ETS TOEIC

Know English. Know Success.

RESEARCH REPORT

Validity: What Does It Mean for the TOEIC® Tests?

Donald E. Powers

January 2010

The ability to communicate in English is often essential in today's global workplace environment. Consequently, many employers must regularly make critical decisions concerning the English language skills of their employees and their prospective employees. The TOEIC® tests are designed specifically to facilitate these decisions. This paper describes the procedures that the developer of the TOEIC tests has undertaken to ensure that the tests fulfill this purpose.

As consumers of many kinds of products, we may all ask ourselves what makes us decide to buy something. Arguably, the main things that we look for are great service and a great product.

The same applies for the TOEIC tests. Great *service* is provided by the ETS Preferred Network members who market the tests, as well as by the ETS TOEIC staff who devote many of their waking hours to responding to requests from customers developing multiple forms of the tests, providing quick turnaround for score reporting and so on.

But what makes a great *product/test*? For the TOEIC tests, some of the appeal is that they are widely available and that their scores are recognized worldwide. But if there is one, single aspect of a test that defines its value, that aspect is the meaningful or *valid* scores that the test yields.

And just what do we mean by *valid* or *validity*? Validity is both a very simple concept and a very complex one. Many articles, chapters and entire books have been devoted to the subject. One of the greatest validity theorists was Sam Messick, a former vice president for research at ETS. Here is Messick's (1989) definition of the complex concept of validity:

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (p. 1.8)

Clearly, Messick's definition is a relatively complex one. So let's see if we can't simplify things a bit.

In simpler terms, validity is the degree to which a test is doing the job it was intended to do, whether that job is certifying that someone has mastered a given body of knowledge, whether it is to facilitate admissions decisions to colleges or universities by measuring a person's readiness for further study, or whether it is to attest to the fact that someone possesses the knowledge, skills and abilities needed to practice medicine or law or to begin teaching. In other words, do the scores mean what we think they mean, and does the test fulfill the purpose for which it was designed?

So, just what is it that the TOEIC tests are supposed to do? There are various statements in the TOEIC program's promotional materials, but basically these statements are all slight variations of this theme: The TOEIC tests measure a person's ability to communicate in English in the context of daily life and the global workplace environment using key expressions and common, everyday vocabulary.

The promotional materials also speak to what the TOEIC tests DO NOT require. For example, the tests do not require the test taker to have a special knowledge of business terms. What else is claimed? The TOEIC tests, which cover reading, listening, speaking and writing, relate to one another in complementary ways and provide unique measurement of particular skills. This latter claim is important because none of the TOEIC tests are regarded as an appropriate substitute for any one of the other TOEIC tests.

The TOEIC tests all meet ETS standards for quality and fairness. This means that:

- different (alternate) forms of the same test yield scores that have the same meaning,
- timing, security and testing conditions are maintained across administrations,
- all test takers are given access to materials that help them to understand the procedures required to take the test,
- the difficulty of test questions is appropriate for efficient testing,
- speaking and writing scores do not depend on who (among many trained readers) scores the tests, and enough aspects of English language proficiency are tested to allow a test taker's ability in all four language skills to be determined.

The essence of validation is defending the claims that are made for a test and test makers do not just *respond* to those who question their claims. Instead, it's almost as if the test makers invite critics to hit them with their best shot to see if the tests are strong enough to hold up. In fact, test makers themselves often design research to challenge their own claims. The way they do that is to gather evidence of various sorts and determine if this evidence is consistent with what they claim about their tests. As discussed below, this evidence can be a variety of different kinds.

One thing that everyone can agree on is that validation is not a static process. It begins when a test is first being planned and continues until the test is retired. Validation is not something that's simply added on at the end of the test development process, nor is there a single moment when a test maker can officially declare that a test has been validated enough. Rather, validation involves gathering evidence and, based on all the evidence gathered, making the best case possible for the claims that are made about the test. Many different kinds of evidence can be useful, depending on the nature of the test.

So, what exactly do we mean by validity evidence? There are many different kinds of evidence that may be appropriate, depending on the particular purpose of the test. As mentioned earlier, validation is considered by some to be never-ending, and so there is usually no clear agreement on just how much evidence is necessary. One thing that everybody does agree on, however, is that more evidence is always better than less. Although one can never have too much evidence, usually a point comes when test makers feel comfortable that they've made at least an adequate case to support the claims they've made about a test.

One clear point on the validity evidence timeline is when the test is first offered to customers. There are certain things that test makers are obligated to do before selling the tests, but much validity research can only be done after people start using the test, when test scores are taken seriously by both test users and test takers.

Often a useful strategy for discussing a complex idea is to approach the idea from several different directions. This is also a useful approach when thinking about validity. So, here is another way to define or think about test validity: It's the extent to which a test measures *exactly* what a test maker intends it to measure, nothing more, nothing less. *Nothing more, nothing less* is the ideal, of course. In practice, one can only *try* to reach this goal. No test is ever perfect, and sometimes it may measure unwanted things. Also, no test is ever so comprehensive that it measures *all* the important aspects of what we are trying to test.

In other words, two general classes of things can weaken the meaning or validity of test scores. Testing experts sometimes use the terms *construct under-representation* and *construct-irrelevant difficulty* to refer to these kinds of threats. These terms suggest that tests don't always measure all that we'd like them to measure (construct under-representation), and that sometimes they measure things that we'd like them not to measure (construct-irrelevant components). The goal is to test as many of the important aspects as possible, while at the same time making sure that unwanted things are not accidentally tested, or that at least these unwanted influences (test anxiety, for example) are held to a minimum.

What the TOEIC tests attempt to measure are the ability to communicate in English, which is a very complex construct. It is impossible to measure every important aspect of this construct in the relatively limited amount of time that is available for testing. The best that can be done is to make sure that the most important things are included in the test, and that they are included in proportion to their importance. The very detailed test specifications (or blueprints) that the TOEIC assessment developers work from are intended to meet this goal for every TOEIC test form that is developed.

Test tasks are designed to elicit aspects of communicative ability that are deemed to be especially important, and when examinee responses for the TOEIC Speaking and Writing tests are scored, raters are trained explicitly to look for these aspects, such as task completion, organization, vocabulary use and correct grammar. The TOEIC Listening and Reading test questions are also designed to measure skills that allow the test taker to function in the real world, such as listening for purpose, reading for required details and so on. All in all, the TOEIC tests provide a very good sampling of the most important aspects of speaking, writing, reading and listening skills.

To reiterate, the issue of construct-irrelevant components arises when a test measures things that we'd rather not have it measure — things that are unrelated to what we are trying to test. This can be thought of as contamination — like something that pollutes the air we breathe or the water we drink. It goes without saying that no test is perfect and to some degree all tests contain some things that make them less "pure" than we'd like. The best that we can ever hope for is to minimize the influence of these factors. This is accomplished in a number of ways, starting with the design of the tests.

Usually (but not always), construct-irrelevant components make a test more difficult — but for the wrong reasons (that is, for the TOEIC test, for reasons other than not knowing how to communicate in English). So, what are some of the factors that the TOEIC developers worry about and attend to?

Well, first of all, they try not to make the test questions so complex that test takers may become anxious and feel incapable of showing what they are able to do. They also make sure that all examinees are given enough time to show what they can do, so that their scores don't simply reflect how quickly they can answer test questions.

Highly technical vocabulary, complicated directions and elaborate test item formats are avoided. In the same way, test takers' lack of experience with computers should not hinder their performance because they can't enter responses or navigate the test efficiently. Thus, the TOEIC tests are designed so that they don't have these problematic factors. But just in case the developers aren't completely successful, pre-test practice materials are made available to make sure that all examinees are familiar with all testing procedures BEFORE they take the tests.

What is done to ensure validity as the tests are being developed? First of all, the TOEIC developers are world-class assessment specialists, who keep up to date on current developments in language learning and language testing. The TOEIC developers also have access to specialists in other areas of testing at ETS as well as access to specialists in the other language testing programs that ETS operates (e.g., the TOEFL® program).

Second, very thorough and detailed test specifications (called test blueprints) are followed to ensure that each form is highly similar to every other one and to ensure that the right content is covered in the same proportion in each test form.

The process that is used for reviewing test items once they are written is exceedingly thorough. For the TOEIC items, some 20 reviewers inspect each and every test question before it is used. Some questions don't survive this scrutiny, and others may undergo extensive revision before they meet the required standards of quality.

And finally, each distinct *type* of question is thoroughly pilot-tested in the design phase to make sure it performs properly — to ensure that test takers know how to deal with each kind of question format and to make sure that questions are appropriately difficult.

What is done to ensure validity as tests are being scored? For the speaking and writing tests, detailed scoring guidelines are developed and used by raters when they evaluate test takers' responses. Only qualified raters are hired, and most importantly, once they are hired, these raters are trained in how to apply the scoring guidelines.

Moreover, all raters must pass a certification test to show that they can score responses accurately. In addition, once they are certified, the performance of raters is monitored continuously to ensure that they maintain their accuracy. Finally, statistics are computed in order to monitor how well raters are agreeing with one another — to make certain that all raters are using the guidelines in the same way.

What do we do as the tests are being offered? Routine statistical analyses are performed after tests are administered to make sure that test items are working properly. Also, ongoing research of various kinds is conducted to support the test. Ideally, as much research as possible is conducted before the

first time a test is given, but sometimes this just isn't practical. Some research can yield meaningful results ONLY after the test is in use and when test takers are motivated to try their best. Otherwise, the result might very well be false readings on the difficulty of the test and the extent to which scores are valid. Ongoing research also provides the opportunity to collaborate with test users to address their concerns and to collect the data that is needed.

Getting back to collecting validity evidence for a moment, validity evidence can take many forms, but there are basically only a few major kinds of evidence. They are:

- Judgmental or logical analyses. Experts may be used to judge the importance of various types of content and test questions.
- Examination of how test takers approach the test questions (the processes that are used to answer test questions). This kind of research may entail having examinees think aloud as they answer questions to make sure that the processes we intend to elicit are the ones that examinees actually use to answer questions. This technique is particularly useful for multiple-choice items, where there is no product to observe, as there is with some tests, such as tests for speaking and writing.
- Examination of differences in test structure, processes, or language use across groups (experts vs. novices, for example). For a test of writing skill, we might contrast the performance of professional writers with that of beginning writers to assure ourselves that writing skills are indeed being tested.
- Investigation of score change in relation to relevant experience, instruction and so on. For instance, knowing how formal instruction, everyday practice, or on-the-job experience affects the TOEIC test scores can inform the meaning of those scores.
- Examination of how test scores relate to, or correlate with, other variables or criteria. For the TOEIC scores, the interest might be in how well the TOEIC scores relate to on-the-job performance for jobs that require English language skills or how well the TOEIC scores relate to other English proficiency tests.

One kind of evidence that has been collected specifically for the TOEIC tests involves self-assessments provided by test takers themselves. These self-assessments have been called *can-do* reports because test takers are asked how well they can perform each of a variety of different language tasks in English. As it turns out, when these questions are asked properly, people are reasonably accurate at reporting how well they are able to read, listen, speak or write in another language. The extent that these self-reports agree reasonably well with the TOEIC scores provides good evidence for the validity of scores (i.e., that the scores mean what the test makers say they mean). Self-assessments also provide a very practical way of helping test-score users to understand what the TOEIC scores mean at each of the TOEIC score levels.

Of course, self-assessments are subject to some criticism because people aren't always completely aware of their own strengths and weaknesses, and they may sometimes intentionally exaggerate their abilities. A good deal of solid research evidence, however, shows that people are fairly accurate when it comes to reporting their skills and abilities in some domains, especially when they have no particular reason to lie about them. This seems to be especially true for language abilities.

In 2007 we asked about 5,000 test takers who took the TOEIC Listening and Reading test to complete a self-assessment (can-do) inventory when they took the test. The inventory asked test takers to consider each of 25 reading tasks and 24 listening tasks and indicate how easily they could perform each task — easily, with little difficulty, with some difficulty, with great difficulty, or not at all. A small sample of the reading tasks that we asked about is shown in Table 1.

TABLE 1

Example for Selected Reading Tasks. Percentage of the TOEIC Test Takers, by Reading Score Level, Who Said They Could Perform Various English Language Reading Tasks Either Easily or With Little Difficulty

Task:	TOEIC Reading Score						
	5–135	140–195	200–255	260–315	320–375	380–435	440–495
Read and understand a train or bus schedule	49	59	70	77	84	90	96
Read office memoranda in which the writer has used simple words or sentences	36	50	61	72	81	88	96
Find information that I need in a telephone directory	23	34	42	52	64	76	89
Read and understand an agenda for a meeting	6	14	22	34	46	62	84
Read English to translate text into my own language (e.g., letters and business documents)	5	12	16	23	36	50	74
Read and understand a proposal or contract from a client	4	7	11	17	25	42	58

The numbers in the table are percentages of test takers who said they could perform a task either easily or with little difficulty. These percentages are shown for test takers at each of the seven TOEIC reading score levels. For instance, for the first (relatively easy) task shown (Read and understand a train

or bus schedule), 49% of test takers at the lowest reading score level (5–135) said they could do this task easily or with little difficulty. At the highest TOEIC reading score level (440–495), nearly all (96%) test takers said they could perform this task either easily or with little difficulty.

For a more difficult task at the bottom of the table (Read and understand a proposal or contract from a client), only 4% of test takers at the lowest score level reported being able to do this task either easily or with little difficulty, while at the highest score level, 58% said they could accomplish the task with this same level of effort.

Note that for each task, without exception, the percentages rise with each of the higher TOEIC reading score levels. In other words, a strong relationship exists between the TOEIC reading scores and test taker reports of their ability to perform a variety of everyday language tasks in English — good evidence for the validity of the TOEIC scores.

Table 2 displays the same kind of information for a small sample of the listening tasks for which test takers were asked to indicate their ability. Again, the listening tasks differ in their difficulty, but the pattern is the same as for reading: Percentages of test takers who reported that they could perform a task rose steadily with each higher level of the TOEIC listening score. The complete details of the study and the results for all listening and reading tasks are available in Compedium Study 6.1 of this compendium and as a report in the ETS Research Report series (Powers, Kim, & Weng, 2008) on the TOEIC website, www.ets.org/toeic.

TABLE 2

Example for Selected Listening Tasks. Percentage of the TOEIC Test Takers, by Listening Score Level, Who Said They Could Perform Various English Language Listening Tasks Either Easily or With Little Difficulty

Task:	TOEIC Reading Score						
	5–135	140–195	200–255	260–315	320–375	380–435	440–495
Understand simple questions in social situations (e.g., "Where do you live?")	57	61	74	82	90	95	97
Understand directions about what time to come to a meeting and where it will be held	20	23	41	55	66	80	91
Take a telephone message for a co-worker	9	15	14	21	37	55	75
Understand lines of argument and the reasons for decisions made in meetings I attend	6	6	7	11	17	34	60
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

Recently, the same kind of self-assessment study was completed for the TOEIC Speaking and Writing tests. A similar self-report can-do inventory of speaking and writing tasks was assembled and administered to the TOEIC test takers in Japan and Korea when they took the TOEIC Speaking and Writing tests in the fall of 2008.

Several additional steps were followed in the development of this inventory of speaking and writing tasks — mainly getting additional input from major clients to ensure that the language tasks were ones that were regarded as important by the TOEIC users. In total, the inventory included 36 speaking tasks and 29 writing tasks.

Some sample results are shown in Table 3 for speaking tasks and in Table 4 for writing tasks. For instance, 19% of test takers at the lowest TOEIC speaking score level said they could, easily or with little difficulty, make/change/cancel an appointment to see a person, while at the highest score level, virtually everyone (100%) felt that they could accomplish this relatively simple task with little or no difficulty. The task shown in the bottom row (Serve as an interpreter…), however, was considerably more difficult for test takers at each score level. Again, each task was considerably easier for test takers at each higher speaking score level.

TABLE 3

Example for Selected Speaking Tasks. Percentage of the TOEIC Test Takers, by Speaking Score Level, Who Said They Could Perform Various English Language Speaking Tasks Either Easily or With Little Difficulty

Task:	TOEIC Speaking Score						
	0–50	60–70	80–100	110–120	130–150	160–180	190–200
Make/change/cancel an appointment to see a person	19	32	43	65	78	91	100
Telephone a company to place (or follow-up) an order for an item	16	22	34	56	67	83	96
Have "small talk" with a guest about topics of general interest (e.g., the weather) before discussing business	10	24	35	57	69	83	94
Explain (to a co-worker or colleague) how to operate a machine or device (e.g., photocopier, PC, audio player) that I am familiar with	11	25	29	43	51	68	86
Serve as an interpreter for top management on various occasions such as business negotiations and courtesy calls	2	3	6	11	18	28	47

TABLE 4

Example for Selected Writing Tasks. Percentage of the TOEIC Test Takers, by Writing Score Level, Who Said They Could Perform Various English Language Writing Tasks Either Easily or With Little Difficulty

Task:	TOEIC Writing Score					
	0 - 80	90-100	110-130	140-160	170-190	200
Write an e-mail requesting information about hotel accommodations	12	37	48	70	81	91
Translate documents (e.g., business letters, manuals) into English	11	18	24	39	54	81
Write discussion notes during a meeting or class and summarize them	11	17	19	37	53	79
Write a formal letter of thanks to a client	11	20	24	38	50	71
Prepare text and slides (in English) for a presentation at a professional conference	7	11	17	28	44	69

The results from a representative sample of the 29 writing tasks exhibit much the same pattern. The complete results for the speaking and writing can-do study are available in Compendium Study 11.1 in this compendium and a report in the ETS Research Report series (Powers, Kim, Yu, Weng, & VanWinkle, 2009), which can be found on the TOEIC website, www.ets.org/toeic.

For each of the TOEIC tests, therefore, there is evidence that links scores to the likelihood that someone can perform, either easily or with little difficulty, a wide variety of everyday or workplace language tasks in English. This information gives meaning to the TOEIC scores in very practical terms and it demonstrates the value of the TOEIC scores to test score users.

As indicated earlier, validation is a never-ending process, and the process still continues for the TOEIC tests. A few of the questions that ETS would like to explore as future projects include:

- How do the TOEIC scores relate to job performance?
- How can we evaluate the trustworthiness of our validity criteria (including self-assessments)?
- How do the TOEIC scores change (improve) over time with relevant experience or instruction?
- How distinct are the speaking, writing, listening and reading sections? What is the unique value of each section score?
- How can we help raters/scorers focus even more consistently on valid aspects of test performance in order to improve their English proficiency?
- Do the tests measure the same things in different countries?
- Can computers automatically (and validly) score some aspects of speaking and writing performances?

To summarize, the main value of the TOEIC tests lies in their validity, which is the extent to which the tests do what we claim they can do - measure a person's ability to communicate in English in a workplace setting. The TOEIC tests yield valid scores in part because of the careful way in which they are designed. Further evidence of the validity of the TOEIC scores comes from special studies such as the can-do self-assessment studies described earlier. More evidence will continue to be generated to make a stronger and stronger case for the value/validity of the TOEIC assessments.

References

Messick, S. J. (1989) Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

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* (ETS Research Rep. No. RR-08-56). Princeton, NJ: ETS.

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 Rep. No. RR-09-18). Princeton, NJ: ETS.