ETS TOEIC

Know English. Know Success.

COMPENDIUM STUDY

The Case for a Comprehensive, Four-Skills Assessment of English Language Proficiency

Donald E. Powers

January 2010

Occasionally, because of time or financial constraints, test users — those who use scores to make decisions about test takers' qualifications for work or study — may be inclined to use a less-than-fully-comprehensive assessment of important skills or abilities. This is true especially when assessing English-language proficiency, where a key question is often, "Can a single measure (typically, a test of speaking ability, or sometimes reading) serve as a sufficient proxy for a test taker's overall proficiency in all modes of communication in English, including listening, reading, writing, and speaking?"

In some contexts, speaking ability seems to be the most important of the four skills; furthermore, test takers' performance in each of the four skill areas is usually very highly related, so this strategy may not be an entirely unreasonable one.

However, if measuring only a single skill (or fewer than four skills) provides a less than adequate estimate of what a person can do in a real-life setting, test users may be dissatisfied, especially if expectations regarding examinees' on-the-job performance are not met. Such criticisms motivated revisions to two of ETS's well-known English-language testing programs: The TOEFL® test, which college and universities use to gauge the language skills of prospective international students, and the TOEIC® test, which employers in a variety of industries use to determine employees' readiness to use English in global communication.

For a variety of practical reasons, the TOEIC program originally offered only a multiple-choice test of listening and reading skills. ETS introduced the TOEIC Speaking and Writing tests in 2006. Similarly, until 2005 the TOEFL test included only listening, writing, and reading components.

A main impetus for adding a speaking component to the current TOEFL battery was criticism that, although students could perform well on the original TOEFL test, some could not communicate orally in academic situations. Similarly for the TOEIC program, many newspapers have reported on cases of TOEIC test takers who, although they obtained very high scores on the TOEIC Listening and Reading test, were seriously deficient with regard to overall communicative ability (Donga Ilbo, 2003; Hangyeorye, 2005; *Jungang Daily*, 2005 as cited by Choi, 2008).

The point here is that, although listening and reading tests can provide an indirect indication of speaking and writing ability, they provide no comprehensive assessment of communicative ability. Thus, the complaints noted above about the TOEIC Listening and Reading test begin to make the case for a more comprehensive assessment of English language skills.

Below, we extend this argument by presenting six (strongly related) reasons for a comprehensive assessment of all four English language skills – reading, listening, writing and speaking. Each component of the argument is discussed in turn. In brief, the reasons are as follows:

1. Users of English language proficiency tests like the TOEIC and TOEFL tests may sometimes be more interested in some language skills (speaking, for instance) than others. However, what they value most often is a person's ability to *communicate in English* in a variety of contexts that is likely to involve the use of *multiple* language skills either singly or in combination.

2. A more accurate estimate of a person's skill in any specific area (speaking, for example) can be attained by testing skills not only in that area but in related areas as well. Because the four aspects of language are inextricably intertwined, a measure of ability in a related domain (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.

3. The four skills are strongly correlated, but not to the degree that a measure of one can substitute perfectly for a measure on another. They are distinct enough, both logically and empirically,

that they have to be measured separately. Failing to measure all of these important aspects of proficiency, therefore, may leave critical gaps in a test taker's language proficiency profile.

4. Related to point 2 above is that, for most kinds of decision making, more information is almost always better than less. More trustworthy decisions are possible when additional relevant information is used to supplement initially available information, whether that decision concerns language abilities or other types of skills.

5. Standardized tests are almost always fairer to those who take them when multiple methods and multiple question formats are used. Some people perform better on some types of test questions than on others, and so it is appropriate to use a variety of methods and question types to assess critical abilities. Obtaining more information about test takers is not only valuable to the test user but also fairer to the test taker.

6. There are long-term societal consequences of testing English-language skills selectively. What is tested can affect what is taught as well as what is learned. Selective testing can result in greater attention paid to some language skills than others, resulting in uneven profiles of proficiency in overall communications skills. Testing all four skills is not only fairer to individuals, but it benefits society as well.

The current trend in language learning and language testing continues to be away from testing individual skills and instead toward a comprehensive, integrated testing of language skills. The six reasons summarized above are, in part, fueling this trend.

1. **What most language test users really value is usually the ability to *communicate in English*, an ability that is likely to involve two or more language skills in combination. For example, the TOEIC users are seeking employees who can communicate effectively in the workplace.**

Although each language skill is distinct and important in its own right, the proficiency of main interest to most users of English language proficiency assessments (like the TOEIC and the TOEFL tests) is usually not speaking, writing, reading, or listening per se. Rather, it is the overall ability to communicate in English. It is important, of course, to be able to understand the written and spoken word, and to produce English both orally and in written form. If, for example, a test user primarily wanted to select candidates who can perform such tasks as making understandable presentations at meetings, then measuring speaking skills would seem paramount. However, successfully performing even this "speaking-dependent" activity typically also depends on having read, understood, and summarized relevant information ahead of time. And during the meeting itself, it may be important to understand the reactions or questions of meeting participants in order to respond effectively. In other words, all four skills are likely to come into play more often than may be first apparent.

We believe, therefore, that users of the TOEIC tests are most interested in a broader construct – *the ability to communicate effectively in English within a workplace setting* (S. Hines, personal communication, Feb. 20, 2009). Communicative competence is a complex construct comprised of many aspects or facets; it may involve speaking, writing, reading and listening in various combinations in different settings or on different occasions. To focus exclusively on some aspects to the exclusion of others might under-represent the construct and thus provide an assessment that was less than sufficiently valid for its intended purpose. While a specific language test may focus in depth on a single skill area and provide very useful information about a test taker's proficiency in that skill, using measures of other skills also will usually allow for a more complete assessment of a test taker's ability to engage in effective communication.

Effective communication is a two-way activity involving both a sender and a receiver of a message. The listener or reader, some believe, has as much responsibility in understanding the message as the speaker or writer has in presenting it. For effective communication to occur, people need not only to speak or write but also to understand how others have perceived their messages if they are to respond in ways that address the concerns and questions of their audience. For instance, a job task might entail reading about a company's product and explaining it to a prospective customer, thus eliciting both reading and speaking skills to *produce* accurate communication, as well as listening skills to *evaluate the success* of the communication. An additional task might require understanding a question from a customer and then reading further in order to provide a satisfactory answer. As another example, as a prelude to speaking at an upcoming meeting, a presenter might need to prepare by reading and writing or taking notes. Thus, multiple skills are often required in combination for successful on-the-job performance.

Finally, communication skills (e.g., following instructions, conversing, and giving and receiving feedback) are becoming increasingly important in today's workplace (See, for example, Maes, Weldy, & Icenogle, 1997) and our communication abilities come into play to an increasing extent with the advance of technology – through, for example, voice mail, e-mail, or teleconferencing. Moreover, as teamwork becomes more and more critical in the workplace, communication skills will assume even greater importance. Stevens (2005) and others have predicted that the ability to communicate effectively — both orally and in writing — will become even more valuable as technology intensifies the influence of messages in the workplace.

2. **Estimating skill in a specific domain (speaking, for example) can be *facilitated* by testing skills in other, related areas as well. However, although this strategy can provide useful *supplemental* information, this is *not* to suggest that testing a skill in a related domain can *substitute* for testing a skill directly.**

This assertion is not surprising, perhaps, given the strong relationships among the four skills, and the common subskills that underlie them. For example, vocabulary figures prominently in speaking and in writing, and one also needs to understand the meaning of words in order to read and to write. Other skills such as word choice and awareness of audience may be similar for speaking and writing, and awareness of the style used by a message sender is important for both listening and reading. Thus, because similar components underlie performance in several domains, the measurement of skill in one domain may indirectly provide information about a test taker's ability in another domain.

This contention also has some empirical support. For example, Wilson (1993) studied the relation of performance on the TOEIC Listening and Reading test to performance on the Language Proficiency Interview (LPI), a well-established direct assessment of oral language proficiency in which examinees respond to a series of increasingly complex questions from expert judges, who evaluate the responses according to standardized criteria. Wilson concluded that the TOEIC Listening and Reading test is a useful *indirect* measure of speaking proficiency. Predictions based on the test takers' TOEIC Listening and Reading test scores yielded reasonably accurate estimates of test takers' speaking skills as measured by the LPI.

In addition, there is some indication that using multiple measures may complement one another in terms of their ability to predict the degree to which test takers are able to perform everyday language tasks. For instance, Powers, Kim, Yu, Weng, and Van Winkle (see Chapter 11 in this compendium) investigated the relationship of the TOEIC speaking and writing measures to test takers' self-assessments of their ability to perform a variety of everyday language tasks in English.

For speaking, a total of 40 tasks of differing degrees of difficulty were included, such as:

- leaving a message on an answering machine to ask a person to call back;
- explaining ongoing troubles (e.g., about flight or hotel accommodations) and make a request to settle the problem;
- serving as an interpreter for top management on various occasions such as business negotiations and courtesy calls.

For writing, a total of 29 tasks such as the following were included:

- Write an e-mail requesting information about hotel accommodations
- Write discussion notes during a meeting or class and summarize them
- Prepare text and slides (in English) for a presentation at a professional conference

Although the scores from the speaking and writing tests were relatively highly correlated, further detailed analysis demonstrates the *unique* value of each test. The TOEIC *speaking* scores were somewhat better predictors of the ability to perform *speaking* tasks, and the TOEIC *writing* scores were better indicators of the ability to perform *writing* tasks. Both speaking and writing scores were reasonably good predictors of the ability to perform various individual language tasks in English.

For instance, consider the speaking task "using a menu, order food at a café or restaurant." For this very easy task, at the lowest TOEIC speaking score level of (0–50), only 21% of test takers said that they could perform the task either easily or with little difficulty. In contrast, at the highest TOEIC speaking score level (190–200), nearly all participants (98%) felt that they could perform this task easily or with little difficulty. At intermediate score levels, the percentages (38%, 52%, 71%, 81% and 93%) also rise consistently with each of the higher TOEIC speaking score levels.

This same consistent pattern was apparent for each and every task, although the percentages are much lower for more difficult tasks such as "serve as an interpreter for top management on various occasions such as business negotiations and courtesy calls," a task that only 2% of the lowest scoring participants indicated they could perform easily or with little difficulty, in comparison to 47% of the highest scoring participants.

Beyond this, however, the prediction of the ability to perform *both* speaking and writing tasks improved when *both* the TOEIC Speaking test *and* the TOEIC Writing test were used *together* to predict the ability to perform these tasks. For instance, when examinees are grouped according to their TOEIC *writing* scores — either as being in the highest third of all examinees or in the lowest third — those who scored *highest* on the TOEIC writing were more likely than those who scored *lowest* to report that they could perform the *speaking* tasks about which they were asked. This was true at each of the four TOEIC speaking score levels for which there were sufficient data (See Table 1).

TABLE 1

Percentage of Examinees Who Said They Could Perform Speaking Tasks, by the TOEIC Speaking and Writing Test Score Levels

TOEIC Writing level	Speaking Level 1-3	Speaking Level 4	Speaking Level 5	Speaking Level 6	Speaking Level 7	Speaking Level 8
Lowest third	---------	13	26	36	54	----------
Highest third	---------	20	35	53	71	----------

The important point here is that at each of the TOEIC speaking levels, the percentage is greater for examinees who had the higher TOEIC *writing* scores, indicating that although the TOEIC speaking scores are highly indicative of test takers' ability to perform speaking tasks, considering information about their TOEIC writing scores *in addition* to their speaking scores significantly increases our ability to forecast their performance on everyday speaking tasks.

The results are even more dramatic when test takers are grouped according to the high and the low TOEIC *speaking* scores and the relationship between self reports of *writing* ability and the TOEIC *writing* scores is examined.

3. The four skills of listening, reading, writing and speaking are distinct.

There is more than ample evidence to suggest that, although the four aspects of communicative ability are highly related, they are nonetheless logically and empirically distinct. Logically, the four skills are related in complementary ways. Both listening and reading are receptive skills — modes of understanding. Speaking and writing are productive skills. Thus, the four basic skills are related to each other by virtue of both the mode of communication (oral or written) and the direction of communication — either receiving or producing messages.

The question of whether language ability is a single, unitary trait or whether it is divisible into distinct components has been of interest to applied linguists for decades. For instance, more than 30 years ago, Oller (1976) posited that language abilities constitute a single language trait. This *unitary trait hypothesis* enjoyed some initial support, and until relatively recently, the issue of unitary vs. divisible traits was still a fairly contentious one. Recent research, however, (e.g., Bachman, Davidson, Ryan & Choi, 1995; Bachman & Palmer, 1981, 1982; Carroll, 1983; Kunnan, 1995; Oller, 1983) has benefited from more advanced data analysis approaches; as a result researchers have concluded that there are multiple components to language skill, and that the so-called *factors* represent both a prominent general language ability that is common to all domains, as well as specific abilities that are unique to each of the four domains. This interpretation is consistent, for example, with a recent investigation of the structure of the TOEFL® iBT test (Sawaki, Stricker, & Oranje, 2008).

Currently, researchers are also undertaking a formal study of the component skills measured by the TOEIC battery (Sinharay & Sawaki, 2009). So, eventually additional empirical evidence will help to inform the question of how distinct the four skills are for the TOEIC test as well. In the meantime, good evidence exists to support the uniqueness of the TOEIC listening, reading, speaking and writing measures. Liao, Qu, and Morgan analyzed data from more than the 12,000 TOEIC test takers, of

whom about 7,500 took all four measures. The following table shows the correlations among the four TOEIC measures. In short, the numbers reveal moderate, but far from perfect, relations among the four measures, suggesting that each measures a unique set of language skills.

TABLE 2

Correlations Among the TOEIC Listening, Reading, Speaking and Writing Scores

Score	L	R	S	W
R	0.76			
S	0.66	0.57		
W	0.59	0.61	0.62	

4. For sound decision making, more information is almost always better than less.

Bachman (2005) has stated that, when building a case for the use of a language test, the two key questions are:

How confident are you about the decisions you make on the basis of test scores?

How sure are you of the evidence you're using to make those decisions?

When making decisions about selecting, hiring, promoting, and so forth, good information is critical, and more information is almost always better than less. Adding relevant assessments to the mix will usually result in more reliable and valid decisions. This is especially true when skills relate as strongly to one another as listening, reading, writing and speaking skills do.

For decisions that involve English language proficiency, using tests of all four language domains provides a more comprehensive basis for decision making and thus results in more trustworthy decisions. Moreover, the use of multiple sources of information gives test score users more flexibility with respect to the kinds of decision making processes that they may use.

When multiple sources of information are available, test score users can employ *compensatory* as well as *non-compensatory* selection strategies. With compensatory selection, a test taker's strengths in one area can compensate for weaknesses in another area. With non-compensatory procedures, however, not all attributes are necessarily considered in decision making, and therefore strengths and weaknesses don't balance each other out. Both strategies may be appropriate, depending on the context, but the problem is that only one is possible if only a single-skills test is used.

The bottom line here is that with multiple test scores, test score users may choose to use either compensatory or noncompensatory procedures in their decision making. When only a single measure is available, compensatory selection is not an option.

5. Standardized tests are almost always fairer to test takers when multiple methods and formats are used.

It is paramount that a test yield trustworthy scores; it is equally important that an assessment is fair to all test takers. By fair we mean that the test methods should be broad enough to allow all test takers to show what they know or can do. Good assessment practice, therefore, demands that multiple formats and methods be employed when assessing important knowledge, skills and abilities. This reduces the

chances of inadvertently disadvantaging some test takers (and inappropriately advantaging others) simply because they do not perform well on a particular method of assessment or on a particular test question format. Toward this end, using language tests such as the TOEIC test to provide test takers with the opportunity to demonstrate their skills directly in all four language domains provides opportunities for test takers to demonstrate their English language skills in different ways.

In other words, good measurement practice dictates that we avoid putting all of our eggs in one basket. To the extent possible, important skills should be assessed by means of different modes, methods or formats so that the results of our assessments don't merely reflect the methods that are employed. The TOEIC Listening and Reading test, for example, employs multiple-choice questions that require test takers to *select* answers from a set of choices, while the speaking and writing measures require them to *produce* answers in response to a variety of different stimuli. Both computer-scored, multiple-choice and human-rated, constructed-response assessment is used, thus decreasing the chances that some test takers may be disadvantaged by the use of a single format or method of assessment. The use of these two very different assessment formats should also broaden the way in which English language skills are taught and thus result in more robust learning of these skills. (See point # 6 below.)

6. There are potentially serious (negative) societal consequences of testing English language skills selectively.

There are potentially serious societal consequences of choosing to test some aspects of language proficiency and not others. *Washback* is a very real phenomenon (Bailey, 1999). It has been alternatively defined as "the connections between testing and learning" (Shohamy, Donitsa-Schmidt & Ferman, 1996, p. 298) and "the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning" (Messick, 1996, p. 241) That is, what is tested is very likely to affect not only *what* is taught, but *how* it is taught — if not immediately, then at least in the longer term. Alderson and Wall (1993) have hypothesized more specifically that a test may influence *what* teachers teach (and *what* students learn) and also *how* it is taught and learned — the rate and sequence, and the degree and depth, for example.

A number of empirical studies have shown that testing can indeed influence what and how English language learners are taught (e.g., Choi, 2008; Cheng, 1997; Wall & Alderson, 1992; Wall & Horák, 2009). For the TOEIC tests specifically, Stoynoff (2009) has suggested that "those examinees who prepare to take the full TOEIC battery will likely experience more positive washback than those who prepare to take the single TOEIC test" (p. 33).

We have seen the effect in some regions of the world of introducing language proficiency tests like the TOEFL test and the TOEIC tests— higher performance over the years, due presumably to instructional emphasis on the language skills that are being tested. So, in light of the observations of Bailey (1999), Messick (1996), Stoynoff (2009) and others, we must recognize that deciding to test only speaking at the expense of other language skills risks encouraging less emphasis on writing, listening and reading, which may eventually result in lower skill levels in these areas.

In addition, it stands to reason that the more that different types of items that are used to test English language skills, the more generalizable the scores based on these item types are likely to be. Conversely, if only a very limited number of test item types are used (say, vocabulary, primarily) then quite predictably, examinees would become proficient at answering these kinds of items, at the exclusion of items that measure other related skills. Thus, requiring the testing of all four skills should also dilute any effect of simply teaching to the test.

In summary, the six reasons discussed here are consistent with the trend in language learning and language testing toward comprehensive, integrated testing of language skills. Market demands play a significant role in determining the nature of the assessments that are offered by test makers. What direction will programs like the TOEIC test, the TOEFL test and the IELTS testing programs take in the future? Traditionally, these programs have followed relatively independent development paths. More recently, however, there has been more collaboration between at least the TOEFL test and the TOEIC language testing programs. For instance, for many years the TOEFL program assessed only reading, listening and writing, the latter in only a multiple-choice format. Eventually, in order to meet market needs, a separate speaking test was developed (*The Test of Spoken English*™), and an essay was added to the writing measure to assess writing more directly. More recently, the TOEFL test was revised to include tasks that integrate multiple language skills. If, as we expect, the TOEIC program eventually follows the same path as the TOEFL program, we should, for the reasons discussed above, see the TOEIC market begin calling for the assessment of all four skills and, eventually possibly, for the integration of these skills (Everson, personal communication, Feb. 20, 2009).

Summary

To summarize briefly, our argument for using all four language skills, as opposed to testing more selectively, is as follows: It is the broader trait of communicative competence, not specific individual skills, that is critical in most academic and workplace settings and of most interest to users of tests like the TOEFL and TOEIC tests. It is important, however, to test for each of these four skills individually because each is a critical aspect of communicative competence. Furthermore, direct evidence of specific individual skills can provide at least indirect evidence of other skills.

Though strongly related, each of the four skills — listening, reading, writing and speaking — are distinct, and each contributes uniquely to an individual's overall communicative ability. When test scores are used to make consequential decisions, the use of several sources of information yields better decisions than does a more selective use of information. Moreover, assessment is fairer to test takers if they are allowed to demonstrate their skills in multiple ways — with different tests, different methods and different question formats. Comprehensive testing also encourages broader and more generalizable teaching and learning of language skills by test takers. All of the reasons given here are consistent with the trend toward more comprehensive, integrated testing of language skills as seen in many prominent language testing programs.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14, 115-129.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1-34.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning*, 31, 67–86.

Bachman, L. F., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16, 449–465.

Bailey, K.M. (1999). *Washback in language testing*. (ETS Research Memorandum No. RM-99-04). Princeton, NJ: ETS.

Carroll, J. B. (1983). Psychometric theory and language testing. In J.W. Oller, Jr. (Ed.), I*ssues in language testing research* (pp. 80–107). Rowley, MA: Newbury House.

Cheng, L. (1997). How does washback influence teaching? *Implications for Hong Kong. Language and Education*, 11, 38-54.

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

Donga Ilbo. (2005). Available from http://www.donga.com/docs/magazine/shin/2003/10/27/2003/10270500015/200310270500015_1.html

Everson, P. (2009, February). *The importance of four skills in English education*. Presentation at the Global Talent Cultivation Symposium, Seoul, Korea.

Hangyeorye. (2005). Available from http://www.hani.co.kr/section-004000000/2005/07/00400000020050708171400l.html

Hines, S. (2009, February). *TOEIC speaking and writing tests*. Presentation at the Global Talent Cultivation Symposium, Seoul, Korea.

Jungang Daily. (2005). Available from http://joongangdaily.joins.com/200512/04/200512042255394979900090409041.html

Kunnan, A. (1995). *Test taker characteristics and test performance: A structural equation modeling approach*. Cambridge, England: Cambridge University Press.

Maes, J. D., Weldy, T. G., & Icenogle, M. L. (1997). A managerial perspective: Oral communication competency is most important for business students in the workplace. *The Journal of Business Communication,* 34, 67-80.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, 241-256.

Oller, J. W. (1983). A consensus for the eighties? In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 351–356). Rowley, MA: Newbury House.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL® internet-based test (iBT): Exploration in a field trial sample* (ETS Research Rep. No. RR 08-09). Princeton, NJ: ETS.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13, 298-317.

Sinharay, S., & Sawaki, Y. (2009). *How much evidence is there in favor of reporting the four scores in TOEIC?* Manuscript in preparation.

Stevens, B. (2005). What communication skills do employers want? Silicon Valley recruiters respond. *Journal of Employment Counseling*, 42, 2-9.

Stoyoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, 42, 1-40.

Wall, D., & Alderson, J. C. (1992). *Examining washback: The Sri Lankan impact study* (ERIC Document Reproduction Services Report ED345512).

Wall, D., & Horák, T. (2009). *The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe: Phase 4, Measuring change.* Manuscript in preparation.

Wilson, K. M. (1993). *Relating TOEIC® scores to oral proficiency interview ratings* (TOEIC Research Summary No. TOEIC-RS-01). Princeton, NJ: ETS.