



*Research
Report*

TOEIC®/LPI Relationships in Academic and Employment Contexts in Thailand

Kenneth M. Wilson

Supalak Komarakul Na Nagara

Robert Woodhead

TOEIC[®]/LPI Relationships in Academic and Employment Contexts in Thailand

Kenneth M. Wilson

ETS, Princeton, NJ

Supalak Komarakul Na Nagara and Robert Woodhead

Center for Professional Assessment, Thailand

April 2004

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:

www.ets.org/research/contact.html



Abstract

The Test of English for International Communication (TOEIC®) is an ETS-developed measure of proficiency in English as a second language (ESL) or foreign language (EFL), designed to assess aspects of English language listening comprehension (LC) and reading comprehension (RC). In general samples of native speakers of Japanese, Korean, French, Spanish, and Arabic, respectively, scores on the TOEIC have been found to be strongly related to performance on the Language Proficiency Interview (LPI) procedure, a widely recognized direct measure of speaking proficiency. In this exploratory study, TOEIC/LPI relationships in subgroups (e.g., by gender, educational level, organizational setting at time of assessment) were assessed within a general national/linguistic sample—native speakers of Thai who are involved in operational assessments under the aegis of the TOEIC representative agency in Thailand, the Center for Professional Assessment. Observed relationships were relatively strong and positive in the total sample and within each of the subgroups studied. Observed coefficients were noticeably larger in subgroups of females than in corresponding subgroups of males and for a sample of university students than for examinees tested in places of employment. Means of observed LPI rating for the respective subgroups were consistent with expectations based on TOEIC scores. Further research is needed to assess the extent to which the subgroup findings may tend to be generalizable in the Center (Thailand) context or elsewhere.

Key words: TOEIC® Test, Language Proficiency Interview, gender differences, native Thai speakers, ESL speaking proficiency

Acknowledgments

This study would not have been possible without the collaboration of Supalak Komarakul and Robert Woodhead, codirectors of the Center for Professional Assessment (Thailand), who conducted assessments involving joint administrations of the TOEIC and the LPI procedure in the academic and employment settings described herein. They also contributed essential details regarding the study context, as well as helpful critiques of earlier drafts. Detailed internal reviews of the draft report were provided by Yong-Won Lee and Xiaoming Xi; Dan Eignor provided helpful editorial suggestions; and Kim Fryer provided helpful editorial assistance in preparation of the manuscript. The ETS Research Division provided essential indirect support. These contributions are acknowledged with appreciation.

Table of Contents

	Page
Introduction.....	1
Developing Guidelines for Interpreting TOEIC Scores	2
Previous Studies of TOEIC/LPI Relationships	3
The Present Study	6
Context of the Study	6
About the Study Sample.....	7
Study Data and Description of the Sample.....	8
TOEIC/LPI Correlations.....	13
Further Assessment of Differences by Gender and Educational Status	15
Observed Criterion Performance in Relation to “Expectation”	16
Concluding Observations.....	20
Assessing Generalizability of “Calibration Sample” Guidelines	22
Potential Usefulness of Self-assessment	22
References.....	24
Notes	27
Appendixes	
A - Evaluating Gender-related Differences in Strength of TOEIC/LPI Relationship	30
B - Self-rated vs. TOEIC-predicted Levels of Speaking Proficiency	32

Introduction

The Test of English for International Communication (TOEIC®) is a test of proficiency in English as a second language (ESL). At the request of the Japanese Ministry of International Trade and Industry (MITI), ETS developed the TOEIC and introduced it in Japan in 1979. The goal was to meet the recognized need of international corporations, businesses, and government agencies for a test of ESL proficiency that would provide a basis for making better-informed decisions involving the recruitment, selection, placement, and/or training of employees for positions requiring the use of English (see, for example, Woodford, 1982; ETS, 1985, 1986; The Chauncey Group International [CGI], 1999).

The TOEIC is currently available worldwide. Each year the TOEIC program at ETS develops several new forms of the test for use in scheduled, national Secure Program (SP) administrations offered originally only in Japan, but subsequently extended to Korea. Secure Program administrations, currently available *only* in Japan and Korea, are scheduled and administered by the respective TOEIC representative agencies. The TOEIC program at ETS, however, is responsible for all SP test scoring and score reporting. Test forms developed for SP administrations are used subsequently in Japan, Korea, and all other countries with representative, national TOEIC agencies; in ad hoc Institutional Program (IP) administrations, conducted in workplace or institutional settings; or in similar, ad hoc assessments involving individual or group administrations conducted at the offices of a local TOEIC representative agency. Assessments typically are conducted by English language professionals associated with a TOEIC representative agency or, in some instances, by resident ESL specialists associated with a client organization. However, in all IP administrations, the TOEIC representative agency is responsible for scoring and reporting scores.

The TOEIC Test provides separate sections designed to assess English language listening comprehension (LC) and reading comprehension (RC). Each of these sections includes 100 multiple-choice items. The listening comprehension section is paced by audiotape. A limited reading load is involved because LC items require examinees to choose among several brief, written answer options. Only written stimulus material is involved in the reading comprehension section. Answers to all questions are recorded by examinees on a scannable answer sheet. For reporting purposes, raw number right scores on the respective test sections are converted to an arbitrarily selected scale ranging between 5 and 495. The reported TOEIC Total score is the

simple sum of the reported LC and RC scores, hence ranges between 10 and 990, inclusive. Scores on the TOEIC have been found to be relatively closely related to scores on the TOEFL (e.g., Woodford, 1982; ETS, 1986; Wilson, Berquist, & Bell, 1998; Hemingway, 1999).

Developing Guidelines for Interpreting TOEIC Scores

The TOEIC Test has face validity as a measure of individual and group differences with respect to English language listening and reading comprehension skills. However, scores on the arbitrarily defined scales used in reporting scores on the TOEIC (indeed, any similarly norm-referenced test) do not in and of themselves permit inferences as to how well test takers who present given scale scores tend to be able to “use English,” for example, for communicative purposes. As has been widely recognized (see, for example, Carroll, 1967a, 1967b; Clark, 1975, 1978, 1981; Woodford, 1982; Wilson, 1989; Boldt, Larsen-Freeman, Reed, & Courtney, 1992), the functional (language-use) implications of scores on norm-referenced tests can be established by correlating test scores with clearly defined “language-use” criterion variables. One such language-use criterion has featured prominently in TOEIC validation research (e.g., Woodford, 1982; Wilson, 1989; Wilson & Chavanich, 1989; Wilson & Lindsey, 1999; Wilson & Stupak, 2001). That criterion is the rated level of functional ability to exchange meaning conversationally using English as the language of discourse, based on direct observation of behavior elicited and rated by professional interviewers/raters using the widely recognized Language Proficiency Interview (LPI) procedure.¹

The LPI procedure (sometimes called the Oral Proficiency Interview procedure) has significant face validity. Interview procedures are described in detail elsewhere (e.g., ETS, 1982). The formal interview involves face-to-face, controlled conversations, usually 30 minutes or less in duration, between a nonnative speaker of English (or other target language) and a trained interviewer/rater who is a native speaker of English (or other native language). Apart from the clear face validity of the conversational interview as a procedure for eliciting pertinent samples of linguistic behavior (involving the mutual exchange of meaning conversationally in a target language) perhaps the most distinctive feature of the LPI model is the “quasi-absolute” scale (after Carroll, 1967a, 1967b) that is used to rate the interview performance. For present purposes, it is sufficient to note that the Foreign Service Institute/Interagency Language Roundtable (FSI/ILR) Speaking scale—and each of three other, conceptually comparable, quasi-absolute FSI/ILR scales (ILR hereafter)—has a total of 11 behaviorally anchored (described)

scale points. The ILR scales have six basic or base levels ranging from Level 0 (no proficiency in English) through Level 5 (proficiency equivalent to that of an educated native speaker) and five behaviorally defined intermediate levels (labeled, in the present instance, Speaking 0 plus, Speaking 1 plus . . . Speaking 4 plus, Speaking 5.² For general interpretive purposes, it is useful to know that interviewers/raters assign “plus” ratings to subjects who are deemed to have exhibited the linguistic behavior described for a particular base level, *plus* many of the aspects of behavior associated with the next higher base level, but fall short of meeting fully, requirements for placement at the next higher base level. For purposes of statistical analysis in TOEIC validation research, base-level and plus ratings are assigned to an 11-point numerical scale such that S(speaking)-0 = 0, S-0+ = .5, S-1 = 1, S-1+ = 1.5, . . . , S-4 = 4, S-4+ = 4.5, and S-5 = 5.

Previous studies, to be reviewed briefly below, have provided empirical evidence regarding the strength and consistency of TOEIC/LPI correlation and the consistency in agreement between means of distributions of observed LPI ratings and corresponding distributions of TOEIC-estimated LPI-criterion ratings, within and across samples of TOEIC takers from several nationally and linguistically diverse populations of EFL users/learners. The present study extends previous research by providing novel empirical evidence regarding relationships between TOEIC Test scores and LPI rating (that is, predictor/criterion relationships) in subgroups of examinees classified by gender, current employment status (employed vs. student) and context of employment, within a particular language group (Thai); also regarding the extent to which, in a general sample of TOEIC examinees in the Thai setting, observed average LPI rating by TOEIC-score interval tends to be consistent with expectation, when the latter is based on previously developed regression-based guidelines.

Previous Studies of TOEIC/LPI Relationships

In the validity study conducted to coincide with the introduction of the TOEIC test (Woodford, 1982), TOEIC scores were found to correlate strongly with LPI ratings (also ratings of samples of reading and writing, using ad hoc procedures) in a sample of Japanese EFL users/learners—observed coefficients centered at approximately the .8 level. Given the potential value of ILR-scale rated LPI performance as a generally applicable “speaking proficiency” criterion a second validity study (Wilson, 1989) was undertaken to assess the strength and stability of TOEIC/LPI (predictor/criterion) relationships (and corresponding predictive equations) within and across samples of TOEIC takers from nationally and linguistically diverse test-taking

populations—several samples of Japanese TOEIC takers ($N = 285$, total), and smaller samples from several other national/linguistic populations ($N = 108$, total: France, Mexico, Saudi Arabia, in descending order by sample size).

The familiar regression model was employed to calibrate (reference, link) scores on the arbitrarily defined TOEIC score scale to directly interpretable levels of ILR-scaled LPI performance (treated as a “context-independent” language use criterion measure) in the samples involved.³ In the combined TOEIC/LPI “calibration sample” ($N = 393$) criterion-related validity coefficients for TOEIC LC, RC, and Total were .74, .68, and .74, respectively. Corresponding coefficients within each of the several, component national/linguistic samples conformed to the foregoing pattern. Moreover, average levels of ILR-scaled (LPI assessed) speaking proficiency by TOEIC score levels tended to be consistent across the several nationally and linguistically differentiated samples. These findings, combined with other evidence reviewed in detail in the report (Wilson, 1989), suggested as a working hypothesis that interpretive guidelines reflecting the regression of LPI-rating on TOEIC scores in the combined “calibration” sample would tend to be generalizable to similarly selected samples from other major national/linguistic populations in the larger TOEIC testing context. Of course, the need for empirical assessment of this working hypothesis was also emphasized.

Subsequent studies were conducted to extend assessment of the stability of prediction involving calibration-sample regression equations. These were (a) a study involving data for a TOEIC-screened (Total score over 500) sample of native Thai speakers in training as cabin attendants for a major Thai-based international airline (Wilson & Chavanich, 1989) and (b) a study involving data for a general sample of Korean TOEIC takers—educated, adult, native speakers of Korean who were primarily academically trained EFL users/learners (Wilson & Stupak, 2001). Data for the Korean sample were analyzed separately and also comparatively, in combination with data for the TOEIC-screened Airline sample, as well as data for the samples of native speakers of Japanese, French, Spanish, and Arabic that comprised the “TOEIC/LPI calibration sample.” In the Airline/Thai and TOEIC/Korea samples, observed TOEIC/LPI coefficients tended to be relatively strong, albeit somewhat lower than those for the calibration sample itself. The lower observed TOEIC/LPI coefficients appeared to be explicable in terms of primarily statistical considerations (restriction of range effects).

Findings regarding the degree of agreement between the average observed LPI rating and the average predicted value based on calibration sample regression equations (see Wilson & Stupak, 2001, Table 4) indicated that (a) the mean criterion (LPI) rating in the TOEIC-screened Airline sample was somewhat higher than was predicted from TOEIC scores (e.g., mean residual = .18 on the 11 point, numerical conversion of the ILR scale, for predictions based on TOEIC LC score) and, (b) in the sample of Korean TOEIC takers, calibration-sample regression equations overestimated LPI performance by approximately half a level (e.g., mean residual = -.41, when prediction was based on TOEIC LC score).

Underestimation in the Airline context appeared to be attributable to the fact that members of the sample were not only positively screened by performance on the TOEIC but, by inference, were also self-selected positively in terms of self-perceived EFL speaking proficiency (as aspirants to positions clearly calling at some point for ability to communicate orally with English-speaking international travelers). The observed average overestimation of LPI performance in the sample of Korean TOEIC takers did not appear to be explicable in terms of “sample selection effects” or other primarily statistical and/or psychometric consideration. Based on lines of reasoning and supporting empirical evidence described in detail in the study report, it was concluded, as a strong working hypothesis, that such a pattern might be expected to obtain more generally in similarly selected samples of EFL users/learners assessed by the TOEIC representative agency in Korea.

While suggesting considerable stability in TOEIC/LPI relationships across samples of TOEIC takers from several nationally and linguistically diverse populations of EFL users/learners, the findings that have been reviewed also reinforce the recognized importance of studying TOEIC/LPI relationships in each national/ linguistic subpopulation being served by TOEIC programs. In evaluating the interpretive implications of observed correlations between TOEIC scores and LPI rating (or any other external criterion variable), and corresponding regression-based guidelines, it is important, of course, to keep in mind that TOEIC scores *only* provide a basis for actuarially circumscribed, empirically verifiable, “predictive inferences” about the average level and range of ILR-scaled speaking proficiency that examinees at designated score levels, from similarly selected samples, may be expected to exhibit.³

The Present Study

As indicated earlier, the present exploratory study extends analysis of TOEIC/LPI relationships to subgroups within a single language group (Thai), namely, subgroups defined by gender, educational level attained (high school graduate, vocational school graduate, enrolled university student, university graduate) academic versus nonacademic status, organizational context of assessment, and TOEIC Test repeater vs. nonrepeater status. Joint TOEIC/LPI assessments were undertaken at a major Thai university (University, hereafter), and in four corporate settings—involving employees of an international oil corporation and employees in three international hotel settings. A brief description of the study context is provided before presenting detail regarding the study sample and data.

Context of the Study

The data for these samples were collected by the Center for Professional Advancement (known until recently as TOEIC Thailand, or TOEIC Center Thailand), the agency that has been responsible for the administration of TOEIC-related affairs in Thailand for more than a decade. TOEIC scores and LPI ratings, along with responses of examinees to several background questions, were collected in operational assessments conducted by the Center for Professional Advancement (Center, hereafter) for corporate clients (business and educational) in that country. Center staff includes ESL professionals certified in use of the LPI procedure. Joint TOEIC/LPI assessments such as those that provided the data for this study may be undertaken at the request of current or prospective TOEIC clients who need guidelines for making informed, English-proficiency-related recruitment, hiring, placement, and/or promotional decisions involving employees in or aspiring to positions calling for the use of English-language skills. In settings such as those involved in the present study, TOEIC scores and LPI ratings may be obtained for representative samples of EFL users/learners in particular organizational settings in order (a) to generate evidence regarding the relationship between TOEIC scores and functionally described (LPI assessed) levels of English speaking proficiency among current employees, and (b) to estimate the average level and range of ILR-defined speaking proficiency that typically is associated with designated levels and ranges of TOEIC scores. Generally speaking, the interpretive implications of such empirical evidence are evaluated by involved Center staff in direct consultation with individuals designated by the client, typically individuals concerned with ESL-proficiency-related personnel matters. Members of the Center staff, of course, also are

cognizant of findings regarding levels of ILR-scaled speaking proficiency that have tended to associated with TOEIC score ranges, based on research such as that reviewed briefly in previous sections of this report.

About the Study Sample

Students at the University who were assessed with both the TOEIC and the LPI procedure were beginning, third-year “Communication Arts” majors, preparing to undertake specific *major-related* English courses. Joint administration of the TOEIC and the LPI procedure was undertaken to establish general guidelines for the operational use of TOEIC scores as a partial basis for placing such students for purposes of EFL instruction.

Joint TOEIC/LPI data were also obtained in four different corporate settings, namely, an international oil corporation (Oil Corp) and three international hotel settings (Hotel A, Hotel B, and Hotel C)—all either current or prospective TOEIC users. In each of these and other similar corporate contexts in the Center setting, management needs to maintain a cadre of “English proficient” employees—EFL users/learners capable of meeting the requirements of positions calling for the use of English as an operational language in interactions involving native and multiple nonnative speakers of English. As indicated earlier, questions of placement for purposes of EFL instruction guided the TOEIC/LPI assessments at the University. In the corporate contexts, the assessments were undertaken to help management evaluate the current level and range of TOEIC-related EFL proficiency in representative cadres of employees and work toward the establishment of operational position-based English guidelines based on TOEIC performance. In conducting the assessments, an effort was made to include employees in a representative range of positions with some level of English use involvement. TOEIC scores and LPI ratings were obtained independently in all settings. Supalak Komarakul and/or Robert Woodhead, codirectors of the Center, conducted the TOEIC and LPI administrations that generated the data involved in this study.

Generally speaking, EFL users/learners in these and other Center assessments may be characterized as “academic” EFL learners—learners who begin to study English as a foreign language in school after reaching the age of native-language maturity (that is, after the first six years of formal schooling) and take formal EFL courses during the grade 7 to grade 12 period (typically involving three hours of study per week, per academic year). Those who go on to postsecondary education typically take EFL courses during the first two years of study. Thus,

university graduates will have studied English for at least eight years of formal EFL instruction (grades 7–12, plus the first two years of university study) typically involving three hours per week, per academic year. This constitutes a common core of formal exposure to the study of English as a foreign language for those with university level education. Those who have completed secondary education, only, typically will have had six years of curriculum-based EFL instruction. In samples of employed EFL users/learners, the amount of formal EFL instruction, of course, will normally tend to vary with level of education. Data were not available for background variables such as amount and type of supplementary (a) EFL instruction (e.g., in a commercial English language institute or an in-company program), (b) self-study, and/or (c) experience in an English-speaking environment, and the like. According to Woodhead (2003, personal communication), programs of EFL instruction for grades 1–12 have been initiated recently in Thailand.

Study Data and Description of the Sample

Three hundred and seven individuals were assessed. Data elements included TOEIC scaled scores (listening comprehension (LC) and reading comprehension (RC)), scaled 5 to 495, and Total, scaled 10 to 990) and LPI rating (numerical values assigned to ILR Speaking- or S-levels were 0 for S-0, or no proficiency, .5 for S-0 plus, 1.0 for S-1, 1.5 for S-1 plus . . . 5.0 for Level 5). These data were available for all examinees. With limited exceptions, responses were also available for background questions regarding (a) the number of times the TOEIC was previously taken (0, 1, 2, 3, 4, or more); (b) gender (female = 2, male = 1); (c) organizational affiliation (four employment contexts and one university setting); and (d) educational status (highest educational level attained at time of assessment).

Figure 1 shows the composition of the Center sample by gender, educational level, specific organizational affiliation, organization type (academic vs. nonacademic), and experience with TOEIC. Figure 2 highlights variation in gender distribution and TOEIC experience across organizational settings. Males slightly outnumbered females in the total sample. However, females substantially outnumbered males at University; females and males were about equally represented among university graduates and in the three hotel samples. Data not shown indicate that the oil corporation (all male) sample included mostly vocational/technical school graduates. Eighty percent of the total sample (including all 123 enrolled undergraduates) was made up of first-time TOEIC takers (T1); most of the remainder were taking TOEIC for the second time.

Means and standard deviations of distributions of TOEIC scaled scores and LPI ratings are shown in Table 1 for the total Center sample and subgroups by educational level, organizational type, specific organization, gender, and TOEIC experience (first-time test takers vs. repeaters—from one to four or more previous administrations). For a comparative perspective, descriptive statistics are also shown for the basic TOEIC/LPI calibration sample (Wilson, 1989), which included examinees tested in Japan, France, Mexico, and Saudi Arabia. Gender differences appear to be slight; differences between the academic (university student) and nonacademic (employee) samples are somewhat more noticeable.

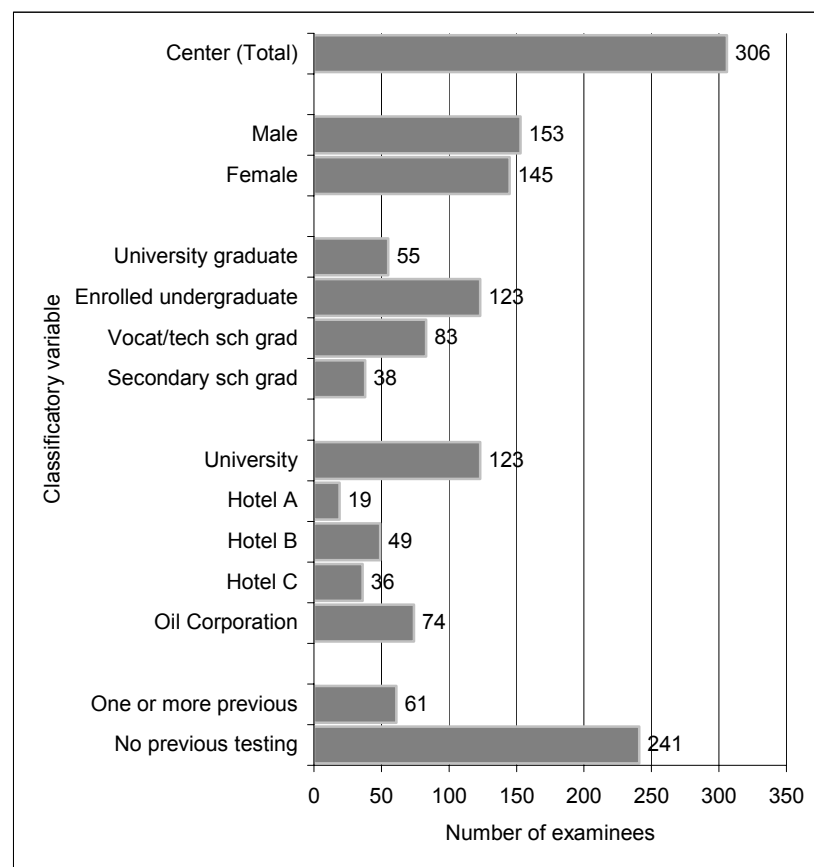


Figure 1. Number of examinees in subgroups defined by gender, educational level, assessment context, and experience with the TOEIC test.

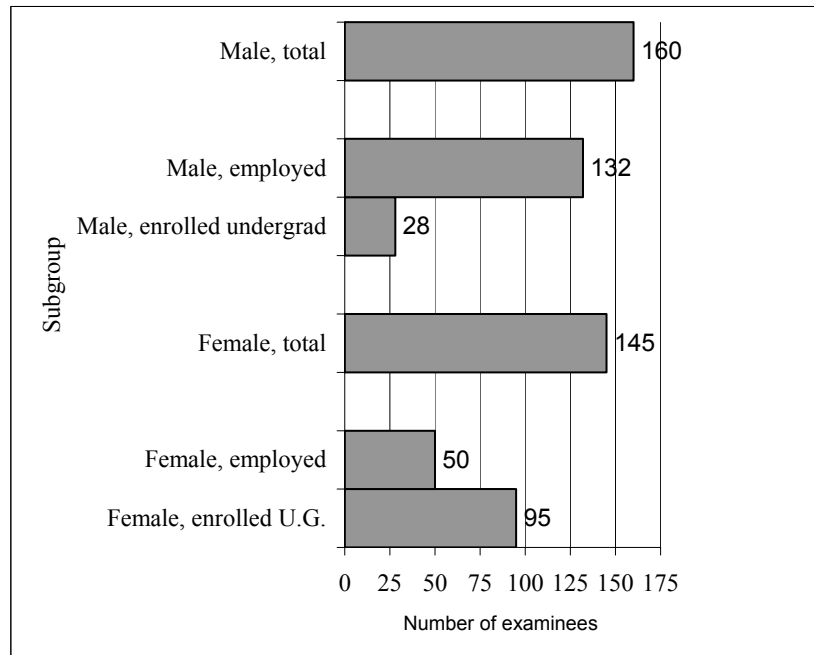


Figure 2. Gender differences with respect to status (employed versus enrolled undergraduate student) at time of assessment.

Table 1

Descriptive Statistics for the Total Center Sample and Designated Subgroups and for the Basic TOEIC/LPI Calibration Sample

Group	N	TOEIC-LC		TOEIC-RC		TOEIC-Total		LPI	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
<i>TOEIC/LPI</i>									
Calibration sample ^a	393	325	96	305	89	630	177	1.91	.68
Center (total)	306	218	84	187	83	405	151	1.24	.44
Educational level									
Secondary sch.	38	175	81	116	65	291	136	1.04	.41
Voc/Tech sch.	83	240	74	158	70	398	135	1.30	.45

(Table continues)

Table 1 (continued)

Group	N	TOEIC-LC		TOEIC-RC		TOEIC-Total		LPI	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Enrolled UG ^b	123	209	82	234	72	443	144	1.21	.44
Univ./Grad.	55	241	88	176	77	416	158	1.35	.43
Organization type									
Academic ^b	123	209	82	234	72	443	144	1.21	.44
Nonacademic	182	225	84	155	74	380	150	1.27	.45
Organization									
C Univ.	123	209	82	234	72	443	144	1.21	.44
Oil Corp ^c	74	243	76	165	73	408	142	1.27	.46
Hotel H	36	169	71	114	52	282	116	1.22	.45
Hotel R	49	237	92	163	73	400	155	1.22	.45
Hotel S	19	247	70	181	94	427	154	1.47	.42
Gender									
Female	145	217	87	212	82	428	154	1.27	.45
Male	160	219	81	164	77	383	145	1.22	.44
Female/Acad.	95	214	83	238	68	452	140	1.24	.43
Male/Acad.	28	191	81	221	84	412	156	1.13	.46
Female/Nonacad.	50	226	94	162	83	389	169	1.35	.50
Male/Nonacad.	132	225	80	152	70	377	143	1.23	.43
Number of previous TOEIC testings ^d									
None	241	215	84	194	86	409	154	1.23	.46
One or more	61	234	82	161	62	397	135	1.27	.40

Note. The basic calibration sample ($N = 393$) was made up of native speakers of Japanese ($N = 285$), French ($N = 56$), Arabic ($N = 10$) and Spanish ($N = 42$). TOEIC/LPI calibration sample 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 K. M.

Wilson, 1989, (TOEIC Research Report No. 1; ETS RR-89-39), p. 44. Copyright 1989 by ETS.

^a Data from Wilson (1989). ^b “Enrolled undergraduate” and “academic” samples are identical, except for label. ^c All academics were first-time test takers, hence constituted approximately one half (123 of 241) of the subgroup with no previous TOEIC Testings. ^d All Oil Corp employees were males, most of whom were vocational-technical school graduates.

The primary academic vs. nonacademic contrast in TOEIC performance alluded to above is shown in Figure 3, which shows the percentage of each subgroup scoring 200 or above on the TOEIC LC and RC, respectively. The percentage scoring above 200 on reading is noticeably greater for academics than for nonacademics; for listening comprehension, differences are in the opposite direction, albeit somewhat less dramatically so than for reading comprehension. When TOEIC Total score is considered, it may be seen in Table 1 that the mean for academics was noticeably larger than that for nonacademics (443 versus 380) whereas, for TOEIC LC, the direction of mean differences was reversed (209 versus 225 for academics versus nonacademics). Mean LPI rating for academics and nonacademics, generally, differed only slightly (1.21 versus 1.27), thus tending to parallel corresponding findings for mean LC (or mean Total) rather than those for mean RC.

Similarly, gender differences in mean LPI rating also appear to be relatively slight (1.27 for females versus 1.22 for males), as do corresponding differences in mean LC (217 versus 219). Differences by educational level are somewhat similar in the sense that subgroups with higher performance in LC appear to be somewhat higher with respect to LPI rating. A more systematic assessment of differences by gender and educational level with respect to criterion (LPI) performance, after controlling for TOEIC performance, will be reported later.

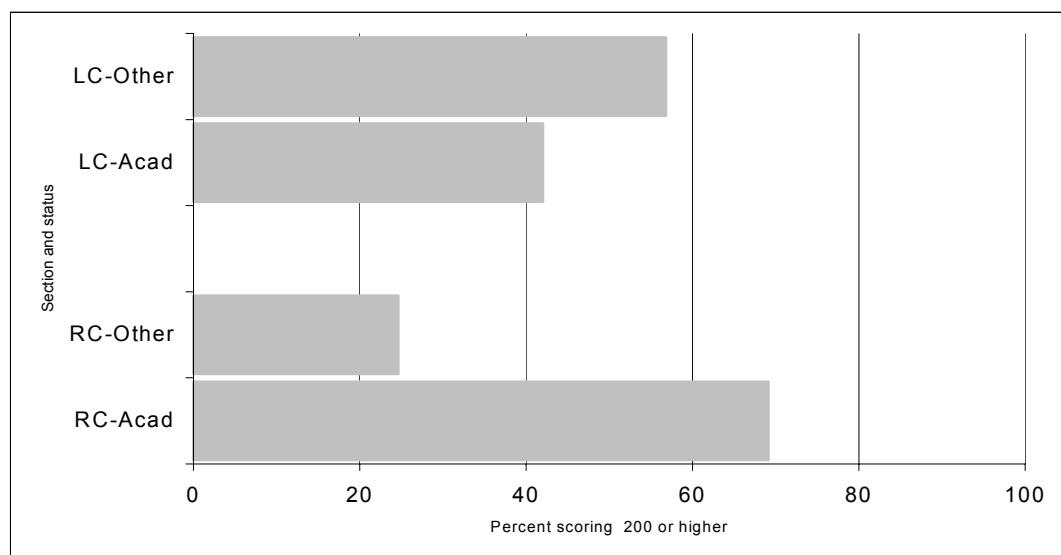


Figure 3. Academics (enrolled university students) perform relatively better on TOEIC RC than do examinees in nonacademic settings, while the opposite is true for TOEIC LC.

TOEIC/LPI Correlations

Table 2 shows zero-order correlations between TOEIC scores (LC, RC, and Total) and LPI criterion rating for the total Center sample and subgroups, including those for which descriptive statistics were shown in Table 1. Coefficients involving TOEIC scores and LPI rating were relatively strong and positive across all classifications in the table. It is noteworthy that the observed LC/LPI coefficients generally tended to be larger than were the corresponding RC/LPI coefficients, and roughly comparable to the corresponding TOEIC Total/LPI coefficients. This is a logical, theoretically consistent pattern given the underlying functional linkage of development of speaking proficiency to the development of aural skills. It also seems noteworthy that observed coefficients generally were consistently somewhat larger for subgroups of female examinees than for corresponding subgroups of male examinees; also for the predominately female, academic (enrolled university students) subgroup than for the nonacademic (employed) subgroup in which males outnumbered females. This pattern of differences does not appear to be explicable in terms of attenuation due to differential restriction of range on the variables involved—for example, standard deviations shown in Table 1 for subgroups by gender and employment status, respectively, appear to be generally comparable. Explication of the pattern of systematically higher observed TOEIC/LPI relationships for females than for males, and for academics than for nonacademics is beyond the scope of this inquiry. And, of course, further research is needed to establish the extent to which this particular gender-related pattern may tend to be generalizable in the Center context, or to TOEIC testing contexts elsewhere.

Table 2

Zero-order Correlation of TOEIC Scores, Gender, and Educational Level, Respectively, With LPI Rating in the Total Center Sample and Designated Subgroups

Group	N	Correlation with LPI		
		TOEIC score		
		LC	RC	Total
		<i>r</i>	<i>r</i>	<i>r</i>
Total center sample	307	.66	.51	.65
Calibration sample	393	.74	.68	.74

(Table continues)

Table 2 (continued)

		Correlation with LPI TOEIC score		
Group	<i>N</i>	LC	RC	Total
		<i>r</i>	<i>r</i>	<i>r</i>
Center subgroups				
Organization type				
Academic ^a	123	.76	.67	.76
Nonacademic	182	.59	.59	.62
Oil Corp. ^b	74	.53	.57	.58
Hotel A	39	.56	.57	.60
Hotel B	50	.72	.73	.77
Hotel C	19	.75	.53	.66
Educational level				
Secondary sch.	39	.55	.58	.60
Voc/Tech sch.	84	.54	.53	.57
Enr. undergrad. ^a	123	.76	.67	.76
Univ./Grad.	57	.58	.60	.62
Gender				
Female	145	.78	.59	.75
Acad./T1 ^a	95	.79	.71	.81
Nonacad./T1	22	.91	.88	.93
Nonacad./T2+	26	.61	.58	.63
Male	160	.54	.45	.54
Acad /T1 ^a	28	.65	.55	.63
Nonacad/T1	34	.52	.53	.56
Nonacad/T2+	94	.53	.51	.55
Previous TOEIC experience				
None (T1) ^a	241	.69	.52	.67
T1 Academic ^a	123	.65	.55	.63
T1 Nonacademic	118	.63	.61	.65
Repeater (T2+)	60	.55	.56	.59

Note. Sum of *N*s for compound categories may not equal corresponding category *N* due to missing background data.

^aAll “academics” were undergraduate students at University, and all were first-time (T1) TOEIC takers—no repeaters (T2+). ^bAll Oil Corp. employees were males.

Further Assessment of Differences by Gender and Educational Status

Among other things, results shown in Table 1 and Table 2 suggest only relatively modest differences in LPI performance by gender and educational status and that individual differences in LPI rating tended to be more closely related to TOEIC LC than to TOEIC RC. Also, the observed coefficient involving LPI and TOEIC Total (including TOEIC RC) tended to be no larger, and in some instances was slightly smaller than, the corresponding LC/LPI coefficient. A systematic assessment was made of the observed subgroup differences in average LPI rating, with control for performance on TOEIC LC and TOEIC Total, respectively. First, members of the sample were classified by educational status *and* gender. Six subgroups representing highest level of education attained and gender were defined as follows: *male/<university*; *female/<university*; *male/university student*; *female/university student*; *male/university graduate*; *female/university graduate*. In the total Center sample, LPI rating was then regressed on TOEIC LC and TOEIC Total, respectively, and the corresponding residuals (labeled LCres and TOTres) were computed for each individual. Mean values of these residuals (reflecting, respectively, differences between the observed LPI rating for each examinee, and the value predicted from the TOEIC LC score and TOEIC Total score) were computed for each of the subgroups named above. Results of an analysis of differences among the six subgroups with respect to mean residuals are shown in Table 3. The table provides mean residual values for the respective equations (LCres for predictions based solely on TOEIC LC and TOTres for predictions based on TOEIC Total, including RC) for the designated subgroups and results of corresponding one-way analyses of variance.

It seems clear from the data in Table 3 that, regardless of the TOEIC score involved, there was relatively close agreement at the mean between distributions of observed and predicted LPI rating across all the subgroups involved. For example, only two mean residual values exceeded .10 (on the 11-point ILR scale) in absolute value. At the same time, however, the results indicate a slightly better fit between subgroup means for the observed and predicted LPI rating based on the measure of listening comprehension alone (TOEIC LC), rather than for a prediction based on TOEIC total (which, of course, includes the TOEIC RC score). Thus, the ILR-defined level of LPI-assessed speaking proficiency in the Center sample may tend to be indexed somewhat more validly by level of performance on the measure of listening comprehension (TOEIC LC) than by level of performance on the more comprehensive measure

(TOEIC Total). This is a theoretically consistent finding. At the same time, from a pragmatic perspective, it is important to recognize that the findings also suggest that the validity of inferences about ILR-defined level of speaking proficiency from TOEIC Total score tends to be comparable to the validity of corresponding inferences based on TOEIC LC.

Table 3

Mean Residuals for Subgroups Defined by Educational Level and Gender

Subgroup	Mean residual		N
	LCres ^a	TOTres ^b	
Male/university	-.04	.04	103
Female/<university	.08	.16	20
Male/university student	-.02	-.13	28
Female/university student	-.01	-.09	97
Male/university grad	.00	.07	29
Female/university grad	.05	.10	28
Total CENTER sample ^c	.00	.00	305

Note. Prediction based on TOEIC Listening Comprehension versus prediction based on TOEIC total (CENTER total sample).

^a $F = .666$, $df (5,299)$, $p = .650$, $Eta^2 = .011$ (for predictions based on LC).

^b $F = 4.493$, $df (5,299)$, $p = < .0007$, $Eta^2 = .070$ (for predictions based on Total).

^c Mean residuals are .00 by definition.

Observed Criterion Performance in Relation to “Expectation”

The primary purpose of this exploratory study was to assess the strength and consistency of TOEIC/LPI relationships across subgroups defined by gender, educational level, loci of assessment, and TOEIC-repeater status. The findings reviewed above provide evidence of relatively strong and theoretically consistent patterns of correlation between TOEIC scores and LPI ratings in designated subgroups; evidence suggesting that average levels of ILR-scaled speaking proficiency for examinees presenting given TOEIC scores, on the average, tend to be about the same regardless of subgroup membership. Specifically, it appears that regression-based guidelines based on data for the total Center sample—guidelines for actuarial inferences

regarding average levels and ranges of ILR-scaled speaking proficiency associated with TOEIC score ranges—may tend to be generalizable (for all practical assessment purposes) across similarly selected samples of university students, employed EFL users/learners, and so on, in the Center setting.

It is also of interest, theoretically and pragmatically, to know the extent to which the observed average level of ILR-scale rated, LPI assessed speaking proficiency in the Thai sample here under consideration (mean statistically scaled ILR level = 1.24) tends to be consistent with expectation for examinees with a corresponding average scores on the TOEIC (Total = 408 or LC = 218, in the present instance) when *expectation is based on patterns of TOEIC/criterion relationships observed in samples of TOEIC takers included in the basic TOEIC/LPI calibration study* (Wilson, 1989). To address this particular issue, two estimated LPI ratings were computed for each member of the Center sample ($N = 307$). One estimated value reflected the regression of LPI rating on TOEIC LC in the TOEIC/LPI “calibration” sample, and the other estimated value reflected the corresponding regression involving TOEIC Total score (see Wilson, 1989, endnote 27 and related material in the text for the equations involved and detail regarding their development and validation). Residual values (LPI rating minus predicted LPI rating) were computed, as were the corresponding means. The mean residual values are shown in Table 4, which also shows intercorrelations and descriptive statistics for the variables involved in the Center sample ($N = 307$), in the “calibration” sample ($N = 393$), and in the combined sample ($N = 700$), respectively. Further interpretive perspective is provided by Figure 4, which permits evaluation of degree of similarity for trends in mean level of ILR Speaking proficiency by TOEIC Total score interval in the Center sample as compared to corresponding trends in the basic calibration sample. In evaluating trends for the Center sample in Figure 4, it is useful to know that data points for the three highest TOEIC Total intervals represent data for a total of only 13 examinees (N s of 8, 2, and 3, respectively, for intervals 700–795, 800–895, and 900–995).

Several aspects of the findings appear to be noteworthy. First, by inspection of data in Table 4, it is evident that the average observed level of LPI-rated, oral English proficiency in the total Center sample was predicted relatively accurately from equations reflecting the regression of ILR-scaled LPI rating S-ratings in the “calibration” sample. The estimated mean values shown in Table 4 (1.34 and 1.29, for equations involving TOEIC LC and TOEIC Total, respectively) may be compared with the observed mean LPI rating (1.24); corresponding mean residuals

(observed minus estimated values, not shown in the table) are -.10 and -.02. These are very modest differences.

Table 4

Relationship Between ILR-scaled Level of Speaking Proficiency and TOEIC Scores in Designated Samples: Intercorrelations and Descriptive Statistics

Variable	LC	RC	Total	ILR-Speak	Mean	SD
	<i>r</i>	<i>r</i>	<i>r</i>	<i>r</i>		
Center sample ^a						
TOEIC LC	--	.64	(.91)	.66	218.0	83.8
TOEIC RC		--	(.90)	.51	186.8	82.8
TOEIC total			--	.65	404.8	150.8
ILR-Speaking (LPI)				--	1.24	0.44
Regression-based estimates						
ILR-S (Est. from LC)					1.34	0.38
ILR-S (Est. from Total)					1.26	0.38
Calibration sample ^b						
TOEIC LC	--	.83	(.96)	.74	325.5	95.8
TOEIC RC		--	(.95)	.68	304.6	89.1
TOEIC Total			--	.74	630.1	176.7
ILR-Speaking				--	1.91	.68
Combined sample ^c						
TOEIC LC	--	.82	(.96)	.78	278.3	105.4
TOEIC RC		--	(.95)	.72	252.7	104.6
TOEIC Total			--	.78	531.0	200.5
ILR-Speaking				--	1.62	.68

Note. “ILR-scaled speaking level” is as assessed directly by trained interviewer/raters using the Language Proficiency Interview procedure.

^a *N* = 307. ^b *N* = 393. ^c *N* = 700.

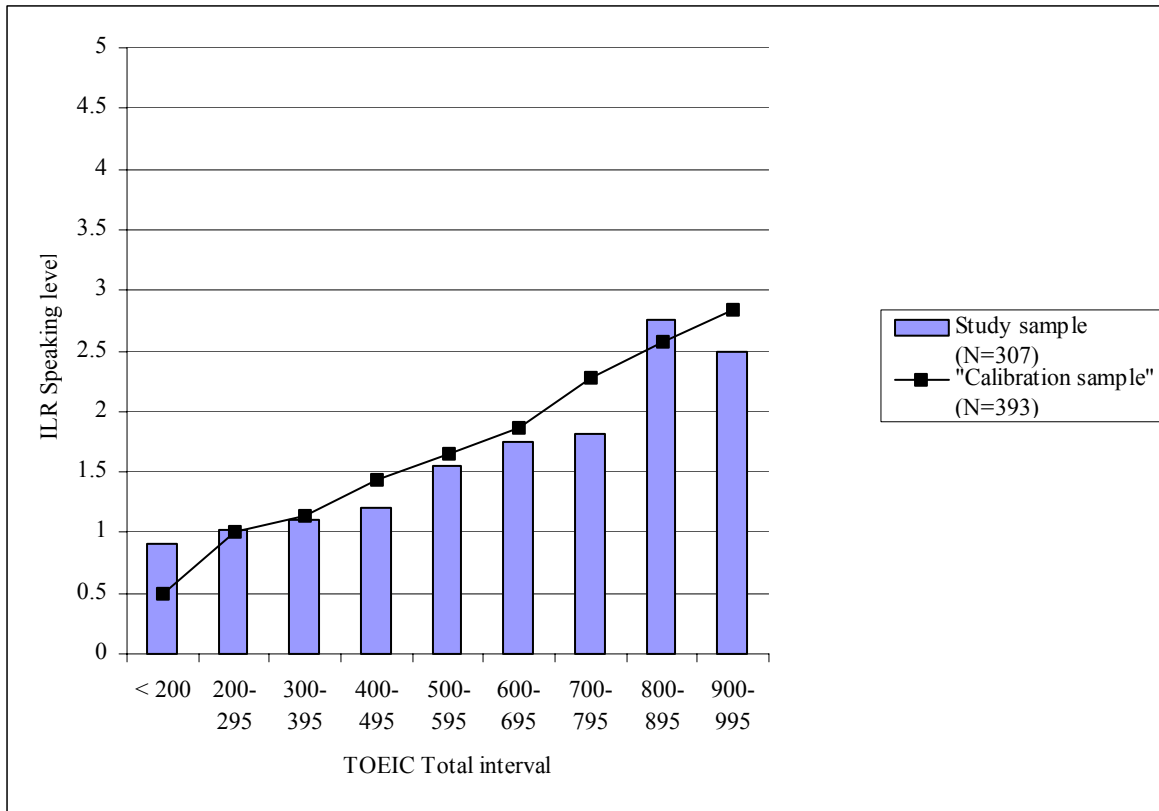


Figure 4. Trends in average level of ILR-defined speaking proficiency by TOEIC Total score interval tend to be similar for the designated samples of EFL users/learners.

The findings suggest that average rated level of speaking proficiency in the Center sample, characterized by an observed TOEIC Total mean of 405, tended to be quite consistent with expectation for TOEIC takers assessed elsewhere with TOEIC Total scores averaging in the vicinity of 400. For the purpose of setting TOEIC/LPI guidelines in the Center context, the data suggest that TOEIC/LPI guidelines based on TOEIC Total score are likely to be as meaningful (useful/valid) as would be comparable guidelines based on TOEIC LC.

Second, generally lower observed criterion-related validity coefficients for TOEIC scores in the Center sample than in the combined appear to be consistent with restriction-of-range effects. Note, for example, that standard deviations for all the variables involved are systematically smaller in the Center sample, than are the corresponding values in the consolidated sample. The fact that TOEIC/criterion correlations were higher in the combined sample than were those observed for the Center sample, only, is consistent with the increase in

range resulting from combining the two samples with divergent, positively correlated means on the predictor and criterion variables involved (see corresponding statistics in Table 4).

Third, combining data for the lower scoring Center sample (less than 10% with TOEIC Total scores of 600 or higher) with data for the higher scoring calibration sample (*averaging over 600 on TOEIC total*) tends to permit better informed evaluation of trends in TOEIC/criterion relationships *at the lower end of the the TOEIC scale* than has been possible heretofore. It would appear to follow, as a strong working hypothesis, that the findings for the combined sample, shown in Table 4, tend to provide a meaningful basis for better informed estimates of trends in TOEIC/LPI relationships for Center-assessed Thai examinees *at the upper end of the TOEIC scale* than do findings for the study sample that includes only a very limited number of examinees at higher TOEIC score levels. As noted above, only 10% of the Center sample scored 600 or higher on the TOEIC. In evaluating trends in Figure 4 it is useful to recall that the *Center data points for the three highest TOEIC Total score intervals are based, respectively, on 8, 2, and 3 cases, only.*

Concluding Observations

The findings that have been reviewed tend to confirm, and extend to the Center setting, previous findings bearing on the strength and consistency of TOEIC/LPI relationships generally.

Results of the exploratory subgroup analyses suggest, as indicated above, that inferences regarding TOEIC/LPI relationships based on general samples of TOEIC takers assessed by the Center may tend to be generalizable to subgroups based on gender, employment status, and so on. At the same time, certain of the findings suggest that TOEIC/LPI relationships may tend to be “moderated” to some extent (that is, may tend to differ systematically in level) by gender, and possibly other variables (e.g., repeater vs. nonrepeater status, student samples vs. employed samples), respectively. These findings, of course, have not been replicated. However, the observed differences in relative size of coefficients involved reach what appears to be a noteworthy level, especially in gender-related analyses, and they do not appear to be explicable in terms of differences in relative degree of selectivity (see, for example, standard deviations for the variables involved, shown in Table 1). Moreover, a similar pattern of gender-related differences has been reported (Wilson & Graves, 1999) for “EFL test”/interview and “EFL test”/essay relationships in samples of native Japanese speaking students—recent secondary school graduates being assessed for EFL placement purposes at Temple University Japan.

One tentative hypothetical rationale, advanced as a basis for further research, is that females may tend to be more generally “amenable to instruction” (e.g., on the average, may tend to work harder, be more inclined to conform to instructional “demands”) than are males. Generally speaking, correlations between tests that measure aspects of the functional ability represented by a given criterion variable should tend to be higher in more highly motivated subgroups than in less highly motivated subgroups. Substantial evidence supportive of this hypothesis, as it applies to academic predictors and academic criteria (especially, grade point average) in U.S. college admission contexts, is available (e.g., Willingham & Cole, 1997; Donlon, 1984; Angoff, 1971). A brief summary of this evidence and the Temple University Japan findings alluded to, above, is provided in Appendix A.

EFL/ESL acquisition/use background differences possibly also could be involved. For example, EFL learning/use backgrounds of females in the study (predominately enrolled university students) plausibly may have tended to be more uniform than those of males (a majority of whom were employees of organizations—hotels and an oil corporation—with international business connections. Thus, it is possible that lower coefficients for males may have been due to effects associated with individual, experiential differences in, for example, opportunity to use English in communicative interaction, possibly differentially impacting performance on the two measures involved. Such differences would more likely tend to be present in samples of employed EFL users/learners than in student samples. There is evidence that more exposure to use of English in an English dominant environment tends to be associated with higher TOEIC performance (e.g., CGI, 1996; Wilson, 1989). Further research is needed to shed light on the “differential experience effects” hypothesis as put forward here.

In any event, questions regarding the degree of generalizability of such gender-related findings and the tenability of the foregoing tentative explanatory rationales might be considered in the design of further research in the Center context, and elsewhere in the TOEIC testing context. More generally, in future assessments involving TOEIC/criterion relationships, it would be useful to collect background data needed to control for differences in English language learning/use background. Such differences may not be a complicating factor in academic contexts, and the strength of observed TOEIC/LPI relationships at University suggests the potential usefulness of the TOEIC Test as a basis for estimating the corresponding criterion distribution—that is, the distribution of EFL speaking proficiency as defined by the ILR scale

that tends to be associated with given distributions of TOEIC scores—in other academic settings in Thailand. To the extent that objectives of EFL instruction in Thailand include the development of aural/oral as well as reading proficiencies, the TOEIC test would appear to be a potentially useful instrument not only for EFL placement purposes, but also for the purpose of estimating levels and ranges of ILR-scaled speaking proficiency that tend to characterize samples of EFL user/learners who have completed designated amounts of EFL instruction (e.g., unselected samples of secondary school graduates generally).

Assessing Generalizability of “Calibration Sample” Guidelines

More generally, it is believed that the findings reviewed above (see especially Table 4 and Figure 4 and related discussion) provide additional evidence supportive of the working proposition (Wilson, 1989) that TOEIC/LPI calibration-sample guidelines will tend to provide a useful basis for estimating average levels and ranges of ILR-defined speaking proficiency likely to be associated with corresponding levels and ranges of TOEIC scores, in samples of adult, educated EFL users/learners, such as those who are likely to take the TOEIC, in diverse, national/linguistic settings served by the TOEIC program. The findings also reinforce the recognized importance of obtaining empirical evidence bearing on the validity of this working proposition in all such settings.

Potential Usefulness of Self-assessment

It may not be feasible to conduct formal LPI ratings in each national/linguistic setting being served by the TOEIC program. However, for TOEIC testing contexts in which it is not feasible to conduct formal Language Proficiency Interviews, there is evidentiary support (e.g., Wilson, 1989; Wilson & Lindsey, 1999) for hypothesizing that useful preliminary working estimates of average levels and ranges of ILR-defined speaking proficiency, associated with observed TOEIC score ranges, may be obtained through self-assessments based on rating schedules directly linked to the ILR scale—that is, schedules with scale points anchored by descriptions adapted from the corresponding ILR Speaking scale descriptions. Specifically, self-ratings according to one such schedule (developed by TOEIC/ETS staff members) have been found to be relatively strongly correlated with TOEIC scores in samples comprised primarily of native speakers of French or German (Total $N = 937$) assessed by the TOEIC representative agency in Switzerland (Wilson & Lindsey, 1999). Zero-order correlation coefficients for TOEIC

LC, RC, and Total scores versus the LPI-scaled self-rating were, respectively, .75, .70, and .75, in the total sample (see corresponding coefficients for TOEIC scores versus formal LPI rating as observed in the calibration sample, shown in Table 4). Moreover,

... average self-placed level on the ILR-referenced (Speaking) scale tended to be generally consistent with expectation for individuals with the observed average level of performance on the TOEIC—that is, the self-assessed ILR-scaled levels, on the average, tended to be consistent with regression-based (calibration sample) guidelines for predicting ILR-scaled LPI rating from TOEIC score(s)” (Wilson & Lindsey, 1999, p. 19).

See Appendix B for detail on trends by TOEIC score interval in average self-rated level of ILR/LPI-scaled speaking proficiency as compared to corresponding trends involving average LPI-rated proficiency, as observed in the calibration sample.

More generally, other research findings also indicate relatively strong and pragmatically meaningful levels of validity for information obtained through variously elicited self-assessments of second-language proficiency (e.g., Oskarsson, 1978; Wangsotorn, 1980; Oscarson, 2001; Clark, 1981; Clark & Swinton, 1979; Ingram, 1985; Hilton, Grandy, & Kline, 1985; Tannenbaum, Rosenfeld, & Breyer, 2000). That evidence is supportive of the proposition that, given stimulus material that provides appropriate foci for organizing and expressing their intuitions, educated, adult EFL users/learners are able to provide valid information about their own level of proficiency in English (or other target language) and that they will tend to do so under nonthreatening, non-high-stakes conditions.

In connection with the suggestion, above, that useful estimates of levels and ranges of ILR-scaled speaking proficiency might be gained through use of the ILR-referenced, global rating schedule alluded to, it is important to emphasize that such estimates need to be evaluated in terms of expectation based on available guidelines, such as those alluded to herein. And, of course, self-assessments should not be thought of as substitutes for valid, external measures that are needed when critical English-proficiency related decisions about particular individuals must be made.

References

- Angoff, William H. (Ed.). (1971). *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.
- Boldt, R. F., Larsen-Freeman, D., Reed, M. S., & Courtney, R. G. (1992). *Distributions of ACTFL ratings by TOEFL score ranges* (TOEFL Research Report no. 41). Princeton, NJ: ETS.
- Carroll, J. B. (1967a). *The foreign language attainments of language majors in the senior year: A survey conducted in United States colleges and universities* (Final Report, Contract OE-4-14-048). Cambridge, MA: Harvard University Graduate School of Education.
- Carroll, J. B. (1967b). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals*, 1(2), 131–151.
- Clark, J. L. D. (Ed.). (1978). *Direct testing of speaking proficiency: Theory and application*. Princeton, NJ: ETS.
- Clark, J. L. D. (1981). Language. In T.S. Barrows (Ed.), *College students' knowledge and beliefs: A survey of global understanding*. New Rochelle, NY: Change Magazine Press.
- Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 10–28). Arlington, VA: Center for Applied Linguistics.
- Clark, J. L. D., & Swinton, S. S. (1979). *An exploration of speaking proficiency in the TOEFL context* (TOEFL Research Rep. No. 4 and ETS RR-79-8). Princeton, NJ: ETS.
- Donlon, Thomas F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.
- ETS. (1997). *SLEP test manual*. Princeton, NJ: Author.
- ETS. (1986). *Guide for TOEIC users*. Princeton, NJ: Author.
- ETS. (1982). *ETS oral proficiency testing manual*. Princeton, NJ: Author.
- Hemingway, M. (1999). *English proficiency tests: A comparative study*. Princeton, NJ: The Chauncey Group International.
- Hilton, T. L., Grandy, J., & Kline, R. G. (with Stupak, S. A. & Woodford, P. E.). (1985). *The oral language proficiency of teachers in the United States in the 1980's: An empirical study*. Princeton, NJ: ETS.

- Ingram, D. E. (1985). Assessing proficiency: An overview of some aspects of testing. In K. Hyltenstam & M. Pieremann (Eds.), *Modelling and assessing second language acquisition* (pp. 215–276). San Diego, CA: College-Hill Press.
- Jones, R. L. (1979). The FSI oral interview. In B. Spolsky (Ed.), *Advances in language testing: Series I--Some major tests* (pp. 104–115). Arlington, VA: Center for Applied Linguistics.
- Lowe, P., & Stansfield, C. W. (Eds.). (1988). *Second language proficiency assessment: Current issues*. Englewood Cliffs, NJ: Prentice-Hall Regents.
- Oscarson, M. (2001). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Language testing and assessment: Encyclopedia of language and education* (vol. 7). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Oskarsson [Oscarson], M. (1978). *Approaches to self-assessment in foreign language learning*. Oxford: Pergamon Press.
- Saegusa, Y. (1985). Prediction of English proficiency progress. *Musashino English and American Literature*, 18, 165–185.
- Tannenbaum, R., Rosenfeld, M., & Breyer, J. (2000). *Linking TOEIC scores to self-assessments of English-language abilities: A study of score interpretation*. Unpublished draft report.
- The Chauncey Group International. (1999). *TOEIC user guide*. Princeton NJ: Author.
- The Chauncey Group International. (1996). *TOEIC report on test-takers worldwide*. Princeton, NJ: Author.
- Wangsotorn, A. (1980). Self-assessment in English skills by undergraduate and graduate students in Thai universities. In J. A. S. Read (Ed.), *Directions in language testing* (pp. 240–246). Singapore: Singapore University Press.
- Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 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 Testing context* (TOEIC Research Rep. No. 1, ETS RR-89-39). Princeton, NJ: ETS.
- Wilson, K. M. (1999). *Validating a test designed to assess ESL proficiency at lower developmental levels* (ETS RR-99-23). Princeton, NJ: ETS.

- Wilson, K. M., Berquist, A., & Bell, I. (1998). *Guidelines for comparing performance on two tests of ESL proficiency: The TOEIC Test and the TOEFL*. Unpublished draft report.
- Wilson, K. M., & Chavanich, K. (1989). *Further evidence of stability in TOEIC/LPI relationships across diverse samples*. Unpublished draft report.
- Wilson, K. M., & Lindsey, R. (1999). *Validity of global self-ratings of ESL speaking proficiency based on an FSI/ILR-referenced scale* (ETS RR-99-13). Princeton, NJ: ETS.
- Wilson, K. M., & Graves, K. (1999). *Validity of the secondary level english proficiency test at Temple University Japan* (ETS RR-99-11). Princeton, NJ: ETS.
- Wilson, K. M., & Stupak, S. S. (2001). *Overestimation of LPI rating in the TOEIC Testing context in Korea: Search for explanation* (ETS RR-01-15). Princeton, NJ: ETS.
- Woodford, P. E. (1982). The Test of English for International Communication (TOEIC). In C. Brumfit (Ed.), *English for international communication* (pp. 61–72). New York: Pergamon Press.

Notes

- ¹ See Wilson, 1989 (pp. 6-10, Sections 1 and 2), for detailed development of the rationale for using LPI performance as a surrogate for workplace or other real-life observations of ability to exchange meaning conversationally using English as the language of discourse. See Lowe and Stansfield (1988), ETS (1982), Jones (1979), Clark (1978); Carroll (1967a, 1967b), for historical perspective and detail regarding the “LPI model”—that is, the interview, and corresponding, behaviorally anchored scale for rating interview performance developed by the Foreign Service Institute (FSI) of the U.S. Department of State for the direct assessment of second-language speaking proficiency. The LPI model and conceptually comparable FSI-developed models were subsequently adopted for use by several U.S. government agencies known collectively as the Interagency Language Roundtable or ILR.
- ² Modified versions of the FSI/ILR scales—versions that provide, for example, finer discriminations at lower ILR-scale levels (Level 0-plus and Level 1)—have been developed by the Association of College Teachers of Foreign Languages (ACTFL) in collaboration with ETS (ETS, 1982) for use in assessing the proficiency of foreign language students. See Boldt, et al. (1992) for a report of research, similar conceptually to that involved in TOEIC/LPI research here under consideration, designed to link level of performance on the TOEFL to ACTFL-scale descriptors for levels of proficiency in *listening, reading, and writing* by correlating TOEFL scores with teachers’ ratings—teachers not trained in use of the scales for rating purposes—of the respective proficiencies according to the corresponding ACTFL scales, based on naturalistic, classroom observations. Relatively strong correlations were observed in each of several university samples. An informal, ad hoc, modified version of the ACTFL *Speaking* scale was used to obtain teachers’ ratings of English speaking proficiency for a sample of secondary-level international students studying or planning to study in the United States (Wilson, 1999). Whether rendered by native-English-speaking EFL teachers or their nonnative-English-speaking counterparts, teachers’ ratings according to the modified speaking scale—based on naturalistic, classroom observation—were relatively strongly correlated with students’ scores on an experimental, “easier” version of the TOEIC. Findings such as the foregoing, suggest that the ACTFL descriptors involved (and by logical extension the ILR scale descriptors from which they were derived), formal or as appropriately modified, tend to portray valid differences in level of functioning in the proficiency domains under consideration; also

that after reading such descriptors, teachers untrained as raters tend to be able to render valid ratings of corresponding proficiencies, based on classroom observation of individual students. Based on findings presented elsewhere (Wilson & Lindsey, 1999) these observations apply as well for *self-ratings by educated, adult EFL users/learners* according to ILR-referenced behavioral descriptions designed to “parallel” (in paraphrased form, for self-rating purposes) corresponding formal descriptors.

³ In a detailed exposition of the rationale for using the regression model (Wilson, 1989, pp. 20–22) for calibrating TOEIC scores to ILR-scaled LPI rating—rather than simple equating models such as those employed with demonstrable interpretive enhancement in national assessments conducted by Carroll (1967a, 1967b) and Hilton et al. (1985)—the following points were made and are reiterated here for background:

It is preferable to employ an approach to linking performance on indirect, norm-referenced tests to levels of performance on functionally scaled criteria that *does not require the assumption of equivalency for working purposes* [italics added].

Given joint distributions of LPI ratings and scores on indirect, norm-referenced measures for a given sample, it is clear that a regression-based calibration model does not require a priori assumptions about the organization of second-language skills, or the psychometric or theoretical equivalence of the measures involved.

At the same time, a regression-based approach to this problem obviously need not be atheoretical. By regressing LPI ratings on measures of listening and reading skills, for example, it is possible to assess the hypothesis of greater correspondence between second-language speaking and listening skills than between speaking and reading skills, while at the same time establishing and evaluating statistically meaningful criterion estimation rules.

In this connection, it is noteworthy that in the regression model, but not in the equating model, the scales of the indirect measures involved are referenced (calibrated) to the functionally scaled criterion variable according to linkage rules that vary directly with the observed level of association between the indirect measures and the functional criterion in calibration samples. Thus, regression-based estimates of criterion behavior are more

explicitly “delimited” than are inferences that derive from the application of simple equating models. And, the usefulness of the regression model for purposes of criterion-referencing is well established.

As a general proposition, regressing a functionally scaled criterion variable of the type represented by LPI performance, on indirect, norm-referenced test scores in samples of test takers from defined populations of second-language user/learners, can be expected, a priori, to provide evidence that permits an informed evaluation of the patterns of relationships among the measures under consideration from both theoretical and practical perspectives, statistically delimited inferences (e.g., estimates, with standard errors), from scores on the indirect test, about probable level of defined language-use behavior, for individuals in samples from the test-taking population involved, and inferences regarding the probable level and dispersion of oral language proficiency in the test-taking population, according to the directly interpretable LPI scale. (pp. 21–22)

Appendix A

Evaluating Gender-related Differences in Strength of TOEIC/LPI Relationship

In evaluating the finding that observed TOEIC/LPI coefficients tended to be larger in samples of females than in corresponding samples of males, it seems noteworthy that generally similar patterns were also found to obtain in a sample of recent postsecondary school graduates in Japan (Wilson & Graves, 1999). The graduates were being tested for purposes of EFL instruction at Temple University Japan, where all instruction is conducted in English. For present purposes it is sufficient to note that the placement battery included the Secondary Level English Proficiency (SLEP) test, an interview conducted according to locally developed procedures for assessing speaking proficiency, and a writing sample (using locally developed prompts). The SLEP test (e.g., ETS, 1997) has sections designed to measure listening comprehension (LC) and reading comprehension (RC), respectively. A total score reflecting overall performance is also reported. Scores on the SLEP are relatively closely related to scores on the TOEFL. As may be seen in Table A1, the observed SLEP/interview and SLEP/essay coefficients were larger in the sample of female students than in the sample of male students.

Table A1

Correlation of Designated SLEP Test Scores with Interview and Essay Ratings, Respectively, at Temple University Japan, by Gender

Males	Males ^a		Females ^b		Difference (F-M)	
	Inter	Essay	Inter	Essay	Inter	Essay
	<i>r</i>	<i>r</i>	<i>r</i>	<i>r</i>		
LC	.60	.56	.65	.59	.05	.03
RC	.49	.57	.56	.60	.07	.03
Total	.61	.62	.66	.65	.05	.03

Note. Based on an ad hoc analysis of data described in Wilson & Graves, 1999.

^a *N* = 840. ^b *N* = 778.

No other comparable gender-related findings involving norm-referenced and direct measures of English proficiency were located during the course of this study. However, generally speaking, in academic settings in the United States, the academic performance of female students has tended to be more predictable from college admission measures than has the academic performance of male students. For example, for men and women in liberal arts and

general programs, based on more than 100 studies involving a general grade point average (GPA) criterion that permitted separate analysis of data for men and for women, validity coefficients for the SAT sections (Angoff, 1971, Table 5.6, p.127) varied systematically by gender. Validity coefficients tended to be systematically higher in samples of female students (median, $r = .41$ for SAT-V) than in corresponding samples of male students (median, $r = .33$). A similar pattern was evident, as well, when grades in English courses, only, constituted the criterion (Angoff, 1971: Table 5.15, p.140). A similar differential-validity pattern was reported by Donlon (1984) after reviewing a similarly developed body of empirical evidence (see Table 8.14, p.156, and related discussion).

Such systematic differences by gender in the strength of association between a measure of “scholastic aptitude” and a criterion reflecting differences in actual academic accomplishment, assuming equal intervening conditions, may tend to reflect gender-related differences in degree of “accomplishment oriented” motivation or effort. In this connection, Willingham and Cole (1997), in evaluating gender-related differences such as those reviewed above, observed that, as compared to males: “Females tend to have stronger academic work habits and more positive indicators of attitude and effort. . . . [T]hey tend to show stronger interests in academic and intellectual endeavor. . . .” (pp. 349–350).

It is plausible that such gender-related differences may tend to obtain in academic settings outside the United States. In any event, based on the Center findings, the generally similar findings in a sample of Japanese college students, and comprehensive evidence from U.S. academic settings, it seems plausible that similar gender-related differences may tend to obtain in diverse national/linguistic/cultural contexts. Empirical answers to questions about the degree of generality of such findings in Thailand, Japan, and elsewhere can be obtained by conducting similar studies in a variety of TOEIC-use settings.

Appendix B

Self-rated vs. TOEIC-predicted Levels of Speaking Proficiency

As indicated in this report, there is reason to posit as a working hypothesis that educated, adult, target (second/foreign) language learners/users have the capacity to provide valid information about their own proficiency in using English and that they will tend do so (under nonthreatening, non-high-stakes conditions) given appropriate self-rating models such as the global scale developed by TOEIC/ETS staff to obtain self-ratings of speaking proficiency on a scale designed to “parallel” the formal ILR Speaking scale (see Wilson, 1989, for historical perspective). Figure B1 provides illustrative evidence bearing directly on the validity of that hypothesis as assessed in a relatively large sample ($N = 937$) of TOEIC examinees tested under auspices of TOEIC Consulting SA, the representative TOEIC agency in Switzerland (Wilson & Lindsey, 1999). In Figure B1, the average level of ILR-defined speaking proficiency by TOEIC Total interval appears to be about the same for TOEIC takers in Switzerland, based on ILR-scaled self-assessment (see dotted line), as that observed for TOEIC takers in the “calibration sample,” based on ILR-rated performance in Language Proficiency Interviews (see bars).

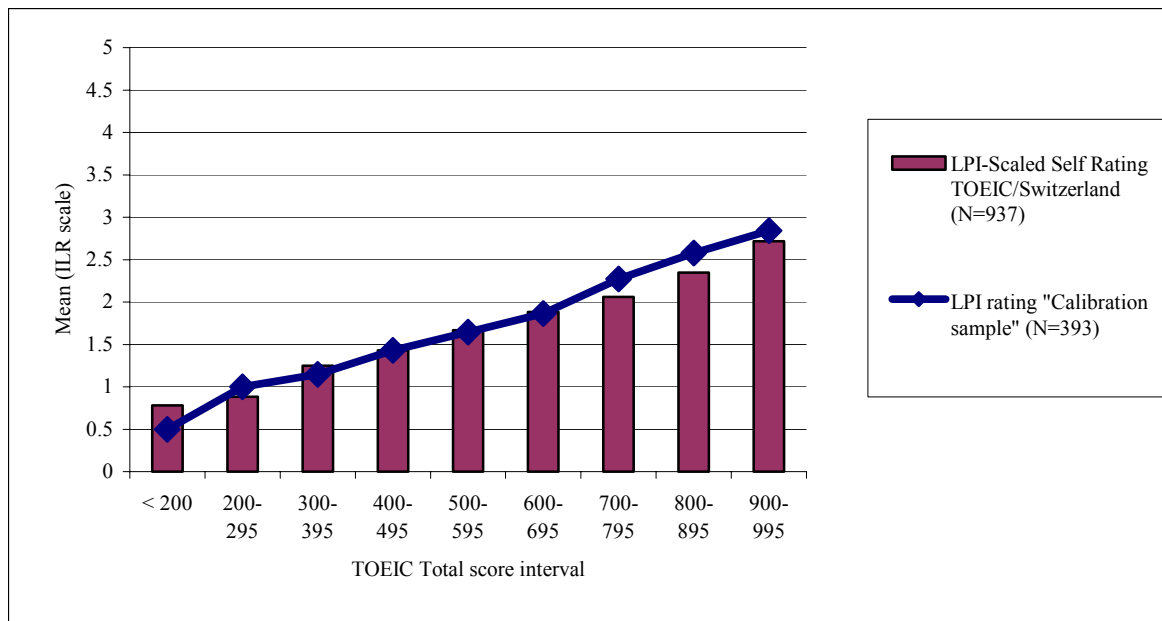


Figure B1. Average level of ILR-defined speaking proficiency by TOEIC total interval.

Note. Based on an ad hoc analysis of data described in Wilson and Lindsey (1999).

As described in detail elsewhere (Wilson & Lindsey, 1999), the educated, adult EFL users/learners involved (primarily native speakers of French or German) provided self-ratings of speaking proficiency in English according to a global schedule designed to parallel the ILR Speaking scale (by paraphrasing the corresponding formal descriptions in a form designed for use in self-assessment). Thus “LPI-scaled” self-ratings were generated. Mean LPI-scaled self-rating was computed by TOEIC Total interval for the Swiss sample (vertical bars in Figure B1) for comparison with mean LPI rating by TOEIC Total interval in the calibration sample (represented by the trend line in Figure B1; see also Figure 4 and related text, herein).

The line in Figure B1 can be thought of as representing the “predicted” or “expected” level of speaking proficiency, based on calibration-sample data. The bars can then be thought of as representing average “criterion” performance, as defined by self-assessment rather than direct assessment. It is evident that there was very close agreement at every TOEIC Total interval between the observed average, ILR-scaled, *self-assessed* level of speaking proficiency and the corresponding average predicted (estimated, expected) LPI-rated ILR-scaled speaking proficiency. Similarly close agreement has also been found to obtain in comparisons involving ILR-scaled self-assessments and LPI ratings when both measures were available for members of the same sample (see Wilson & Lindsey, 1999, Figure 1, adapted from findings reported by Hilton et al., 1985).

Findings of this nature attest to the potential value of the ILR-scaled self-assessment model as a basis for research concerned with estimating the level and range of ILR-defined proficiencies that tend to be associated with TOEIC scores in diverse, national/linguistic TOEIC-use settings.