence from several sources: teacher judgments, students’ self-assessments, and student performance on other academic achievement measures.

Test Administrations

A form of TOEIC Bridge was administered to more than 6,000 examinees in three South American countries—Colombia, Ecuador, and Chile—at the request of the ministries of education in these countries. The test takers were students at educational institutions in Colombia and Chile (public and private high schools and vocational schools), and Ecuador (public and private high schools and a police academy). In Colombia, 1,958 11th-grade students in several cities took the test in April 2006. In Ecuador, 408 5th and 6th grade students took the test in November 2006. In Chile, students from four different institutions took the test: 2,304 students at Vocational Institute 1 took the test in November 2006; 205 students Vocational Institute 2 took the test in November 2007; 1,447 high-school students from different institutions in Chile, selected by the Ministry of Education (henceforth referred to as Ministry) took the test in November 2006; and 239 students at Private Institute took the test in November 2006.

In Colombia, a two-stage sampling procedure was used to sample examinees. First, 65 schools were chosen from 17 cities in Colombia according to a stratified random sampling procedure (the strata denoting different levels of performance on English in the Colombian State Examination in 2005). Then, within each school, 30 students were chosen using a simple random sampling procedure from all eleventh grade students.

In Ecuador, a two-stage sampling procedure was used to sample examinees. First, five institutions were chosen. Then, a number of classrooms were selected within each institution and all students from those classrooms were asked to participate in the study.

In Chile, all students participating in English courses at Vocational Institute 1 were invited to participate. At Vocational Institute 2, all students finishing their programs where English courses are required were invited to participate. The Ministry of Education selected the sample of high-school students, and this sample is assumed to be representative of Chile's high-school population. At Private Institute, all students finishing the college preparation program were invited to participate.

Summary Statistics

Table 1 lists the summary statistics for the scores for the different examinee groups. For each of the listening comprehension and reading sections, the table shows the mean score, standard deviation (SD) of scores, reliability (Cronbach's α), and average biserial correlation for each examinee group. In all the analysis henceforth, we will deal with the raw scores of the examinees in the two sections of TOEIC Bridge: listening comprehension and reading. The table shows that test performance differs substantially among the examinee groups. For example, the Colombian examinees score 20.2 on average in reading; compared to 37.2 for examinees at Private Institute. The test reliability, however, is quite high for all the examinee groups. The minimum reliability observed in Table 1 is 0.88 and the maximum is 0.92. The average biserial correlation (ranging between 0.48 and 0.66) is respectably high, so that the test items discriminate well among the examinees.

Table 1
Summary Statistics of the Scores

Examinee group	Listening comprehension				Reading			
	Mean	SD	Reliability	Av. bis.	Mean	SD	Reliability	Av. bis.
Colombia	22.1	9.0	0.88	0.51	20.2	8.9	0.88	0.50
Ecuador	32.6	9.8	0.91	0.57	28.4	9.8	0.90	0.54
Chile Vocational Institute 1	33.3	8.1	0.88	0.52	28.6	8.3	0.88	0.48
Chile Vocational Institute 2	35.0	10.9	0.92	0.66	31.1	11.4	0.91	0.59
Chile Ministry	23.8	9.8	0.88	0.55	20.8	10.3	0.89	0.51
Chile Private Institute	40.0	9.7	0.92	0.62	37.2	10.2	0.92	0.61

Note. Av. bis. = average biserial correlation of the test items.

Correlations and Factor Analysis

Study of the correlation structure within the test is often relevant in judging validity. High correlations between test parts that measure the same construct and low correlations between test parts that measure different constructs provide favorable validity evidence. The correlation coefficients (both with and without attenuation) and other summary measures for the scores in the five parts of the test were computed for each examinee group. These values are shown in Table 2 for the Colombian examinees.

Table 2

Summary Statistics of the Colombian Examinee Scores in the Different Parts of the Test

Type	Part	Listening			Reading	
		Photographs	Stimulus-response	Short cons/talks	Incomplete sentences	Reading comprehension
Mean		7.91	8.48	5.67	10.60	9.62
Reliability		0.73	0.74	0.68	0.80	0.79
Listening						
Correlation	Photographs	1.00	0.63	0.62	0.63	0.66
	Stimulus-response	0.86	1.00	0.69	0.71	0.65
	Short cons/talks	0.88	0.97	1.00	0.70	0.64
Reading						
Correlation	Incomplete sentences	0.82	0.92	0.94	1.00	0.68
	Reading comprehension	0.87	0.85	0.87	0.86	1.00

Note. The correlation coefficient between listening comprehension and reading scores is 0.94, after correcting for attenuation (or unreliability of the tests). The shaded entries in the table are the corresponding correlation coefficients after correcting for attenuation. The first three parts belong to listening comprehension and the last two to reading. The within-section correlations are marked in bold. Cons = conversations.

Most of the correlations exhibit patterns of higher within-section correlations and lower between-section correlations, which supports the interpretation of distinct measures of reading and listening. However, a within-section correlation is smaller than a corresponding between-section correlation after correction for attenuation,² in a few instances. For example, in Table 2, the corrected correlation of 0.86 between incomplete sentences and reading comprehension (two parts of the reading section) is lower than that between stimulus-response (a part of the listening comprehension section) and incomplete sentences (0.92), and that between short conversations/talks and incomplete sentences (0.94). The same patterns (of few within-section correlations being smaller than corresponding between-section correlations) were also observed for Ecuador and the four student populations in Chile. Appendix A, which shows a table with correlations for the combined sample, demonstrates a similar phenomenon. Messick (1989) recommended that such anomalies be followed up empirically, because they are frequently the source of new insights in science.

Factor analysis has long been held to be a powerful tool in test score validation (Messick, 1989). For example, the simultaneous emergence of separate factors corresponding to different constructs (in this context, that means separate listening comprehension and reading factors) provides discriminant evidence (Messick, 1989). Exploratory and confirmatory factor analyses were performed on the matrix of tetrachoric correlation coefficients among the 100 items, to study the factor structure of the TOEIC Bridge test scores for each of the six examinee groups described in Table 1. In exploratory factor analysis (EFA), we examined the eigenvalues of the tetrachoric correlation matrix using scree plots and parallel analysis, and used the varimax and promax rotation methods to search for interpretable factors. In confirmatory factor analysis (CFA), we used several goodness-of-fit measures such as the comparative fit index (CFI), the standardized root mean square residual (SRMR), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA).

For Vocational Institute 2, separate listening and reading factors were observed in the factor analysis. However, the TOEIC Bridge scores of all the other five examinee groups are primarily described by a general factor that is a combined listening and reading factor. For example, in EFA, the two-factor solution did not turn out to have listening and reading factors, and in CFA, a two-factor model with listening and reading factors did not provide better fit than the one-factor model (details of these analyses are available on request). This finding is in consonance with those of the

well-known study related to second language proficiency by Oller (1979), who proposed that a single general language proficiency factor underlay performance on a variety of language tests.

The correlation coefficients and the factor analysis results demonstrate that the listening comprehension and reading sections of the TOEIC Bridge test do not measure very distinct skills for the examinee groups that were studied. In other words, the listening and reading skills measured by the test do not seem well-separated for these beginning learners of English. That the English language skills of beginning learners are difficult to distinguish is not uncommon. As learners develop their proficiency in English, reading and listening skills may become more highly differentiated empirically. We recommend monitoring the correlational and factorial structure of the results of possible future administrations of the test to these examinees.

Comparison of TOEIC Bridge Scores to Scores on Other Tests of English Proficiency

We tried to collect scores of other tests from the examinees who took the TOEIC Bridge test. For the different countries, we received scores from different types of tests. Hence, we will discuss results for the different countries separately. No such scores could be obtained for the examinees at Private Institute.

Colombia

Scores were available on different content areas (such as English and mathematics) from the 2006 Colombian State Examination for 198 students who took the TOEIC Bridge test. The Colombian State Examination includes an English test that consists of 24 multiple-choice items (and is thus expected to be much less reliable than the TOEIC Bridge test; the reliability estimate itself was unavailable) and involves some reading comprehension, but no listening comprehension. These 198 students are from schools whose academic calendars are like those from schools in the United States. From a study of their previous Colombian State English Examination scores in English, these schools do not differ systematically in their English skills from those following another calendar, so these schools can be considered representatives of all the schools involved in our study. These 198 students, however, scored higher on average in TOEIC Bridge (with average scores of 29.7 in listening comprehension and 28.6 in reading) than the rest of the examinees.

Figure 1 compares the scores on English and mathematics³ in the Colombian State Examination versus the TOEIC Bridge test scores for the 198 students. The TOEIC Bridge

(listening comprehension in the two leftmost panels or reading in the two rightmost panels) scores are plotted along the x -axis and the scores on English (two top panels) or mathematics (two bottom panels) in the Colombian State Examination are plotted along the y -axis, in each panel. Each point in any plot denotes one examinee. The line in each plot, shown for convenience of viewing, is the regression line of the variable plotted along the y -axis on the variable plotted along the x -axis. The correlation coefficient for each pair of variables plotted is also shown in each panel. Both the TOEIC Bridge listening comprehension and reading scores have higher association with scores on English in the Colombian State Examination than with scores on mathematics in the Colombian State Examination. The TOEIC Bridge listening comprehension and reading scores correlate highly with scores on English in the Colombian State Examination (correlations of 0.88 and 0.89, respectively) and correlate moderately with scores on mathematics in the Colombian State Examination (correlations of 0.69 and 0.73, respectively) and with the other content areas in the Colombian State Examination (results not shown). These values point to the respectable external validity (Messick, 1989) of the TOEIC Bridge test scores. Ideally, the scores of a valid test should correlate highly with similar measures and correlate less with dissimilar measures. That is exactly what is observed for this test.

Average scores were available on several content areas of the Colombian State Examination in 2005 for all of the 65 schools whose students took the TOEIC Bridge test. Figure 2 shows the average 2005 Colombian State Examination scores (on English and mathematics) versus the average TOEIC Bridge test scores for these schools. The average school-level TOEIC Bridge (listening comprehension or reading) scores are plotted along the x -axis and the average school-level scores on English or mathematics in the Colombian State Examination are plotted along the y -axis, in each panel in the figure. Each point in each panel denotes one school. The line in each plot, shown for convenience of viewing, is the corresponding regression line. The correlation coefficient for each pair of variables plotted is also shown in each panel. Figure 3 shows a three-dimensional scatter-plot of the school-level scores on TOEIC Bridge (listening comprehension and reading) and school-level scores on English in the Colombian State Examination. The average school-level TOEIC Bridge listening comprehension scores are plotted along the x -axis, the average school-level TOEIC Bridge reading scores are plotted along the y -axis, and the average school-level scores on English in the Colombian State Examination are plotted along the z -axis. Each point in the plots denotes one school.

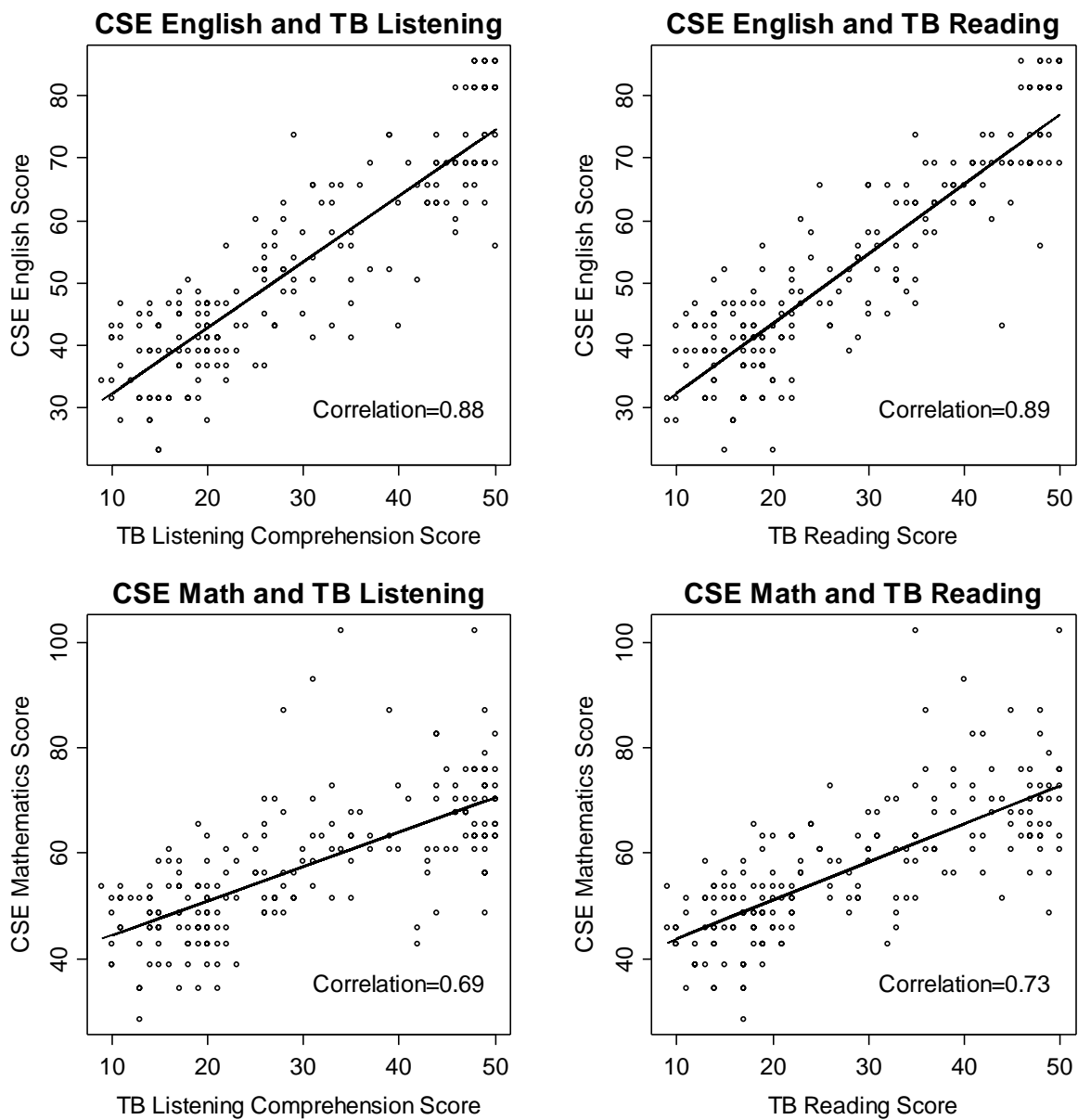


Figure 1. Comparison of the Colombian State Examination scores and TOEIC Bridge test scores for about 10% of the Colombian students who took the TOEIC Bridge test.

Note. CSE = Colombian State Examination, TB = TOEIC Bridge.

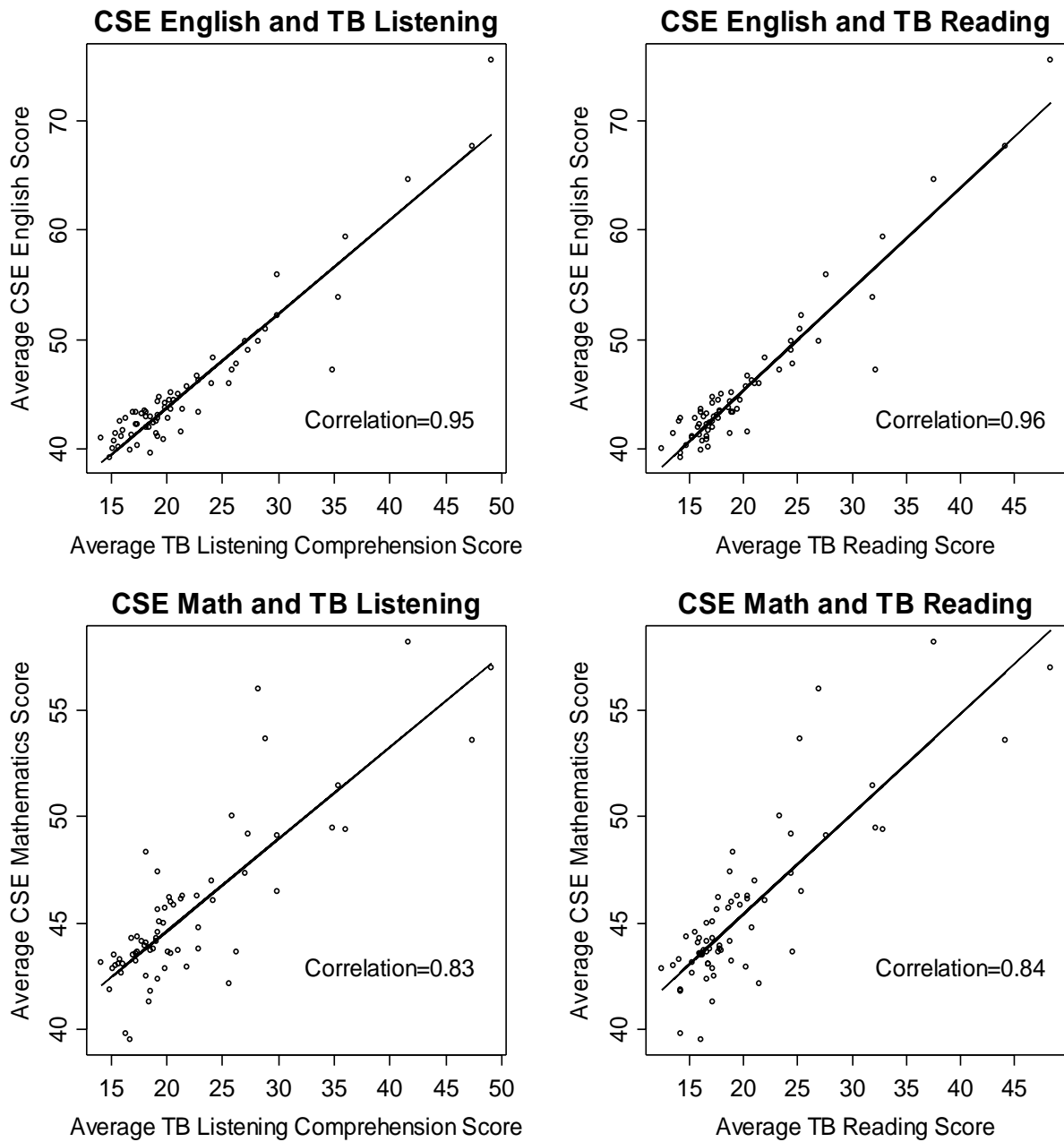


Figure 2. Comparison of average school-level Colombian State Examination scores and average school-level TOEIC Bridge scores.

Note. CSE = Colombian State Examination, TB = TOEIC Bridge.

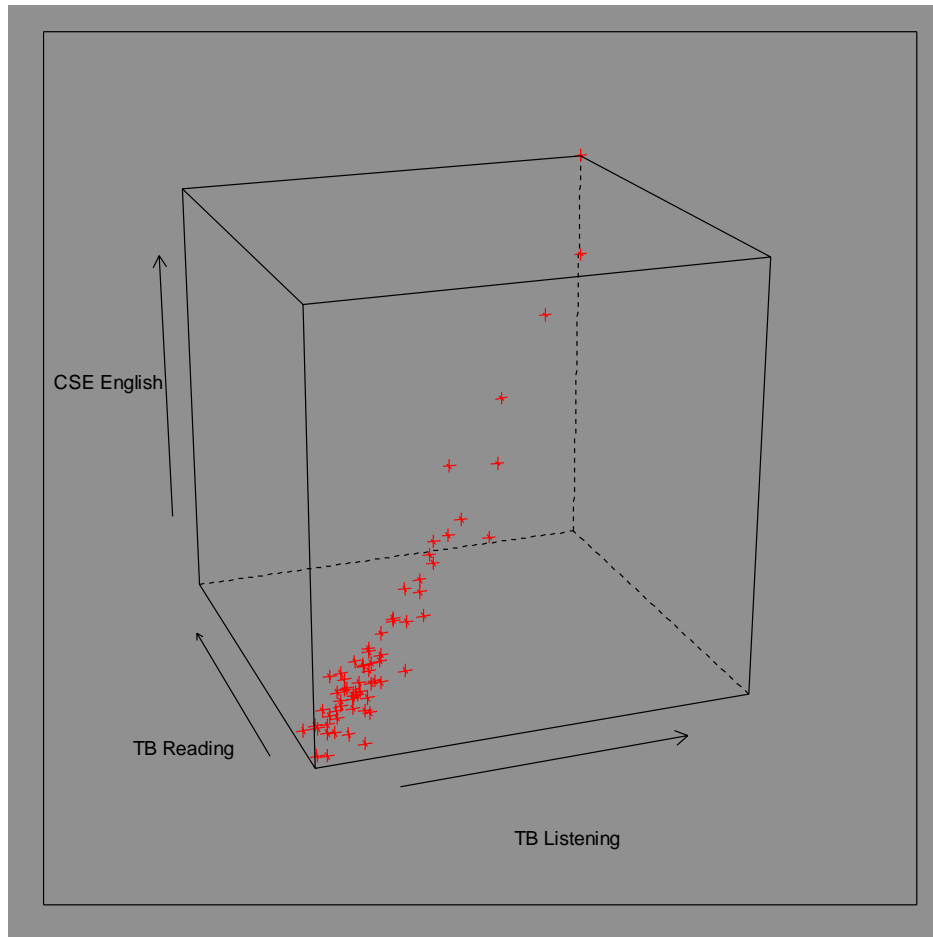


Figure 3. Comparison of average school-level Colombian State Examination (CSE) scores and average school-level TOEIC Bridge (TB) scores.

Figures 2 and 3 demonstrate that the TOEIC Bridge listening comprehension and reading average scores correlate very highly with those on English in the Colombian State Examination (correlations of 0.95 and 0.96, respectively) and correlate somewhat less highly with the average scores on mathematics in the Colombian State Examination (correlations of 0.83 and 0.84, respectively) and with other content areas in the Colombian State Examination (plots not shown). These values provide additional support to the respectable external validity (Messick, 1989) of the TOEIC Bridge test scores. The study of the schools with high scores on English in the Colombian State Examination and low scores in the TOEIC Bridge, and vice versa, may provide additional valuable information.

Also, for both student-level scores and average school-level scores, the TOEIC Bridge reading score has a slightly higher correlation than the TOEIC Bridge listening comprehension score with the English score in the Colombian State Examination, which provides additional validity evidence, as the English examination in the Colombian State Examination has some reading, but no listening.

Ecuador

Scores, as integers between 0 and 20, were available on English and mathematics for the most recent classroom examinations in Ecuador for 292 students who took the TOEIC Bridge test. These students belonged to three schools out of the five who participated in the TOEIC Bridge pilot in Ecuador. Examinees whose classroom examination scores are available have an average score of about 12 points higher on the TOEIC Bridge test (listening comprehension and reading combined) than the other examinees. Table 3 shows the correlations of the classroom examination scores with the TOEIC Bridge scores for the three schools. The table also shows the mean listening comprehension and reading scores for the three schools, and the weighted average of the correlations, weights being proportional to the sample size from each school.

Both the TOEIC Bridge listening comprehension and reading scores have higher association with scores on English in the classroom examination than with scores on mathematics in the classroom examination. This indicates to respectable external validity (Messick, 1989) of the TOEIC Bridge test scores for the Ecuadorian examinees.

Table 3

Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Ecuador

School	Sample size	Mean listen. comp.	Mean reading	Correlation with listen. comp.		Correlation with reading	
				Classroom English	Classroom mathematics	Classroom English	Classroom mathematics
1	135	40.3	35.6	0.47	0.29	0.46	0.27
2	23	31.0	26.6	0.68	0.42	0.79	0.52
3	134	29.5	24.7	0.28	0.16	0.32	0.25
Weighted average	292			0.40	0.24	0.42	0.28

Chile

Scores from classroom examinations of several examinees from Chile (2,246 students out of 2,304 at Vocational Institute 1, 199 out of 205 at Vocational Institute 2, and 874 out of 1,447 at Ministry) who took the TOEIC Bridge test were available. There was (a) an English score, which is the average of scores on all English examinations throughout the year, and (b) an overall score, which is the average of scores on all examinations throughout the year. Both of these scores range between 1 and 7, at intervals of 0.1. Different programs/campuses belonging to each institution administer different sets of classroom examinations. The English tests involve reading and writing, but no listening.

Tables 4, 5, and 6 show, respectively, for Vocational Institute 1, Vocational Institute 2, and Ministry, the correlations of the classroom examination scores with the TOEIC Bridge scores for the programs/campuses from which at least 20 students took the TOEIC Bridge test. The tables also show the mean listening comprehension and reading scores for the programs/campuses, and the weighted average of the correlations, weights being proportional to the sample size from each program/campus. Though there is considerable variability among the programs/campuses, there is a clear pattern. Except for a few programs/campuses (the last two programs listed in Table 4, the fourth program in Table 5, and Campuses 7 and 8 in Table 6, both the TOEIC Bridge listening comprehension and reading scores have higher association with scores on English in the classroom examination than with the overall scores in the classroom examination. For Vocational Institute 1 only, the TOEIC Bridge reading score has a higher correlation with the classroom English scores than the TOEIC Bridge listening comprehension score.

Comparing TOEIC Bridge Scores to the Self-Assessment Reports of Students

Although there are good reasons not to trust self-assessments *completely*, there are reasons for putting at least some stock in them as a validity criterion, especially with regard to language skills (e.g., Upshur, 1975; Shrauger & Osberg, 1981). In many instances, self-assessments may provide a perspective on an individual's language skills that is not captured by other means. Hence, we compared the TOEIC Bridge scores of the examinees to their self-assessments.

Table 4***Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Vocational Institute 1***

Program	Sample size	Mean listen. comp.	Mean reading	Correlation with listen. comp.		Correlation with reading	
				English	Overall	English	Overall
1	109	32.1	28.1	0.45	0.18	0.61	0.35
2	120	29.4	25.8	0.39	0.09	0.33	0.11
3	38	29.0	26.8	0.64	0.38	0.69	0.41
4	186	35.0	29.9	0.55	0.28	0.57	0.33
5	60	34.3	29.0	0.54	0.44	0.57	0.45
6	51	32.7	28.1	0.62	0.43	0.71	0.48
7	71	34.0	28.5	0.59	0.33	0.60	0.34
8	198	35.1	30.9	0.54	0.26	0.53	0.29
9	122	35.6	30.6	0.53	0.21	0.62	0.24
10	50	35.8	30.4	0.55	0.19	0.67	0.20
11	49	33.5	29.7	0.40	0.30	0.47	0.32
12	196	34.2	28.5	0.40	0.32	0.45	0.39
13	37	31.7	24.6	0.44	0.15	0.38	0.28
14	153	31.9	26.5	0.41	0.09	0.46	0.35
15	241	34.5	30.7	0.48	0.18	0.51	0.28
16	100	36.2	30.3	0.37	0.23	0.41	0.39
17	40	31.3	26.9	0.60	0.22	0.59	0.27
18	185	30.6	26.7	0.52	0.30	0.53	0.32
19	38	30.4	22.5	0.26	0.07	0.44	0.36
20	40	30.0	25.1	0.38	0.52	0.46	0.57
21	53	34.9	29.4	0.07	0.21	0.22	0.42
Weighted average	2,137			0.47	0.24	0.51	0.32

Table 5***Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Vocational Institute 2***

Program	Sample size	Mean listen. comp.	Mean reading	Correlation with listening comp.		Correlation with reading	
				English	Overall	English	Overall
1	29	32.7	28.7	0.46	0.28	0.67	0.53
2	21	33.7	30.1	0.42	0.21	0.51	0.40
3	69	29.0	22.8	0.45	0.31	0.37	0.28
4	25	36.6	36.0	0.30	0.31	0.22	0.06
Weighted average	144			0.42	0.29	0.42	0.31

Table 6***Correlations Between the Classroom Examination Scores and TOEIC Bridge Scores for Chile Ministry***

Campus	Sample size	Mean listen. comp.	Mean reading	Correlation with listen. comp.		Correlation with reading	
				English	Overall	English	Overall
1	53	23.2	16.0	0.68	0.34	0.60	0.40
2	89	20.2	17.0	0.40	0.22	0.37	0.32
3	57	16.1	13.6	0.47	0.13	0.42	0.03
4	107	18.4	13.7	0.28	0.24	0.27	0.21
5	135	32.3	31.3	0.58	0.34	0.65	0.50
6	260	29.7	26.8	0.30	0.17	0.44	0.27
7	27	15.1	11.6	0.31	0.34	0.26	0.20
8	24	23.9	13.0	0.46	0.52	0.55	0.64
9	43	20.0	19.3	0.47	0.34	0.57	0.37
10	55	25.4	21.4	0.34	0.15	0.43	0.20
Weighted average	850			0.40	0.24	0.46	0.30

Ecuador and Chile

A few days before they took the test, all of the students in Ecuador and Chile were asked to complete a self-assessment inventory, and to submit their completed inventories to us on the day of testing. The inventory included a series of common language tasks (*can-do* statements; ETS, n.d.), for both listening (25 tasks) and reading (25 tasks). The questions, as well as the instructions associated with them, are included in Appendixes B and C. Most items were chosen from previous studies (e.g., Duke, Kao, & Vale, 2004; Powers, Roever, Huff, & Trapani, 2003). Some new items were developed, and some previously used ones were modified for this effort. The main modification was to include more putatively easy tasks for this population of test takers. The inventory included 12 items (6 for reading and 6 for listening) that were developed to correspond to the six levels of the CEF of Reference for Language. These are the last 6 questions in both Appendixes B and C. The CEF-related statements were included because the CEF is a fairly well-established structure that has been used to describe the language proficiency of foreign-language learners across Europe. The inventory was translated and administered in Spanish. Directions to participants were also given in Spanish.

Test takers were asked to respond to each *can-do* statement using a 5-point scale, with responses as follows: 1 = not at all, 2 = with great difficulty, 3 = with some difficulty, 4 = with little difficulty, and 5 = easily. Respondents were allowed to omit an item if they felt that it did not apply to them, or if they were unable to make a judgment.

Figure 4 shows the self-assessment scores on the above-mentioned 5-point scale versus the TOEIC Bridge test scores for the listening comprehension *can-do* statement, “Using English, I can understand headline news broadcasts on the radio,” given to Chile Ministry students. In the figure, the average listening comprehension scores are plotted (using small circles) along the *y*-axis for each self-assessment score for the item plotted along the *x*-axis. The vertical line around each average TOEIC Bridge score shows a 95% confidence interval around the average value.⁴ The figure shows that students with higher scores on the question have higher TOEIC Bridge scores on average.

Table 7 shows the summary of the correlations among each TOEIC Bridge score (listening and reading) and self-assessment scores for the different administrations in Ecuador and Chile. We computed correlations of the listening *can-do* scores with the TOEIC Bridge Listening Comprehension score and of the reading *can-do* scores with the TOEIC Bridge

Reading score. For each examinee population, we reported the minimum, maximum, and median of the 25 correlations.

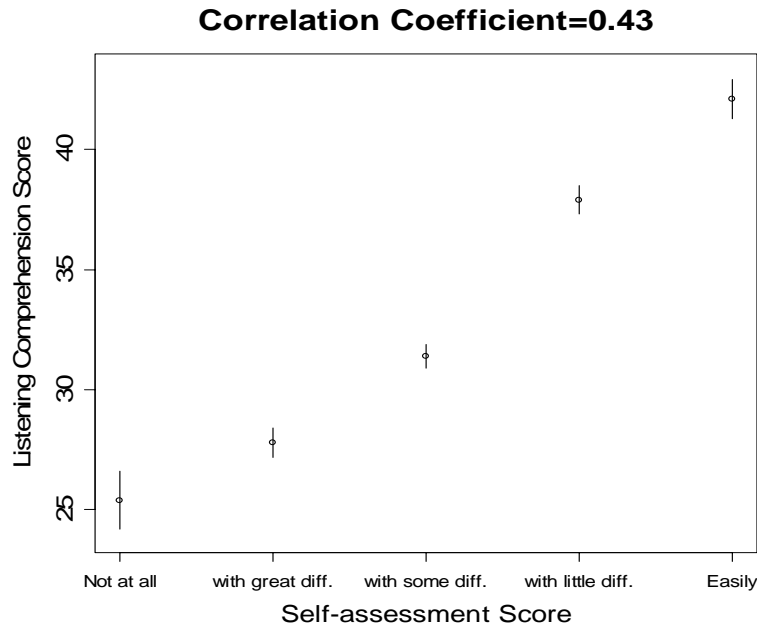


Figure 4. Average TOEIC Bridge Listening scores and corresponding 95% confidence intervals for the different answer choices to the statement, “Using English, I can understand headline news broadcasts on the radio,” for Chile Ministry.

Table 7

Summary of the Correlations Between the Can-Do Scores and the TOEIC Bridge Scores for Ecuador and Chile

Student population	Listening comprehension			Reading		
	Minimum	Maximum	Median	Minimum	Maximum	Median
Ecuador	0.13	0.50	0.35	0.18	0.49	0.38
Chile Vocational Institute 1	0.18	0.43	0.36	0.13	0.43	0.32
Chile Vocational Institute 2	0.08	0.70	0.50	0.04	0.67	0.45
Chile Ministry	0.20	0.46	0.37	0.05	0.55	0.40
Chile Private Institute	0.15	0.67	0.52	0.36	0.62	0.50

Figure 4 and Table 7 demonstrate that, on average, the students who thought that they had good English skills performed better on the TOEIC Bridge tests than those who thought they did not. The level of correlation between the self-assessment scores in the can-do questionnaire and the TOEIC Bridge scores compares favorably with those typically found in other validity studies. For example, in the context of graduate admissions, Graduate Record Examinations[®] (GRE[®]) General test scores generally correlate in the 0.2–0.4 range with graduate grade averages and in the 0.3–0.5 range with such criteria as faculty ratings and performance on comprehensive examinations. Powers et al. (2003) found moderate agreements, that is, correlation coefficients between 0.2 and 0.5, between self-assessment results and student scores in language tests, which they declared to be favorable.

Of special interest are the results for the can-do tasks based on the various CEF proficiency levels, given that the framework has been so widely used as a common basis to describe the language proficiency of foreign-language learners across Europe. These results are based on all the examinees of Chile and Ecuador, combined (total sample size of 4,603). For the listening tasks, we observed a moderate, though not perfect, correspondence between tasks representing each CEF level and test takers' reports of their ability to perform the tasks. For instance, the mean ratings given to the six CEF-related tasks were as follows for levels A1, A2, B1, B2, C1, and C2, respectively: 4.29, 4.19, 4.15, 3.05, 2.62, and 2.42. That is, the tasks are perfectly ordered by study participants, according to their expected difficulty according to the CEF levels, but the difference between any two successive levels varies substantially. For reading tasks, the same perfect correspondence was noted between mean responses and the CEF levels with which statements corresponded: 4.36, 4.17, 3.92, 3.17, 2.68, and 2.46, respectively.

To assess the consistency of test taker responses, we re-administered seven of the can-do statements immediately after the students had taken the TOEIC Bridge test. All of these statements had been administered previously in the full inventory that students completed before they came to the testing site. By repeating the seven statements on two different occasions, several days apart, we were able to assess the stability of self-assessment reports, to some extent. The retest correlations were 0.54, 0.55, and 0.56 for the three listening comprehension statements that were repeated. The four reading statements exhibited retest correlations of 0.50, 0.56, 0.57, and 0.60. These correlations suggest reasonably consistent results across occasions for single-

item measures. More confidence can be placed in the validity of our self-assessment reports because test takers rank the tasks similarly on different occasions.

For more details on the can-do study for Chile and Ecuador, see Powers et al. (2008).

Colombia

We could ask only six can-do questions (three listening and three reading) of the examinees in Colombia, due to logistical constraints. The correlations between the TOEIC Bridge scores and can-do scores are 0.30, 0.38, and 0.34 for listening and 0.35, 0.40, and 0.37 for reading, which are favorable.

Discussion of the Self-Assessment Results

The above results have provided some evidence of the validity of TOEIC Bridge scores by linking them to test takers' assessments of their ability to perform a variety of everyday English language activities. The relationships are far from perfect, but they are believed to be practically meaningful. Although examinee self-reports cannot be accepted uncritically as a validity criterion, they do have characteristics that seem to suggest their trustworthiness, especially since examinees in this low-stakes research study had no incentive to intentionally distort their reports. The results of the study reported here are consistent with previous meta-analytic summaries (e.g., Ross, 1998) that have documented substantial correlations between the self-ratings of learners of English as a second language and a variety of criterion measures.

Comparing TOEIC Bridge Scores to Teacher Ratings

Powers and Simpson (2008) provided details about the comparison of the TOEIC Bridge scores and teachers' ratings for Chile. We discuss some of their important findings here. A 14-item can-do inventory, consisting of eight everyday English language reading tasks and six everyday English language listening tasks, was created. Each of the 14 can-do tasks was included in the 50-item self-assessment inventory discussed in the previous section.

The 14-item inventory was administered in an interview setting to each of 10 teachers at seven educational institutions (four secondary-level and three technical institutes) in Chile, in the spring of 2007. Each of these teachers was asked to complete the inventory for each of several students in his/her class, indicating whether or not he/she thought each student could perform each of the tasks. Response choices were on a 5-point scale, as follows:

1. Cannot do at all.
2. Can do with great difficulty.
3. Can do with some difficulty.
4. Can do with little difficulty.
5. Can do easily.

Teachers were also allowed to indicate that they were unable to judge the student's ability to perform a task. While the responses were in English, the questions were posed in Spanish (verbally and written), to ensure understanding.

Students were selected by the interviewer to represent students who had scored high, middle, or low on the TOEIC Bridge test. The TOEIC Bridge scores were not known to the teachers. These 10 teachers rated a total of 59 students. Four of the teachers rated 4 students, two rated 5 students, two rated 6 students, and two rated either 10 or 11 students.

Table 8 shows the mean (and SDs) of teacher ratings for each of the 14 can-do tasks. The ratings made by the teachers were estimated to be highly reliable, with Cronbach alpha reliability estimates of 0.94 for the reading tasks and 0.92 for the listening tasks. Furthermore, as shown in Table 8, teacher ratings for the tasks correspond relatively well with our expectations of the difficulty of the tasks. For example, for the reading tasks, read and understand a popular novel was rated as the most difficult task for these students, while recognize memorized words and phrases (for example, "Exit," "Entrance," and "Stop") was rated as easiest. Thus, we have some confidence in the reliability of these ratings.

Table 8 also contains, for each language task, the correlation of teacher ratings with TOEIC Bridge scores. As can be seen, for individual reading tasks, the correlation of teachers' ratings ranged from 0.32 to 0.58. The correlation of a composite index based on all 8 reading tasks was 0.47. For individual listening tasks, the correlations ranged from 0.32 to 0.63, and 0.60 for a composite index based on all 6 listening tasks.

Each of the 59 students rated by teachers also completed the can-do inventory (25 reading tasks and 25 listening tasks) given in Appendixes B and C. Table 9 presents the correlations among teacher ratings, student self-ratings, and TOEIC Bridge scores, for these students. As can be seen, the correlation between teacher and student ratings, although statistically significant, was modest (0.35 for listening and 0.28 for reading). However, both teacher and student ratings

correlated moderately well with TOEIC Bridge scores for both listening (0.60 and 0.51) and reading (0.47 and 0.49) tasks.

Table 8

Means (and Standard Deviations) for Teacher Ratings of Can-Do Tasks and Correlations With TOEIC Bridge Scores

Can-do statement	Mean	SD	Correlation
			with TOEIC Bridge
			Reading
Read and understand a popular novel.	2.57	1.2	.58
Understand the viewpoints expressed in articles and reports about contemporary issues or problems.	2.81	1.1	.49
Read and understand a travel brochure.	3.66	1.1	.40
Understand short, simple texts (e.g., personal letters).	3.98	1.0	.39
Understand the main point of simple messages and short, clear announcements.	4.03	0.9	.41
Read and understand a simple postcard from a friend.	4.20	0.9	.44
Understand any (a) reading and (b) written instructions (in English) required as part of his/her schoolwork.	4.36	0.8	.47
Recognize memorized words and phrases (for example, “Exit,” “Entrance,” and “Stop”).	4.46	0.8	.32
Total reading	3.79	0.8	.47
			Listening
Understand extended speech and lectures, and follow complex arguments on familiar topics.	2.47	1.2	.59
Understand an extended debate on a relatively complex topic.	2.80	1.1	.58
Understand a person in social situations talking about his/her background, family, or interests.	4.07	0.9	.48
Understand someone who is speaking slowly and deliberately about his or her hobbies and interests.	4.10	1.0	.41
Understand discussions and oral instructions (in English) required as part of his/her schoolwork.	4.14	0.9	.63
Understand simple questions in social situations such as “How are you?” and “Where do you live?”	4.53	0.7	.32
Total listening	3.68	0.8	.60
Total	3.74	0.8	.54

Note. All but two correlations are significant at the .01 level; the other two are significant at the .05 level. $N = 59$.

Table 9

Correlations among Student Self Ratings, Teacher Ratings, and TOEIC Bridge Scores for the Examinees in Chile

Skill		Method		
		Rating		TOEIC Bridge (TB)
		Teacher (T)	Self (S)	
Listening	T	1.00	.35*	.60**
	S		1.00	.51**
	TB			1.00
Reading	T	1.00	.28*	.47**
	S		1.00	.49**
	TB			1.00

* $p < .05$. ** $p < .01$

We also examined this relationship for each of the 10 teachers, even though the correlations are based on only a small number of data points for each teacher. For individual teachers, the correlation of TOEIC Bridge scores with the sum or ratings over all tasks ranged from 0.29 to 0.93 for reading (median = 0.87), and from 0.28 to 0.99 for listening (median = 0.84). These correlations are high, on average, but should not be over-emphasized, because they are based on very small samples.

The results above suggest that the TOEIC Bridge scores relate moderately well to teacher reports of students' English language skills. Within the limits of the small samples available, these results provide reasonably compelling evidence of the validity of the TOEIC Bridge as an indicator of English language proficiency.

Discussion/Implications

Scores on the TOEIC Bridge test meet several important conditions that show the test scores to be valid as measures of English language proficiency for these examinees. These conditions are as follows:

- Reliable scores.
- Items delineate well between the low-scoring examinees and the high-scoring examinees as shown from the relatively strong biserial correlations of the items with total score.

- Scores have a moderate to high association with related measures, as follows:
 - Scores correlate moderately to highly with English scores in other examinations administered in the examinees' own countries.
 - Scores moderately agree with student self-assessment scores.
 - Scores moderately agree with teacher self-assessment scores.

While this study has not *proved* conclusively that TOEIC Bridge scores are valid as a measure of English language proficiency for examinees in several South American countries, evidence collected in this study is consistent with such an interpretation.

The factor analysis results and the correlations between scores on parts of the TOEIC Bridge test show that the listening comprehension and reading scores are not well-separated, at least not in the sample of beginning learners that were studied here. However, compared to listening comprehension, the reading scores mostly correlated higher with scores on other English tests that cover only reading and no listening comprehension, which is consistent with listening comprehension and reading factors that are separated. We recommend monitoring the correlational and factorial structure on the results of possible future administrations of the test to these examinees.

We plan to perform a collection of expert judgments to determine score ranges on the TOEIC Bridge test for job performance and higher education uses. For the job performance use, that would mean an expedited activity, like a standard setting included in a job analysis. For the higher education use, that would mean an expedited activity, like a standard setting included in a curriculum survey. These results would provide a way to review the data described in this paper, to ensure that the test performs well for score ranges where decisions will be made.

This study has several limitations. First, all the data were collected from trial administrations of the TOEIC Bridge test, so one could argue that the lack of motivation of the examinees could have affected our results. The student sample also may not have been representative of any real examinee population of interest. For example, all the examinees were from South America, but the largest South American countries such as Brazil and Argentina were not involved, making the examinee population not representative of the South American student population. Also, though we were able to collect some scores from tests given in the students' countries, they were not always adequate, relevant, or of high quality. For example, for the

students of Private Institute, we could not obtain any scores on an English test. In addition, we did not have any information on reliability or validity of these scores. Finally, because students often belonged to institutions, a hierarchical model with students as first-level observational units and institutions as second-level observational units could also have been applied.

References

- Duke, T., Kao, C., & Vale, D. C. (2004, April). *Linking self-assessed English skills with the Test of English for International Communication (TOEIC)*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- ETS. (2007). *TOEIC Bridge examinee handbook*. Retrieved May 1, 2008, from <http://www.ets.org/Media/TOEIC%20BridgeExam.pdf>
- ETS. (n.d.). *TOEIC can-do guide: Linking TOEIC scores to activities performed in English*. Princeton, NJ: Author.
- Gulliksen, H. (1950). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.
- Oller, J. W., Jr. (1979). *Language tests at school*. London: Longman.
- Powers, D. E., Bravo, G. M., Sinharay, S., Saldivia, L. E., Simpson, A. G., & Weng, V. Z. (2008). *Relating scores on the TOEIC Bridge to student perceptions of proficiency in English* (ETS Research Memorandum No. RM-08-02). Princeton, NJ: ETS.
- Powers, D. E., Roeber, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge Courseware scores against faculty ratings and student self-assessments* (ETS Research Rep. No. RR-03-11). Princeton, NJ: ETS.
- Powers, D. E., & Simpson, A. G. (2008). *Validating TOEIC Bridge scores against teacher and student ratings: A small-scale study* (ETS Research Memorandum No. RM-08-03). Princeton, NJ: ETS.
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1-20.
- Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others of psychological assessment. *Psychological Bulletin, 90*, 322-351.
- Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping test scores onto the Common European Framework* (ETS Research Rep. No. RR-05-18). Princeton, NJ: ETS.

- Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT Research Series Rep. No. TOEFLiBT-06) Princeton, NJ: ETS.
- Upshur, J. (1975). Objective evaluation of oral proficiency in the ESOL classroom. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967-1974* (pp. 53-65). Washington, DC: TESOL.

Notes

¹ The Common European Framework (CEF) of Reference for Language has six levels – from basic user (levels A1 and A2), to independent user (levels B1 and B2), to proficient user (levels C1 and C2).

² See, for example, Gulliksen (1950), p. 101, for a description of correcting correlations for attenuation. The correlation corrected for attenuation is an estimate of the correlation that would be obtained if the test lengths were increased to infinity (which would make the tests perfectly reliable) and is obtained by dividing the simple correlation between the scores in two tests by the square root of the product of the reliabilities of the two tests.

³ The mathematics test in the Colombian State Examination has 24 multiple choice items.

⁴ Note that due to multiple comparison issues, suitable adjustments should be made to the figure to perform a pairwise 95% level comparison, or a 95% overall comparison.

Appendix A

Summary Statistics for the Combined Population in the Different Parts of the Test

Type	Part	Listening			Reading	
		Photo- graphs	Stimulus- response	Short cons/talks	Incomplete sentences	Reading compre- hension
Mean		9.54	10.14	6.96	12.08	11.37
Reliability		0.83	0.86	0.79	0.86	0.88
		Listening				
	Photographs	1.00	0.63	0.62	0.63	0.66
	Stimulus- response	0.90	1.00	0.69	0.71	0.65
	Short cons/talks	0.91	0.98	1.00	0.70	0.64
		Reading				
	Incomplete sentences	0.85	0.92	0.95	1.00	0.68
Correlation	Reading comprehension	0.87	0.85	0.88	0.87	1.00

Note. The correlation coefficient between listening comprehension and reading scores is 0.94, after correcting for attenuation. The shaded entries in the table are the corresponding correlation coefficients after correcting for attenuation. The first three parts belong to listening comprehension and the last two to reading. The within-section correlations are marked in bold. Cons = conversations.

Appendix B
Listening Comprehension Can-Do Questionnaire

Below you will find several statements about English-language listening activities. For each statement, please circle the one number that you believe best represents your ability to perform the activity in English. If you have never actually performed the activity that is described, please rate how easily you believe you could perform the activity if you had to do so in English.

Circle the one number that is most true for you.		Not at all	With great difficulty	With some difficulty	With little difficulty	Easily
<i>Using English, I Can. . .</i>						
1	understand simple questions in social situations such as “How are you?” and “Where do you live?”	1	2	3	4	5
2	understand a salesperson when she or he tells me prices of various items	1	2	3	4	5
3	understand a classmate discussing a simple problem that arose in class	1	2	3	4	5
4	understand someone who is speaking slowly and deliberately about his or her hobbies and interests	1	2	3	4	5
5	understand directions about what time to come to a class and where it will be held	1	2	3	4	5
6	understand headline news broadcasts on the radio	1	2	3	4	5
7	understand a complex presentation or demonstration given in a classroom	1	2	3	4	5
8	understand an explanation given over the radio of why a road has been temporarily closed	1	2	3	4	5
9	understand a person’s name when she or he gives it to me over the telephone	1	2	3	4	5
10	understand play-by-play descriptions on the radio of sports events that I like (e.g., soccer)	1	2	3	4	5

Circle the one number that is most true for you		Not at all	With great difficult y	With some difficulty	With little difficulty	Easily
<i>Using English, I Can. . .</i>						
11	understand the days of the week and the months of the year	1	2	3	4	5
12	understand an extended debate on a complex topic such as public health	1	2	3	4	5
13	understand explanations about how to perform a routine task related to my class	1	2	3	4	5
14	understand a person in social situations talking about his/her background, family, or interests	1	2	3	4	5
15	understand some memorized words and phrases	1	2	3	4	5
16	understand a discussion of current events taking place among a group of persons speaking English	1	2	3	4	5
17	understand public announcements that are broadcast	1	2	3	4	5
18	understand someone speaking slowly and deliberately, who is giving me directions on how to walk to a nearby location	1	2	3	4	5
19	understand the main ideas in news reports broadcast on the radio or TV	1	2	3	4	5
20	recognize familiar words and simple phrases, when people speak slowly and clearly	1	2	3	4	5
21	understand the main point of simple messages and short, clear announcements	1	2	3	4	5
22	understand the main points of standard speech on familiar matters, when people speak slowly and clearly	1	2	3	4	5
23	understand extended speech and lectures, and follow complex arguments on familiar topics	1	2	3	4	5
24	understand extended speech, even when it is not clearly structured	1	2	3	4	5
25	understand any kind of spoken language, even when people speak quickly	1	2	3	4	5

Appendix C
Reading Can-Do Questionnaire

Below you will find several statements about English-language reading activities. For each statement, please circle the one number that you believe best represents your ability to perform the activity in English. If you have never actually performed the activity that is described, please rate how easily you believe you could perform the activity if you had to do so in English.

Circle the one number that is most true for you		Not at all	With great difficulty	With some difficulty	With little difficulty	Easily
<i>Using English, I Can</i>						
1	read and understand a restaurant menu	1	2	3	4	5
2	read highly technical material related to my classes with little use of a dictionary	1	2	3	4	5
3	read and understand magazine articles without using a dictionary	1	2	3	4	5
4	read and understand simple, step-by-step instructions (for example, how to operate an electronic machine, like a CD player)	1	2	3	4	5
5	find information that I need in a telephone directory	1	2	3	4	5
6	read, on storefronts, the type of store or services provided (e.g., “dry cleaning,” “book store”)	1	2	3	4	5
7	read a newspaper editorial and understand its meaning as well as the writer’s intent	1	2	3	4	5
8	read entertainment-related information (for example, tourist guides)	1	2	3	4	5
9	read information about products (for example, advertisements)	1	2	3	4	5
10	read English to translate text into my own language (for example, letters and technical documents)	1	2	3	4	5

Circle the one number that is most true for you		Not at all	With great difficulty	With some difficulty	With little difficulty	Easily
<i>Using English, I Can</i>						
11	recognize memorized words and phrases (for example, “Exit,” “Entrance,” and “Stop”)	1	2	3	4	5
12	identify inconsistencies or differences in points of view in two newspaper interviews with politicians of opposing parties	1	2	3	4	5
13	read and understand directions and explanations presented in manuals written for beginning users	1	2	3	4	5
14	read and understand a popular novel	1	2	3	4	5
15	read and understand a letter of thanks from a friend or a relative	1	2	3	4	5
16	read and understand a train or bus schedule	1	2	3	4	5
17	read and understand traffic signs	1	2	3	4	5
18	read and understand the main points of an article on a familiar academic topic	1	2	3	4	5
19	read and understand a simple e-mail from a friend	1	2	3	4	5
20	understand familiar words and very simple sentences	1	2	3	4	5
21	understand short, simple texts (e.g., personal letters)	1	2	3	4	5
22	understand texts that consist mainly of everyday words or vocabulary	1	2	3	4	5
23	understand the viewpoints expressed in articles and reports about contemporary issues or problems	1	2	3	4	5
24	understand long, complex texts, even technical ones that do not relate to my classes	1	2	3	4	5
25	understand virtually all forms of written language, including abstract and linguistically complex texts	1	2	3	4	5