TOEIC®

Number 1

Research Summaries

Relating TOEIC Scores to Oral Proficiency Interview Ratings

by Kenneth Wilson

Educational Testing Service

Edited by Lorraine Luciano from "Enhancing the Interpretation of a Norm-Referenced Second-Language Test through Criterion Referencing: A Research Assessment of Experience in the TOEIC Testing Context" by Kenneth Wilson.

Introduction

The Test of English for International Communication (TOEIC®) was developed to measure English language skills used in international corporations around the globe. The test is especially useful for making decisions regarding hiring, job placement, promotions, and placement in courses of business English-language training. It is also helpful in monitoring the progress made by individuals and groups as they seek to improve their abilities in English through various forms of study.

The four basic skills in most languages are listening and reading (passive skills) and speaking and writing (active skills). Directly measuring active skills is both time-consuming and costly. Even when using highly trained people to score active skills tests, the process tends to be subjective. TOEIC measures the passive skills of reading and does so in a demonstrably reliable and cost-effective way.

Many corporations that find TOEIC to be a valuable tool in making personnel decisions are also interested in the speaking ability of their employees and recruits. Dr. Kenneth Wilson, a researcher at Educational Testing Service, examined the relationship between TOEIC scores and performance on a direct test of speaking ability. His research indicates group patterns of predictive association between the two measures.

This Research Summary briefly describes Dr. Wilson's research. For a copy of the complete research study, contact your local TOEIC representative, or write to:

TOEIC Service International
Educational Testing Service
Rosedale Road
Princeton, New Jersey, 08541
U.S.A.

Criterion- and Norm-Referenced Tests

Tests, including second-language proficiency tests, can be either of two types: criterion-referenced tests or norm-referenced tests. Both supply needed and valuable information, but in different ways. Criterion-referenced tests are used to identify an individual's status with respect to an established standard of performance. For example, "X" must correctly answer 70 percent of the questions on a given test in order to demonstrate his/her competence in that area, i.e., to "pass" the test. Or, "X" may earn a particular score or grade and is, as a result, expected to possess certain abilities or to have mastered specific tasks within the area tested.

Norm-referenced tests, on the other hand, measure the learner's proficiency in relation to the performance of other individuals on the same measure, i.e., "X" performed better than "Y" but not as well as "Z." These scores do not, in and of themselves, establish, define, or explain distinct levels of ability associated with various scores.

TOEIC is an example of a norm-referenced test. The TOEIC test measures English-language proficiency in the international work environment, unlike other tests that commonly focus on English as it is used in an academic setting, such as a college or university. TOEIC is designed to evaluate, in the context of real-life, business-world situations, the English language listening comprehension and reading ability of those adults whose native language is not English. It consists of 200 multiple-choice questions divided into two sections — 100 listening comprehension items, administered by audiotape, and 100 reading items. Examinees indicate their answers to the questions by marking the letter A, B, C, or D with a pencil on a scannable answer sheet. The test takes approximately two hours to administer. The number of correct answers on each of the two sections is translated into a standard score with values ranging from 5 to 495 for each section, plus a total test score of between 010 and 990.

TOEIC is a "user-defined" test in that scores can be considered in a variety of ways depending upon the requirements of a particular individual or client. For example, one client may use TOEIC scores as a consideration in recruiting, while another uses TOEIC as a tool to place employees within various language training programs. Appropriate scores or score ranges for any given situation can be established and reestablished as needed. This flexibility — due to its wide scale — has been one of many reasons for TOEIC's acceptance by the international business community.

Correlating TOEIC Scores and Language Proficiency Interview Ratings

In an effort to expand the interpretation of TOEIC scores, the present study was undertaken using a standard research procedure that links the scores on one test (or type) to scores on another. In this case, TOEIC was linked to the scores of the Language Proficiency Interview (LPI), 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 TOEIC (Listening Comprehension and Reading) 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.

LPI ratings range from 0 (no proficiency) through 5 (proficiency equivalent to that of a well-educated native speaker). (See Table 1 for a summary of each level of proficiency as reflected by function, context, and accuracy. These descriptions are meant to provide the reader with a quick overview of each level; the actual descriptions used as part of the LPI are exceedingly more involved and complex.) The interview consists of a face-to-face conversation of approximately 10-30 minutes between the interviewee and a trained interviewer who is a native or near-native speaker of the target language. The conversation begins at a fairly simple level and becomes increasingly more difficult, with an increased rate of speech on the part of the interviewer, use of more complex structures, and more specialized vocabulary. Each of the six principle points (0, 1, 2, 3, 4, 5) is characterized by a clearly defined pattern of language-use behavior or ability. Additionally, a "+" (or ".5") is added to a rating for individuals whose performance is judged to substantially exceed that for a given level, but not to meet fully the requirements for the next higher level. This method results in a more discriminating 11-point scale (0, 0+ [or .5], 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5).

By establishing the connection between TOEIC and the LPI, it is possible to examine the overall trends and relationships that exist between these two different measures. To determine whether consistent patterns of association exist between TOEIC and LPI ratings, a regression-based calibration model is employed. Simply, scores and ratings are obtained from examinees who have taken both TOEIC and the LPI. Analyses involved TOEIC scores for Listening Comprehension (LC) and for Reading (R), as well as the combined Total Score (LC + R). TOEIC scores are then referenced or matched (correlated)

with the LPI ratings. This regression-based approach examines the strength and consistency of relationship between scores on the TOEIC and ratings on the LPI, as indicated by "correlation coefficients" that can range between .00 (indicating no relationship) and 1.00 (indicating a perfect relationship). It also yields equations that can be used to estimate, with known statistical accuracy, the most probable LPI level for individuals with particular scores on the TOEIC. In other words, if these relationships are shown to be consistent and mathematically significant, we can estimate, from TOEIC scores alone, the linguistic behaviors that would most likely be exhibited by examinees in an LPI. Please be aware that these findings are meant to be examined in a research setting only. Conclusions are based on information about specific groups and are, therefore, not appropriate for making predictions about any one individual's possible performance. Nor can the assumption be made that these two measures should be viewed as interchangeable. While we can establish certain relationships between the tests, each employs different methods of assessment to evaluate specific, unique abilities.

The present study focused primarily on data pertaining to the TOEIC testing context in Japan, where TOEIC/LPI data have been generated for several years. Similar, but less extensive, TOEIC/LPI data sets were also available from France, Mexico, and Saudi Arabia. Assessment was made of the level and pattern of relationships between the LPI and TOEIC scores (LC, R, and Total [LC + R]) in the Japanese sample, in each of the several non-Japanese samples, in the total non-Japanese sample, and in the combined sample of Japanese and non-Japanese examinees.

For the Japanese sample, data was collected from four sources. One data-set, consisting of TOEIC scores and LPI ratings for 122 individuals, was collected in 1985 for the explicit purpose of evaluating TOEIC/LPI relationships. Three additional data sets (163 individuals) involving the joint use of TOEIC and LPI scores were collected during 1984, 1986, and 1987 by the Institute for International Studies and Training (IIST), a graduate-level business school, and made available for the present study.

Summary statistics are shown for the four Japanese subsamples in Table 2. Each of the subsample groups had average LPI ratings of 1+. A speaker at level 1+ exhibits a fair amount of LPI level 2 language behavior but, as the plus level implies, the speaker is not able to consistently sustain this level of language. Level 2 speakers can participate fully in casual conversations, express facts, give instructions, describe, and provide narration about current, past and future activities (See Table 1).

TABLE 1: ORAL PROFICIENCY LEVELS AS DEFINED BY FUNCTION, CONTEXT, AND ACCURACY

ORAL PROFICIENCY LEVEL	FUNCTION (Tasks accomplished, attitudes expressed, tone conveyed)	CONTEXT (Topics, subject areas, activities, and jobs addressed)	ACCURACY (Acceptability, quality, and accuracy of message conveyed)
5	Functions equivalent to an educated native speaker (ENS).	All subjects.	Performance equivalent to an ENS.
4	Able to tailor language to fit audience, counsel, persuade, negotiate, represent a point of view, and interpret for dignitaries.	All topics normally pertinent to professional needs.	Nearly equivalent to an ENS. Speech is extensive, precise, appropriate to every occasion with only occasional errors.
3	Can converse in formal and informal situations, resolve problem situations. Deal with unfamiliar topics, provide examinations, describe in detail, offer supported opinions, and hypothesize.	Practical, social, professional, and abstract topics, particular interests, and special fields of competence.	Errors never interfere with understanding and rarely disturb the native speaker. Only sporadic errors in basic structures.
2	Able to fully participate in casual conversations, can express facts, give instructions, describe, report, and provide narration about current, past, and future activities.	Concrete topics such as own background, family, interests, work, travel, and current events.	Understandable to native speaker not used to dealing with foreigners. Sometimes miscommunicates.
1	Can create with the language, ask and answer questions.	Everyday survival topics and courtesy requirements, participate in short conversations.	Intelligible to native speaker used to dealing with foreigners.
0	No functional ability.	None.	Unintelligible.

TABLE 2: SUMMARY STATISTICS FOR THE JAPANESE SAMPLE(S)

Sample	Number	MEAN TOEIC SCORES			MEAN LPI SCORES
		Listening Comprehension	Reading Comp.	Total	
TOEIC-85	122	311	300	611	1.89
IIST-84	66	313	295	608	1.78
IIST-86	55	329	310	640	1.87
IIST-87	42	315	309	624	1.93
TOTAL	**285**	**316**	**302**	**618**	**1.86**

Table 3 shows the intercorrelations, or degree of relationship, between the study variables for the four individual Japanese subsamples as well as for the total Japanese sample. In general: 1) TOEIC-TOTAL was most closely related to LPI; 2) TOEIC-LC was almost as closely related to LPI as was TOEIC-TOTAL; and 3) TOEIC-LC was more closely associated with LPI than TOEIC-R.

TABLE 3: INTERCORRELATIONS OF VARIABLES FOR THE JAPANESE SAMPLE(S)

LANGUAGE PROFICIENCY INTERVIEW (LPI)	NUMBER	TOEIC-LC	TOEIC-R	TOEIC-T (LC+R)
TOEIC-85	(N=122)	.79	.72	.80
IIST-84	(N= 66)	.67	.68	.71
IIST-86	(N= 55)	.80	.65	.76
IIST-87	(N= 42)	.73	.70	.75
TOTAL	**(N=285)**	**.75**	**.69**	**.76**

NOTE: A perfect correlation (relationship) between two variables = 1.0.

Data for the four subsamples (N = 285) were analyzed to establish how closely the predicted or estimated LPI ratings matched actual LPI ratings for all TOEIC score ranges in the study sample. Table 4 shows: (a) designated score ranges for TOEIC-TOTAL and TOEIC-LC scores, (b) midpoint of each score range, (c) the number of examinees in each of the ranges, (d) the mean ESTIMATED LPI level for individuals at the midpoint of each range, based on the regression equation, and (e) the mean ACTUAL LPI ratings for examinees in each score range.

TABLE 4: ESTIMATED AND OBSERVED LPI LEVELS ASSOCIATED WITH DESIGNATED LEVELS OF PERFORMANCE ON TOEIC-TOTAL AND TOEIC-LISTENING COMPREHENSION

RANGE OF SCORES	MIDPOINT	N	MEAN LPI RATING	
			Estimated	Actual
TOEIC-Total				
200 – 299	250	5	.62	.90
300 – 399	350	16	.96	1.00
400 – 499	450	41	1.30	1.38
500 – 599	550	68	1.64	1.59
600 – 699	650	67	1.97	1.89
700 – 799	750	48	2.31	2.29
800 – 899	850	31	2.65	2.68
900+	950	9	2.99	3.06
TOEIC-LC				
100 – 149	125	4	.71	.63
150 – 199	175	18	1.01	1.19
200 – 249	225	37	1.32	1.34
250 – 299	275	62	1.62	1.58
300 – 349	325	72	1.92	1.85
350 – 399	375	38	2.23	2.11
400 – 449	425	37	2.53	2.65
450+	475	17	2.83	2.85
TOTAL SAMPLE		**285**	**1.86**	**1.86**

Tables 5.1 and 5.2 provide more comprehensive information, including the actual distributions of LPI ratings (in percent) for each specific TOEIC score interval, i.e., for TOEIC-Total and TOEIC-LC, respectively. The actual number of examinees falling into each category is also listed (in parentheses) next to these percentages.

TABLE 5.1: RELATIONSHIP BETWEEN TOEIC-TOTAL SCORE AND LPI RATING FOR JAPANESE SAMPLE

TOEIC-TOTAL	PERCENT OF EXAMINEES AT LPI LEVEL							
	0+	1	1+	2	2+	3	>3	N
900+					33 (3)	33 (3)	33 (3)	9
800 – 895				16 (5)	45 (14)	29 (9)	10 (3)	31
700 – 795			12 (6)	38 (18)	31 (15)	16 (8)	2 (1)	48
600 – 695		9 (6)	22 (15)	57 (38)	8 (5)	4 (3)		67
500 – 595		19 (13)	46 (31)	34 (23)	2 (1)			68
400 – 495	7 (3)	24 (10)	56 (23)	10 (4)	2 (1)			41
300 – 395	25 (4)	50 (8)	25 (4)					16
200 – 295	40 (2)	40 (2)	20 (1)					5
	3.2 (9)	**13.7 (39)**	**28.1 (80)**	**30.9 (88)**	**13.7 (39)**	**8.1 (23)**	**2.5 (7)**	**285**

TABLE 5.2: RELATIONSHIP BETWEEN TOEIC-LC SCORE AND LPI RATING FOR JAPANESE SAMPLE

TOEIC-LC	PERCENT OF EXAMINEES AT LPI LEVEL							
	0+	1	1+	2	2+	3	>3	N
450+				6 (1)	41 (7)	35 (6)	18 (3)	17
400 – 445			3 (1)	19 (7)	38 (14)	30 (11)	11 (4)	37
350 – 395		3 (1)	18 (7)	42 (16)	29 (11)	8 (3)		38
300 – 345		11 (8)	21 (15)	58 (42)	6 (4)	4 (3)		72
250 – 295	2 (1)	13 (8)	56 (35)	26 (16)	3 (2)			62
200 – 245	8 (3)	35 (13)	41 (15)	14 (5)	3 (1)			37
150 – 195	11 (2)	44 (8)	39 (7)	6 (1)				18
<150	75 (3)	25 (1)						4
	3.2 (9)	**13.7 (39)**	**28.1 (80)**	**30.9 (88)**	**13.7 (39)**	**8.1 (23)**	**2.5 (7)**	**285**

In addition to research with the Japanese sample, analyses were also conducted which included data from France, Mexico, and Saudi Arabia. A total of 393 subjects participated in these extended analyses (the original 285 from Japan, plus 56 from France, 42 from Mexico, and 10 from Saudi Arabia). These data were collected in 1987 and 1988 by TOEIC/ETS staff in conjunction with corporate clients in these countries. Identical statistical methods were applied to this larger and more diverse sample. Again, performance on the Language Proficiency Interview was strongly and consistently related to specific TOEIC performance. As before, TOEIC-LC was more highly correlated with LPI ratings than was TOEIC-R; TOEIC-LC/LPI correlations were extremely close to TOEIC-TOTAL/LPI correlations. Table 6 provides summary statistics for the combined sample.

TABLE 6: SUMMARY STATISTICS, INCLUDING INTERCORRELATIONS OF VARIABLES, FOR THE COMBINED STUDY SAMPLE

SAMPLE	N	MEAN SCORES				CORRELATIONS WITH LPI		
		TOEIC LC	TOEIC R	TOEIC TOTAL	LPI	TOEIC LC	TOEIC R	TOEIC TOTAL
France (F)	56	428	389	817	2.30	.62	.58	.65
Mexico (M)	42	262	237	499	1.71	.78	.70	.76
Saudi Arabia (S)	10	304	184	489	1.95	.85	.86	.87
Total (F, M, S)	**108**	**352**	**311**	**663**	**2.04**	**.74**	**.67**	**.73**
Japan	285	316	302	618	1.86	.75	.69	.76
Combined	**393**	**325**	**305**	**630**	**1.91**	**.74**	**.68**	**.74**

The relationship between mean ACTUAL LPI ratings and mean LPI ratings predicted from TOEIC-LC scores for this combined sample is presented in Figure 1. *(NOTE: The "x" 's represent the mean actual LPI ratings; the solid line represents the mean ratings predicted by the regression equation.)* The mean actual and estimated LPI ratings reflect the average of those ratings occurring within the specific TOEIC score ranges, i.e., between 150-195, between 200-245, etc. Table 7 shows the actual distributions of LPI ratings for the specific score ranges of the TOEIC-LC for this larger sample. Analyses with TOEIC-TOTAL scores yielded similar, consistent results.

FIGURE 1: RELATIONSHIP BETWEEN MEAN ACTUAL LPI RATINGS AND MEAN LPI RATINGS PREDICTED FROM TOEIC-LC SCORES, FOR THE COMBINED SAMPLE

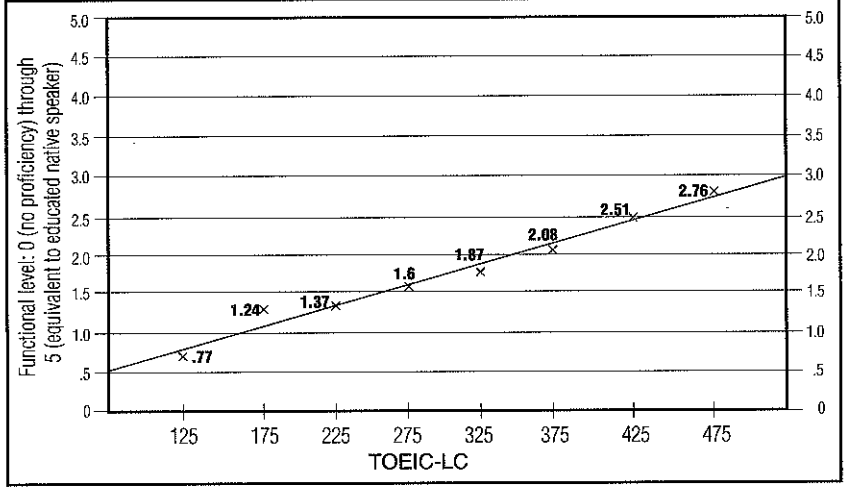


TABLE 7: RELATIONSHIP BETWEEN TOEIC-LC SCORE AND LPI RATING FOR COMBINED SAMPLE

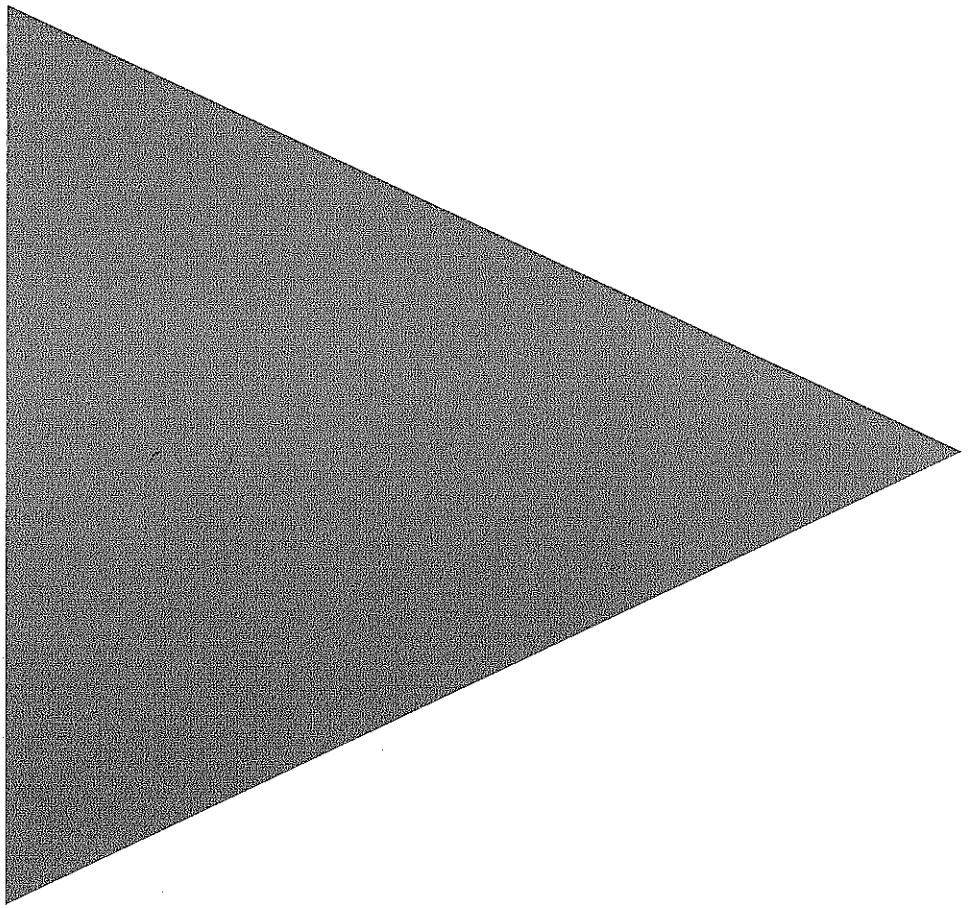
TOEIC-LC	PERCENT OF EXAMINEES AT LPI LEVEL							
	0+	1	1+	2	2+	3	>3	N
450+				17 (8)	30 (14)	40 (19)	13 (6)	47
400 – 445			9 (5)	21 (11)	36 (19)	26 (14)	8 (4)	53
350 – 395		4 (2)	18 (9)	45 (23)	26 (13)	8 (4)		51
300 – 345		9 (8)	20 (18)	61 (54)	6 (5)	3 (3)		88
250 – 295	3 (2)	11 (8)	55 (40)	26 (19)	6 (4)			73
200 – 245	7 (3)	33 (15)	42 (19)	16 (7)	2 (1)			45
150 – 195	12 (3)	40 (10)	36 (9)	12 (3)				25
<150	64 (7)	18 (2)	18 (2)					11
	3.8 (15)	**11.5 (45)**	**26.0(102)**	**31.8(125)**	**14.2 (56)**	**10.2 (40)**	**2.6 (10)**	**393**

▶ ▶ ▶ ▶ ▶ ▶ ▶ Conclusion

Although this study was designed to explore the extent to which TOEIC scores could imply active speaking ability in English (through the TOEIC-LPI relationship), it has also shed additional light on test *validity*. A test is valid if it tests what it is supposed to test. 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. Although the LPI requires a response in spoken English, while the TOEIC requires an examinee to answer questions printed in English in the test booklet, both the LPI and TOEIC-LC measure the underlying ability to comprehend spoken English. The fact that correlations proved to be consistently high between LPI and TOEIC-LC strongly suggests that both tests are, in fact, effectively measuring the common ability to understand and use spoken English.

In conclusion, the present study has demonstrated that trends in TOEIC/LPI relationships exhibit consistent patterns of association, relationship, and predictability. The research has found that TOEIC-TOTAL scores, followed closely by TOEIC-LC scores alone, are significantly related to Language Proficiency Interview ratings. We anticipate that additional research will confirm and expand the findings here, and will verify that these relationships hold true for the larger TOEIC population. We view this report as a guidepost for further research, and as one of many possible areas open to future investigation.

Information about additional and prospective TOEIC research reports is available from your local TOEIC representative.



TOEIC®

Test of English for International Communication

85071–09780 • S103M20 • 275686 • Printed in U.S.A.