ETS TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

The Redesigned TOEIC® Listening and Reading Test: Relations to Test Taker Perceptions of Proficiency in English

Donald E. Powers, Hae-Jin Kim, and Vincent Z. Weng

January 2010

The TOEIC® test assessment was developed to measure the ability to listen and read in English, using a variety of contexts from real-world settings. Recently, a redesign of the test was undertaken, in order to better align test questions with everyday workplace language scenarios and to provide test takers with more information about their listening and reading proficiency levels.

Although many of the question types are the same as in the previous version of the TOEIC Listening and Reading test, there are some significant modifications. These modifications were undertaken in order to articulate more exactly various aspects of the construct. Specifically, the listening section now has:

- fewer questions that involve photographs,
- *both* recorded *and* written questions to assess understanding of conversations and short talks,
- fewer *individual* questions and more *sets* of questions to assess the understanding of conversations, and
- a range of different English accents, as spoken in the United States, Great Britain, Canada, and Australia.

The new reading section has the following major changes:

- the elimination of questions that require the recognition of grammatical errors,
- the addition of text completion questions,
- an increase in the number of reading comprehension questions, and
- the inclusion of sets of questions based on two interrelated passages.

In summary, these changes are intended to align the test more closely with theories of communicative competence (see, for example, Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, et al., 2000). For instance, the use of interrelated passages now actually *requires* the use of strategies to comprehend and connect information in order to answer some of the questions. In addition, the Redesigned TOEIC Listening and Reading test is believed to better reflect international business communication styles and real language contexts. The redesign is thought to be a valid measure of international communication today.

The effort described here was intended to provide evidence of the validity of the Redesigned TOEIC Listening and Reading test as a measure of English language proficiency. ETS hoped to accomplish this by establishing the relationship between scores on the Redesigned TOEIC Listening and Reading test and test-taker reports of their ability to perform selected, everyday language tasks in English.

Method

In order to accomplish our objective, we assembled and administered (in the summer of 2007) a self-report *can-do* inventory to TOEIC test takers in Japan and Korea immediately after they had taken the test. The inventory included a series of common language tasks ("can-do" statements) for both listening (24 tasks) and reading (25 tasks). Tasks were adapted from previous studies (e.g., Duke, Kao, & Vale, 2004; Powers, Roever, Huff, & Trapani, 2003; Tannenbaum, Rosenfeld, Breyer, & Wilson, 2007). Tasks were translated from English into Japanese and Korean (and also back-translated), so as to convey, to the extent possible, the same meaning as the original text. The translations were performed by ETS field representatives in Japan and Korea, with subsequent reviews provided by ETS

staff and an external consultant. Directions, which were also translated into Japanese and Korean, were as follows:

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

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

Two putatively parallel forms of the inventory were assembled, each with approximately half of the can-do statements. Both the test and the inventory were administered via computer, with each form of the inventory administered to a random half of the total examinees.

Results

Test scores and can-do reports were obtained from 7,292 test takers from Japan and 3,626 from Korea. Nearly 5,400 participants completed one form of the inventory, and approximately 5,500 completed the other form.

Table 1 shows the correlations between TOEIC listening and reading scores and test takers' assessments of their ability to perform the can-do tasks, as defined by the sum of responses to (a) all listening can-do tasks and (b) all reading can-do tasks. For both of the listening can-do forms, the Cronbach alpha reliability estimate was .94. For the reading can-do forms, these estimates were .95 and .94. For the TOEIC test scores, the KR20 reliability estimates were .93 for reading scores and .92 for listening scores. As can be seen from Table 1, the correlations between TOEIC Listening and Reading test scores are high (.80 for the sample taking one form of the inventory and .81 for those taking the other form), as are the correlations between the listening and reading can-do reports (.80 for one form and .77 for the other). Can-do listening reports and TOEIC Listening test scores correlate relatively strongly (.53 for each form). The corresponding correlation between reading can-do reports and TOEIC Reading test scores is only slightly lower (.47 and .48). (Individually, the correlations of reading statements with TOEIC Reading test scores range from .08 to .48, with a median of .39. For listening statements, the correlations range from .30 to .50, with a median of .44.) The correlations between *reading* can-do reports and TOEIC Listening test scores (.47 and .46), and between *listening* can-do reports and TOEIC Reading test scores (.43 and .45), are slightly lower, thus suggesting some *discriminant* validity of the two TOEIC test scores, even though they correlate highly with one another, as do the listening and reading can-do reports. This result is confirmed when correlations are corrected for attenuation, as the correlation between TOEIC Listening and Reading test scores is estimated to be very high (.86 to .88), but not perfect. The same is true for the listening and reading can-do reports, whose disattenuated correlations are .82 to .85. The effect of disattenuating the correlations between can-do reports and TOEIC test scores was to increase the correlations systematically, by .03 to .04.

To allow a better indication of how test performance relates to each can-do activity individually, we have also presented (in Table 2 for listening and Table 3 for reading) item-by-item results, ordered by the degree of difficulty of each can-do task (mean response on the 5-point scale). Because the samples that completed the two can-do forms were randomly equivalent, we have merged the results into two tables—one for listening and one for reading. The percentages shown are the proportions of test takers at each of several score intervals who said that they could perform the task either easily or with little difficulty. An arbitrary TOEIC test score range of 55 points was chosen for each interval, except for the lowest one. For this lowest interval, a range of 130 points was used because there were very few test takers in this lowest score range and the percentages would have been

extremely unstable with any fewer test takers. Table entries are shaded in various colors, according to magnitude, in order to enable the reader to see at a glance the overall pattern of results. The mean shown for each item is the average response to the item on the 1-to-5 response scale. The correlation of each individual can-do item with either the TOEIC Reading or Listening test score is also shown in the two tables.

TABLE 1

Correlations Among Can-Do Self-Assessments and TOEIC Test Scores

Measure	M (SD)	TOEIC listening score	TOEIC reading score	Can-do listening task	Can-do reading task
Can-Do Form A					
TOEIC score					
Listening	325.1 (86.8)	1.00	(.86)	(.57)	(.50)
Reading	273.3 (91.6)	.80*	1.00	(.46)	(.50)
Can-do task					
Listening	38.3 (9.3)	.53*	.43*	1.00	(.85)
Reading	43.4 (9.7)	.47*	.47*	.80*	1.00
Can-Do Form B					
TOEIC score					
Listening	322.3 (86.7)	1.00	(.88)	(.57)	(.49)
Reading	272.0 (93.9)	.81*	1.00	(.48)	(.51)
Can-do task					
Listening	38.1 (8.9)	.53*	.45*	1.00	(.82)
Reading	42.2 (9.0)	.46*	.48*	.77*	1.00

Note. Ns are approximately 5,400 for Form A and approximately 5,500 for Form B. Numbers in parentheses above the diagonal have been corrected for attenuation.

****p < .001.***

TABLE 2

Percentages of TOEIC Test Takers, by Listening Score Level, Who Indicated That They Could Perform Various English Language Listening Task Either Easily or With Little Difficult

Task:	Listening Score Level							M	SD	Corr. with TOEIC test listening scaled score
	5-135	140-195	200-255	260-315	320-375	380-435	440-495			
Understand the days of the week and the months of the year	73	82	85	88	89	93	95	4.45	0.76	.20
Understand simple questions in social situations (e.g., "How are you?" and "Where do you live?")	57	61	74	82	90	95	97	4.35	0.84	.37
Understand someone who is speaking slowly and deliberately about his or her hobbies and interests	35	38	57	70	79	89	94	3.98	0.87	.43
Understand someone speaking slowly and deliberately, who is giving me directions on how to walk to a nearby location	30	37	51	64	74	84	91	3.86	0.90	.40
Understand some memorized words and phrases	43	43	52	59	65	75	85	3.77	0.84	.28
Understand directions about what time to come to a meeting and where it will be held	20	23	41	55	66	80	91	3.71	0.94	.46
Understand a person's name when she or he gives it to me over the phone	31	34	47	57	61	69	80	3.70	0.98	.30
Understand a salesperson when she or he tells me prices of various items	16	28	35	49	60	77	89	3.67	0.95	.45
Understand a person in social situations talking about his/her background, family, or interests	11	16	22	31	46	66	82	3.34	0.98	.49
Understand public announcements that are broadcast	18	17	22	28	41	54	72	3.28	0.90	.39
Understand explanations about how to perform a routine task related to my job	2	11	13	21	36	52	76	3.14	0.95	.50
Take a telephone message for a co-worker	9	15	14	21	37	55	75	3.10	1.03	.49

Task:	Listening Score Level							M	SD	Corr. with TOEIC test listening scaled score
	5-135	140-195	200-255	260-315	320-375	380-435	440-495			
Understand play-by-play descriptions on the radio of sports events that I like (e.g., soccer, baseball)	14	11	15	19	21	29	50	2.89	0.97	.32
Understand a co-worker discussing a simple problem that arose at work	6	7	9	15	25	43	68	2.88	1.05	.50
Understand the main ideas in news reports broadcast on the radio or TV	7	11	9	14	23	33	53	2.87	0.95	.40
Understand an explanation given over the radio of why a road has been temporarily closed	6	4	8	14	20	37	63	2.81	1.08	.49
Understand lines of argument and the reasons for decisions made in meetings that I attend	6	6	7	11	17	34	60	2.77	1.01	.48
Understand a discussion of current events taking place among a group of persons speaking English	5	7	6	10	18	29	53	2.70	0.98	.46
Understand headline news broadcasts on the radio	6	7	8	10	14	24	46	2.69	0.95	.42
Understand a client's request made on the telephone for one of my company's major products or services	5	8	6	12	20	29	51	2.65	1.03	.46
Understand discussions in a workplace meeting with several people	6	3	3	8	13	25	51	2.64	0.98	.49
Understand an extended debate on a complex topic such as equality in the workplace	0	6	5	6	12	22	45	2.60	0.92	.46
Understand the details of a fast-breaking news event on the radio	0	6	5	8	12	19	39	2.60	0.91	.40
Understand a complex presentation or demonstration in an academic or work-related setting	6	3	4	6	8	14	32	2.36	0.97	.41
N for score interval	46-49	304-336	937-1,047	1,312-1,324	1,252-1,284	830-830	673-694			

Note. In previous, similar can-do studies, a less conservative coding may have been used; here, we coded only "can do easily" and "can do with little difficulty" as evidence that a person can perform a task. The percentages shown would have been considerably higher if we had used a less stringent standard and included "can do with some difficulty" in the calculations. Table entries (percentages) have been shaded to indicate their magnitude as shown in the key below.

[0-29]	[30-49]	[50-70]	[70-80]	[80-90]	[90-100]

TABLE 3

Percentages of TOEIC Test Takers, by Reading Score Level, Who Indicated That They Could Perform Various English Language Reading Tasks Either Easily or With Little Difficulty

Task:	Reading Score Level							M	SD	Corr. with TOEIC test reading scaled score
	5-135	140-195	200-255	260-315	320-375	380-435	440-495			
Read the letters of the alphabet	91	95	96	95	96	97	99	4.81	0.61	.08
Read and understand a restaurant menu	65	72	79	83	86	87	95	4.22	0.88	.23
Recognize memorized words and phrases (e.g., "Exit," "Entrance" and "Stop")	63	72	78	82	87	92	97	4.16	0.84	.27
Read and understand a train or bus schedule	49	59	70	77	84	90	96	4.00	0.91	.34
Read, on storefronts, the type of store or services provided (e.g., "dry cleaning," "book store")	47	64	69	72	81	90	91	3.95	0.95	.31
Read and understand a simple postcard from a friend	43	58	65	75	83	90	97	3.94	0.92	.37
Read office memoranda in which the writer has used simple words or sentences	36	50	61	72	81	88	96	3.83	0.92	.39
Read and understand traffic signs	40	51	61	68	77	86	90	3.81	0.98	.33
Read tables, graphs and charts	31	40	54	64	73	83	93	3.69	0.94	.38
Read and understand directions and explanations presented in technical manuals written for beginning users	26	34	46	58	66	78	87	3.56	0.97	.40

Task:	Reading Score Level							M	SD	Corr. with TOEIC test reading scaled score
	5-135	140-195	200-255	260-315	320-375	380-435	440-495			
Read and understand simple, step-by-step instructions (e.g., how to operate a copy machine)	24	34	45	55	64	79	90	3.52	0.97	.39
Find information that I need in a telephone directory	23	34	42	52	64	76	89	3.48	1.00	.39
Read and understand a letter of thanks from a client or customer	18	26	39	53	66	81	94	3.45	0.97	.47
Read entertainment-related information (e.g., tourist guides)	15	25	32	45	57	72	85	3.34	0.97	.41
Read information about products (e.g. advertisements)	14	22	29	40	52	68	88	3.27	0.98	.42
Read and understand a travel brochure	10	18	26	38	51	68	86	3.22	0.98	.44
Read and understand an agenda for a meeting	6	14	22	34	46	62	84	3.09	1.00	.48
Read and understand the main points of an article on a familiar topic in an academic or professional journal	10	17	23	30	40	53	79	3.07	0.96	.37
Read English to translate text into my own language (e.g., letters and business documents)	5	12	16	23	36	50	74	2.92	1.01	.39

Task:	Reading Score Level							M	SD	Corr. with TOEIC test reading scaled score
	5-135	140-195	200-255	260-315	320-375	380-435	440-495			
Read and understand a popular novel	7	10	15	23	31	43	67	2.91	0.92	.40
Identify inconsistencies or differences in points of view in two newspaper interviews with politicians of opposing parties	7	8	13	20	30	43	69	2.82	0.97	.43
Read highly technical material in my field or area of expertise with little use of a dictionary	5	10	14	19	27	40	59	2.76	1.01	.38
Read a newspaper editorial and understand its meaning as well as the writer's intent	6	7	10	17	25	35	57	2.71	0.95	.41
Read and understand a proposal or contract from a client	4	7	11	17	25	42	58	2.68	1.01	.44
Read and understand magazine articles like those found in *Time* or *Newsweek*, without using a dictionary	3	5	5	11	19	30	47	2.60	0.91	.42
N for score interval	395-443	845-915	1,179-1,183	1,161-1,187	945-981	604-679	199-202			

Note. In previous, similar can-do studies, a less conservative coding may have been used; here, we coded only "can do easily" and "can do with little difficulty" as evidence that a person can perform a task. The percentages shown would have been considerably higher if we had used a less stringent standard and included "can do with some difficulty" in the calculations. Table entries (percentages) have been shaded to indicate their magnitude as shown in the key below.

[0-29]	[30-49]	[50-70]	[70-80]	[80-90]	[90-100]

To illustrate how to read Tables 2 and 3, consider the first can-do statement in Table 2 ("Understand the days of the week and the months of the year"). For this very easy task, at a TOEIC Listening test score level of 5–135, a total of 73% of study participants responded that they could do the task either easily or with little difficulty. In contrast, at the highest TOEIC Listening test score level (440–495), nearly all participants (95%) felt that they could perform this task easily or with little difficulty. At the intermediate score levels, the percentages [82, 85, 88, 89, and 93] also rise slightly with each higher score level. A much different pattern is apparent for the last, very difficult task listed in Table 2 ("Understand a complex presentation or demonstration in an academic or work-related setting"), for which only 6% of the lowest scoring participants indicated that they could perform this task, in comparison to 32% of the highest scoring participants. (Tables 2 and 3 have been color-coded. Higher percentages have been indicated in darker shades, as indicated in the key at the bottom of the tables. Numbers of examinees at each score level are indicated by the *N*s at the bottom of each score level column.)

An alternative way in which to utilize the table is to use the TOEIC test score level as the reference point and read down any given column. For example, in Table 2, a reader might be interested in the perceptions of test takers at a particular score level, say, a listening score level of 320–375. Reading down this score interval column shows the responses of test takers who scored at this level on the TOEIC Listening test section. For instance, a total of 90% of these test takers indicated that they could "Understand simple questions in social situations" (e.g., "How are you?" and "Where do you live?"). However, for the last, most difficult task listed ("Understand a complex presentation or demonstration in an academic or work-related setting"), only 8% indicated that they could perform this task easily or with little difficulty.

As can be seen, for nearly all of the tasks, higher test performance is associated with a greater likelihood of reporting successful task performance. For the listening statements in Table 2, percentages increase, with few exceptions, for each item with each higher score interval. Of the total number of pairs of percentages[1] that can be compared in the table (24 statements x 6 pairs of comparisons of adjacent percentages for each can-do statement = 144), only 11 do not show increases when going from a lower to the next higher score level. All 11 of these inconsistencies involve very small discrepancies, and all occur at the three lowest score levels, suggesting that the test may be slightly less discriminating at these levels than at other levels, possibly because of the occurrence of chance scores at these levels. For reading tasks (Table 3), there is only one very slight inconsistency of the 150 (25 statements x 6 pairs of comparisons of adjacent percentages for each can-do statement) that are possible.

Discussion/Implications

One kind of evidence that has proven useful in elucidating the meaning, or validity, of language test scores has come from examinees themselves, in the form of self-assessments of their own language skills. Although self-assessments may sometimes be susceptible to distortion (either unintentional or deliberate), they have been shown to be valid in a variety of contexts (see, for example, Falchikov & Boud, 1989; Harris & Schaubroeck, 1988; Mabe & West, 1982), especially in the assessment of language skills (LeBlanc & Painchaud, 1985; Shrauger & Osberg, 1981; Upshur, 1975). For instance, it has been asserted (e.g., Shrauger & Osberg; Upshur) that language learners often have more complete knowledge of their linguistic successes and failures than do third-party assessors. This may

be particularly true for skills like reading and listening, which are not directly observable by third parties.

For this study, a large-scale data collection effort was undertaken to establish links between test taker performance on the Redesigned TOEIC Listening and Reading test and self-assessments of their ability to perform a variety of common, everyday language tasks in English. Results revealed that, for both listening and reading, TOEIC test scores were moderately strongly related to test takers' self-assessments, both overall and for each individual task. The correlations that were observed compare very favorably with those typically observed in validity studies using other kinds of validation criteria, such as course grades, supervisor ratings and self-reports.

In addition, the pattern of correlations among the measures also suggested modest discriminant validity of the listening and reading components of the Redesigned TOEIC Listening and Reading test. This result is consistent with a recent factor analytic study of a similar test (the TOEFL® iBT Test) by Sawaki, Stricker, and Oranje (2008), in which the correlation (r =.89) suggested highly related, but distinct, reading and listening factors.

In the present study, we were not able to evaluate the soundness of test taker self-reports as a validity criterion. However, in comparable studies that we have conducted recently in other similar contexts, can-do self-reports have exhibited several characteristics that suggest that they are reasonably trustworthy validity criteria, especially for low-stakes research, in which examinees have no incentive to intentionally distort their reports. For example, we have found that examinees rank-order the difficulty of tasks in accordance with expectations (Powers, Bravo, & Locke, 2007; Powers, Bravo, Sinharay, Saldivia, Simpson et al., 2008), and that they exhibit reasonably stable agreement about task difficulty when self-reports are collected again on later occasions (Powers et al., 2008). In addition, the results of the study reported here are consistent with previous meta-analytic summaries (e.g., Ross, 1998) that have documented substantial correlations between a variety of criterion measures and the self-ratings of learners of English as a second language.

In conclusion, the study has provided evidence of the validity of redesigned TOEIC test scores by linking them to test takers' assessments of their ability to perform a variety of everyday English language activities. The relationships that were detected are practically meaningful ones.

References

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. MS-19). Princeton, NJ: ETS.

Duke, T., Kao, C., & Vale, D. C. (2004, April). *Linking self-assessed English skills with the Test of English for International Communication (TOEIC)*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series No. MS-17). Princeton, NJ: ETS.

ETS. (n.d.). *TOEIC can-do guide: Linking TOEIC scores to activities performed in English.* Princeton, NJ: Author.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59,* 395–430.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43–62.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19,* 673–687.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280–296.

Powers, D. E., Bravo, G., & Locke, M. (2007). *Relating scores on the Test de français international™ (TFI™) to language proficiency in French* (ETS Research Memorandum No. RM-07-04). Princeton, NJ: ETS.

Powers, D. E., Bravo, G. M., Sinharay, S., Saldivia, L.E., Simpson, A. G., & Weng, V. Z. (2008). *Relating scores on the TOEIC Bridge™ to student perceptions of proficiency in English* (ETS Research Memorandum No. RM-08-02). Princeton, NJ: ETS.

Powers, D. E., Roever, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge Courseware scores against faculty ratings and student self-assessments* (ETS Research Rep. No. RR-03-11). Princeton, NJ: ETS.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15,* 1–20.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample* (TOEFL iBT Research Series No. TOEFLiBT-04). Princeton, NJ: ETS.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others of psychological assessment. *Psychological Bulletin, 90,* 322–351.

Tannenbaum, R. J., Rosenfeld, M., Breyer, J., & Wilson, K.M. (2007). *Linking TOEIC scores to self-assessments of English-language abilities: A study of score interpretation.* Unpublished manuscript.

Upshur, J. (1975). Objective evaluation of oral proficiency in the ESOL classroom. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967-1974* (pp. 53–65). Washington, DC: TESOL.

Notes

[1]Because this computation may not be entirely intuitive, we give this example. In any given row (i.e., for any given *can-do* task), there are six pairs of percentage comparisons that can be made. Take, for example, the percentages for the first *can-do* listening task in Table 2. The percentage in the lowest score interval (57) can be compared with the percentage (61) in the next higher score interval. This percentage (61) can be compared with the percentage (74) in the *next* higher score interval, which can in turn be compared with the percentage (82) in the next higher score interval, and so on. Six such comparisons are possible in each row (*can-do* task statement). Inconsistencies are those instances where the percentage at the next higher score interval is *lower* than the percentage at the immediate previous lower score interval.