Compendium Study

The *TOEIC*® Listening, Reading, Speaking, and Writing Tests: Evaluating Their Unique Contribution to Assessing English-Language Proficiency

Donald E. Powers, Feng Yu, and Fred Yan
With the assistance of IIBC, Japan and YBM, Korea

September 2013

The *TOEIC*® test was developed some 30 years ago in order to measure the ability to listen to and read English in a variety of real-world contexts. In 2006, the TOEIC program also developed direct tests of speaking and writing in order to assess the ability to speak and write in English in a workplace setting. This addition was in response to multinational corporations' needs for employees with high-level speaking and writing skills. The then new measures thus complemented the *TOEIC*® Listening and Reading tests. Together, the four components of the TOEIC test battery now provide a comprehensive assessment of English-language communication skills in all four language domains.

Occasionally, however, because of time or financial constraints, test users may be inclined to use a less than comprehensive assessment of important knowledge, skills or abilities. This seems especially true when assessing English-language proficiency where a key question is often, "When making decisions about test takers' qualifications for work or study, can a single measure (a test of speaking ability, for instance) adequately substitute for a more complete assessment of a test taker's overall proficiency in English?" In some contexts, the answer may be "Yes, less comprehensive testing may be sufficient," while in other instances, it may be an emphatic "No."

In any case, considering only a single skill (or fewer than four skills) may provide an imprecise estimate of a person's ability to use English. As a result, test users' expectations regarding examinees' global communication skills and their on-the-job performance may not be met. The resulting dissatisfaction has in fact been chronicled in several newspapers, which have reported on numerous cases of TOEIC test takers who had obtained high scores on the TOEIC Listening and Reading test but were seriously deficient with regard to their overall communicative ability (*Jungang Daily*, 2005, as cited by Choi, 2008). Such criticisms were, at least in part, the motivation for the TOEIC program's development of Speaking and Writing measures. The point here is that although listening and reading tests may provide an *indirect* indication of speaking and writing ability, they do not provide a comprehensive assessment of communicative ability.

The primary objective of the study presented here was to test the following hypothesis: A more accurate estimate of English proficiency in any specific language domain (speaking, for instance) can be attained by assessing skills not only in that domain but in other related domains as well (listening, for instance). Because the four domains of language are related in complementary ways, a measure of ability in one (e.g., listening) can, when used in conjunction with a measure of the target ability (e.g., speaking), add nuance/depth and accuracy to the measurement of the target ability.

The TOEIC Tests

For the listening section of the TOEIC test, test takers listen to a variety of questions and short conversations recorded in English and then answer multiple-choice questions based on what they have heard. Stimuli include photographs, question-responses, conversations, and short talks.

The reading section of the TOEIC test requires test takers to read a variety of materials and answer multiple-choice questions based on incomplete sentences, error recognition, or text completion, and reading passages.

Speaking is assessed by six different kinds of tasks requiring various types of responses, which are evaluated by trained raters according to the following criteria: pronunciation, intonation and stress, grammar, vocabulary, cohesion, and the content's relevance and completeness. Writing is assessed by three different task types, with responses evaluated according to the following criteria: grammar, relevance of the response to the stimulus, quality and variety of sentences, vocabulary, organization, and the extent to which the examinee's opinion is supported by reasons and examples.

Method

In the fall of 2011, we assembled a 24-item can-do self-assessment inventory, which included six statements for each of the four language domains. These statements were selected from longer self-assessment inventories that had been administered in previous studies of the TOEIC Listening and Reading test (Powers, Kim, & Weng, 2008) and the *TOEIC*® Speaking and Writing tests (Powers, Kim, Yu, Weng, & VanWinkle, 2009). These studies provided support for the utility of each of the separate TOEIC test measures. The construction of these longer inventories is described elsewhere (Powers et al., 2008; Powers et al., 2009). Suffice it to say, their development drew heavily from previous research by Ito, Kawaguchi, and Ohta (2005); Duke, Kao, and Vale, (2004); and Tannenbaum, Rosenfeld, Breyer, and Wilson (2007).

For the current effort, the sample of tasks was selected from the longer assessments so as to represent a range of activities of varying degrees of difficulty. The resulting abbreviated 24-item can-do inventory (see the appendix for a list of tasks) was translated into Japanese and Korean by TOEIC test in-country representatives and administered to individuals who had taken all four TOEIC test measures in Japan and Korea between approximately June 2010 and June 2012. Test takers used a 5-point scale to rate how easily they could perform each task of the 24 can-do tasks: 1 (*not at all*), 2 (*with great difficulty*), 3 (*with some difficulty*), 4 (*with little difficulty*), or 5 (*easily*).

Respondents were encouraged to respond to each statement. If, however, they had never actually performed the activity described, they were asked to rate how easily they believed they could perform it if they were to try.

Correlations were computed among the four TOEIC scores and the four can-do self-assessment scores, with scores determined for each language domain by summing over the responses (1 to 5) for each of the six statements written for a domain. In addition, for individual can-do statements, correlations with the corresponding TOEIC test measure were computed.

Our main objective was to determine if self-assessed performance in a given domain, such as speaking, could be better predicted by considering not only the TOEIC test measure that corresponded to the domain (in this case, *TOEIC*® Speaking scores), but other TOEIC test measures as well. To do so, we performed hierarchical regression analyses for each domain to assess the incremental contribution of additional TOEIC test measures to the prediction of self-assessed performance. The increment in the multiple *R* was taken as an indication of the value added by considering additional TOEIC scores. Because TOEIC Listening and Reading scores are always available in tandem, they were considered in combination as predictors here (instead of separately), as our focus was more on providing information for test score users than on advancing language theory. The results are discussed in terms of increased proportion of variance explained. They are also presented in a more nonstatistical manner, showing self-reported performance in each domain in relation to scores on multiple TOEIC test measures.

Results

We obtained data from 974 test takers in Korea and 1,351 in Japan. For test takers who had taken the same tests on more than one occasion, we selected scores from the test administrations that were closest in time. About 83% of our sample had taken all of the four tests within a one-year period (64% within six months), but a small number (less than half of 1%) had taken the tests slightly more than two years apart. Approximately 58% of the participants were females. Participants' ages ranged from 18 to 69; the median age was approximately 30 years.

Each of the four 6-item can-do measures proved to be highly reliable. Cronbach alpha internal consistency estimates were .93 for both the can-do listening and reading measures and .95 for both the can-do speaking and writing measures. The correlations among can-do measures ranged from .68 (between reading and speaking) to .82 (between listening and speaking), with a median correlation among all four measures of .77. Correlations among TOEIC test scores ranged from .58 (between listening and writing) to .74 (between reading and listening), with a median correlation of .62 among all four test scores.

Table 1 provides the correlations between TOEIC scores and test takers' assessments of their ability to perform English-language tasks in each domain. As Table 1 shows, the correlation between self-assessed performance on can-do tasks and the corresponding TOEIC score ranged from .44 for writing to .51 for speaking. Moreover (reading down each column), each TOEIC score correlated very slightly higher with can-do reports in its domain than with reports in any of the other three domains. (Reading across columns, however, each can-do report did not always correlate most strongly with the test corresponding to its domain.)

Table 1

Correlations Between Can-Do Self-Assessments and TOEIC Scores

Self-assessment	M (SD)	TOEIC test			
		L	R	S	W
Listening	23.4 (4.7)	0.46	0.40	0.50	0.42
Reading	24.3 (4.2)	0.40	0.47	0.41	0.40
Speaking	21.0 (5.4)	0.42	0.34	0.51	0.40
Writing	21.0 (5.4)	0.39	0.43	0.45	0.44
	Mean	423	373	133	147
	SD	63	78	27	24

Note. All correlations are significant at $p < .001$. Correlations in bold are those relating each TOEIC test to the self-assessment to which it corresponds. L = TOEIC Listening score, R = TOEIC Reading score, S = TOEIC Speaking score, W = TOEIC Writing score.

Individually, the correlations of each TOEIC score with can-do reports in the domain that it corresponds to ranged from .32 to .42 for reading, from .34 to .43 for listening, from .42 to .48 for speaking, and from .36 to .40 for writing. The means for each task on the 5-point difficulty scale ranged from 3.51 to 4.40 for listening tasks, from 3.20 to 4.52 for reading tasks, from 2.99 to 3.99 for speaking tasks, and from 3.08 to 3.97 for writing tasks. Thus, although we accomplished the goal of including tasks that varied in difficulty, overall, tasks were generally rated by the sample as being relatively easy on average.

Table 2 displays the results of the hierarchical regression analyses. In short, for each domain, self-reported performance was best explained by the TOEIC test measure that corresponded to that domain. In addition, for each domain, the consideration of each of the other noncorresponding TOEIC test measures also contributed (above and beyond the corresponding score) to explaining self-reported performance. For example, 22.2% of the variation in self-reported performance on the six listening tasks was accounted for by performance on the TOEIC Listening and Reading test. An additional 6.9% was accounted for when TOEIC Speaking scores were considered. Furthermore, an additional statistically significant, but very small, portion of variance (0.5%) was explained when *TOEIC*® Writing scores were added. The next biggest incremental contribution (4.1% of variance over and above an initial 19.3%) occurred when TOEIC Listening and Reading scores were considered along with TOEIC Writing scores to predict self-assessed writing ability. Smaller, but statistically significant increments (of about 2%) were also noted when, in addition to TOEIC Listening and Reading scores, TOEIC Speaking scores were used to predict self-assessed performance in reading. Likewise, a similar increase (2%) was observed when, in addition to TOEIC Speaking scores, TOEIC Listening and Reading scores were used to predict self-assessment of speaking ability.

Table 2

Results of Hierarchical Regression Analyses

Explanatory variable	Cumulative R^2	Increase in R^2	F increase	*df*
		Listening self-assessment		
L, R	0.222	0.222	332.2	2, 2322
S	0.291	0.069	225.8	3, 2321
W	0.296	0.005	16.5	4, 2320
		Reading self-assessment		
L, R	0.225	0.225	338.7	2, 2322
S	0.246	0.021	64.6	3, 2321
W	0.250	0.004	12.4	4, 2320
		Speaking self-assessment		
S	0.260	0.260	888.5	1, 2323
L, R	0.280	0.020	32.2	3, 2321
W	0.285	0.005	16.2	4, 2320
		Writing self-assessment		
W	0.193	0.193	555.1	1, 2323
L, R	0.234	0.041	62.1	3, 2321
S	0.260	0.026	81.5	4, 2320

Note. All *F*s are significant at $p < .001$. L, R = TOEIC Listening and Reading test, S = TOEIC Speaking test, W = TOEIC Writing test.

To provide a more intuitive understanding of the practical implications of the results of the hierarchical regression analyses, we have displayed the means of self-assessment reports according to both (a) TOEIC scores in the corresponding domain and (b) TOEIC Listening and Reading score levels. Table 3 (and Figure 1) shows the results for speaking; Table 4 (and Figure 2) shows the results for writing. Means are shown only for those cells having at least five test takers. As can be seen in each table, the means for self-assessments increase consistently with each higher level of the corresponding TOEIC score, regardless of TOEIC Listening and Reading level. In addition (and more relevant to our objective), the means also increase slightly but consistently for each of the three levels of TOEIC Listening and Reading score for each level of the TOEIC score (Speaking or Writing) that corresponds to the domain of the self-assessment. The results of the hierarchical regression are even more dramatic when we display the same kind of cross-tabulation for listening self-assessment means classified according to (a) TOEIC Listening plus Reading score levels and (b) broad TOEIC Speaking score levels (Table 5 and Figure 3).

Table 3

Mean Self-Assessment Rating for Speaking Tasks by TOEIC Speaking Score Level and TOEIC Listening + Reading Score Level

TOEIC Listening + Reading score level	TOEIC Speaking score level						
	40–50	60–70	80–100	110–120	130–150	160–180	190–200
Top third[a]	-	-	2.96 *n* = 12	3.31 *n* = 66	3.72 *n* = 373	4.20 *n* = 262	4.55 *n* = 60
Mid third[b]	-	3.03 *n* = 5	2.80 *n* = 49	3.26 *n* = 180	3.54 *n* = 431	3.95 *n* = 100	4.50 *n* = 5
Lowest third[c]	2.46 *n* = 16	2.61 *n* = 32	2.74 *n* = 181	3.10 *n* = 240	3.39 *n* = 277	3.83 *n* = 29	-

Note. SDs for cell entries range from 0.85 to 1.04.

[a]For top third, mean = 917, SD = 40, *n* = 773.
[b]For mid third, mean = 796, SD = 34, *n* = 773. [c]For lowest third, mean = 618, SD = 97, *n* = 779.

Table 4

Mean Self-Assessment Rating for Writing Tasks by TOEIC Writing Score Level and TOEIC Listening + Reading Score Level

TOEIC Listening + Reading score level	TOEIC Writing score level					
	80 or below	90–100	110–130	140–150	170–190	200
Top third[a]	-	-	3.27 *n* = 41	3.80 *n* = 424	4.35 *n* = 285	4.49 *n* = 23
Mid third[b]	-	2.83 *n* = 14	3.28 *n* = 136	3.46 533	4.03 *n* = 84	-
Lowest third[c]	-	2.76 *n* = 77	3.01 *n* = 296	3.21 *n* = 354	3.72 *n* = 21	-

Note. SDs for cell entries range from 0.91 to 1.06.

[a]For top third, mean = 917, SD = 40, *n* = 773. [b]For mid third, mean = 796, SD = 34, *n* = 773.
[c]For lowest third, mean = 618, SD = 97, *n* = 779.

Table 5

Mean Self-Assessment Rating for Listening Tasks by TOEIC Listening + Reading Score Level and TOEIC Speaking Score Level

TOEIC Speaking score level	TOEIC Listening and Reading score level					
	340–450	450–560	560–670	670–780	780–890	890–990
Top third[a]	-	-	4.09 *n* = 9	4.04 *n* = 51	4.17 *n* = 204	4.52 *n* = 411
Mid third[b]	-	3.94 *n* = 14	3.53 *n* = 67	3.76 *n* = 202	3.93 *n* = 312	4.12 *n* = 212
Lowest third[c]	3.03 *n* = 35	3.20 *n* = 63	3.34 *n* = 73	3.47 *n* = 264	3.72 *n* = 187	3.94 *n* = 54

Note. SDs for cell entries range from 0.62 to 1.02.

[a]For top third, mean = 164, SD = 14, *n* = 677. [b]For mid third, mean = 135, SD = 5, *n* = 860.
[c]For lowest third, mean = 104, SD = 18, *n* = 788.

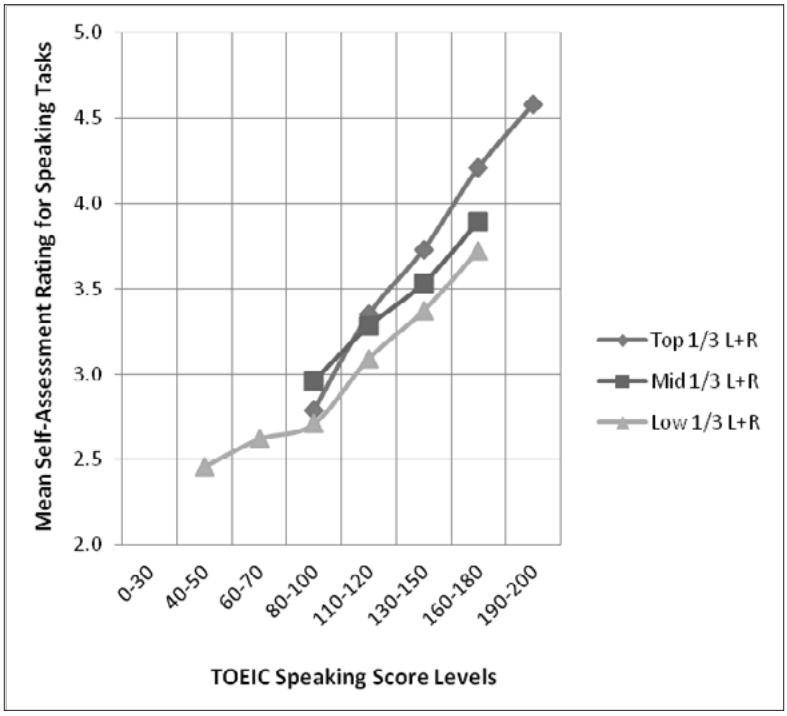

***Figure 1*. Mean self-assessment rating for speaking tasks by TOEIC Speaking score level and TOEIC Listening + Reading score level.**

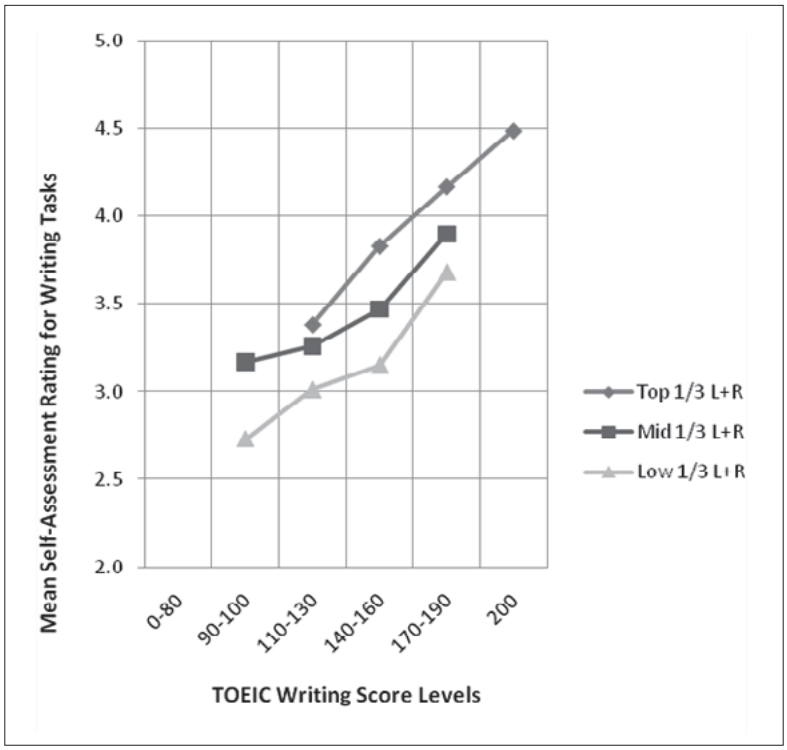

***Figure 2*. Mean self-assessment rating for writing tasks by TOEIC Writing score level and TOEIC Listening + Reading score level.**

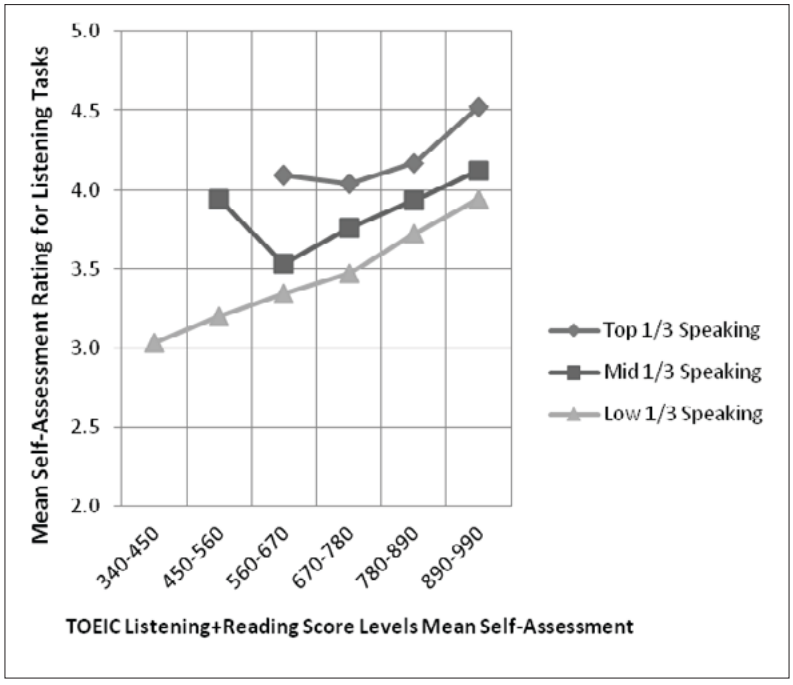

***Figure 3*. Rating for listening tasks by TOEIC Speaking score level and TOEIC Listening + Reading score level.**

Discussion

For this study, a large-scale data collection effort was undertaken to establish links between (a) test takers' performance on each of the TOEIC tests (Listening and Reading, Speaking, and Writing) and (b) self-assessments of test takers' self-reported ability to perform a sample of common, everyday language tasks in each of the four corresponding English-language domains. The study results provide corroboration of earlier studies of the validity of TOEIC scores that have used test-taker self-assessments as a criterion. In the current study, the correlations between TOEIC scores and self-assessments were comparable to those found previously (Powers et al., 2008; Powers et al., 2009) and can, according to social science conventions (Cohen, 1988), be described as being on the threshold of *large*. Thus, the utility of each of the four separate TOEIC scores, when used alone, was confirmed.

We also found that examinees rank-ordered the difficulty of tasks in accordance with expectations from previous studies (Powers et al., 2008; Powers et al., 2009). This finding constitutes additional evidence for the trustworthiness of self-assessments as a validity criterion for TOEIC scores. This result is generally consistent also with evidence on the use of self-assessments in such diverse fields as personality research (Ackerman, 2002), higher education (Falchikov & Boud, 1989), organization psychology (Mabe & West, 1982), and language assessment (Ross, 1998).

More importantly, by collecting information from each of the four language domains, we were able to demonstrate the utility of using all four TOEIC test measures for assessing English-language proficiency. It is relatively easy to argue that listening, reading, speaking, and writing skills are logically distinct, and there are numerous empirical studies to support this argument (e.g., In'nami & Koizumi, 2012; Sawaki, Stricker, & Oranje, 2008). However, there are, to our knowledge, far fewer, if any, studies that support the empirical *utility* of employing multiple measures of language skills for decision making, at least with regard to the use of TOEIC scores.

In conclusion, results provide reasonably strong support for the initial hypothesis: More precise estimates of English proficiency in a specific language domain are possible by assessing skills not only in that domain but in other related domains as well. Possibly because the four domains of language are related in such intricate ways, a measure of ability in one can, when used in conjunction with a measure of the target ability, add nuance/depth and accuracy to the measurement of the target construct.

Furthermore, although the contribution due to using additional measures is relatively small (when compared with the contribution of the measure that is most closely aligned with the performance domain of interest), it is statistically significant. We suggest that the results are also *practically* meaningful by virtue of the proportion of additional variance (of self-assessed performance) that is explained by considering multiple test measures. The results also reveal that the contribution may be greater for some domains than others.

Like most research studies, this study has certain limitations. Chief among them is that we were unable to control the length of time between test administrations. As a result, a significant minority of the study's participants had taken the four TOEIC tests over a relatively long time period. The import of this fact is as follows: The relations among test scores, on which analyses depended, were only rough proxies for the relations we sought to estimate; that is, the correlations among TOEIC scores at a single point in time. The extent to which this misalignment may have depressed correlations and affected interpretations is uncertain.

Second, the results of this study are based on data from only two, somewhat similar countries. It is not completely clear, therefore, that the findings presented here generalize to other countries that may, for instance, employ different methods of teaching or place different emphases on performance in each of the four language domains.

Finally, in order to minimize participants' response burden and to increase the likelihood of their responding, we relied on a less than comprehensive sample of English-language tasks. Although our self-assessment measures proved to be quite reliable from an internal consistency viewpoint, they may not have fully supported the kind of generalizability that we sought. Nonetheless, we believe, to the degree that the language tasks studied here are important for success in a global business environment, that the findings of this study lend support to the notion of using TOEIC test measures, either individually or in combination, to recruit, hire, and train prospective employees who are required to use English in an international workplace.

References

Ackerman, P. L., Beir, M.B., & Bowen, K.R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences*, *34*, 587-605.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25*, 39–62.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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.

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

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised *TOEIC* test: A multiple-sample analysis. *Language Testing, 29*, 131–152.

Ito, T., Kawaguchi, K., & Ohta, R. (2005). *A study of the relationship between TOEIC scores and functional job performance: Self-assessment of foreign language proficiency* (TOEIC Research Rep. No. 1). Tokyo, Japan: Institute for International Business Communication.

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., Kim, H.-J., & Weng, V. Z. (2008). *The redesigned TOEIC (listening and reading) test: Relations to test-taker perceptions of proficiency in English* (Research Report No. RR-08-56). Princeton, NJ: Educational Testing Service.

Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & VanWinkle, W. (2009). *The TOEIC Speaking and Writing tests: Relations to test-taker perceptions of proficiency in English* (ETS Research Report No. RR-09-18). Princeton, NJ: Educational Testing Service.

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* (Research Report No. RR-08-09). Princeton, NJ: Educational Testing Service.

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.

Appendix

List of Can-Do Self-Assessment Tasks

Listening Tasks

Understand someone speaking slowly and deliberately, who is giving me directions to a nearby location.

Understand directions about what time to come to a meeting and where it will be held.

Understand explanations about how to perform a routine task related to my job.

Take a telephone message for a coworker.

Understand a coworker discussing a simple problem that arose at work.

Understand lines of argument and the reasons for decisions made in meetings that I attend.

Reading Tasks

Read office memoranda in which the writer has used simple words or sentences.

Read and understand simple, step-by-step instructions (e.g., how to operate a copy machine).

Read and understand a letter of thanks from a client or customer.

Read and understand an agenda for a meeting.

Read English to translate text into my own language (e.g., letters and business documents).

Read highly technical material in my field or area of expertise with little use of a dictionary.

Speaking Tasks

Make, change, or cancel an appointment to see a person.

Telephone a company to place (or follow up) an order for an item.

Tell a foreign colleague or newly employed person how to perform a routine task.

Translate (e.g., conversations) in an informal setting.

Comment on or react to someone's opinion during a discussion.

Discuss (in English) world events with a guest.

Writing Tasks

Write a brief note to a coworker explaining why I was not able to attend a meeting.

Send an email or letter to a public organization to request needed information.

Translate documents (e.g., business letters, manuals) into English.

Write a memorandum to my supervisor (or instructor) to describe progress on a current project or task.

Prepare text and slides (in English) for a presentation at a professional conference.

Write a brief, several-page (formal) report explaining the progress being made on a project.