論　文

A Review of TOEIC: A Critical Analysis of the Structure, Methods, and Purpose of TOEIC, Including its Effect on Teaching and Learning English as a Second or Foreign Language

Ryoko　MIYATA

INTRODUCTION

The TOEIC (Test of English for International Communication) is one of the largest-scale language tests in the world, administered internationally with almost two million test-takers every year. The test was created as a result of Japanese Ministry of International Trade and Industry requests to the ETS (Educational Testing Service) in the mid 1970's. Along with the TOEFL (Test of English as a Foreign Language) also developed by ETS, TOEIC is becoming more popular and is frequently used throughout Asia and Europe.

This paper is primarily a review of the TOEIC with particular focus on the test methods used in the different sections of the test. Initially the TOEIC test structure and general test information are briefly reviewed. It then goes on to summarize the key concepts in test design, such as reliability and validity, related to the TOEIC. The paper finishes by discussing potential advantages and disadvantages of test methods in relation to realization of test theory foundations.

TEST STRUCTURE

The TOEIC is a two-hour, multiple-choice test that consists of 200 questions. It is designed to evaluate the examinee's ability to "communicate" in English by measuring their receptive English skills. The general contexts from which TOEIC questions are taken are business settings (ex: corporate development, manufacturing, personnel) and real-life situations (ex: entertainment, health, travel, purchasing, dining out).

The listening section is divided into four parts. Candidates listen to a variety of statements, questions, short conversations and short talks delivered by audiocassette. In the reading section, two subsections evaluate the examinee's ability to use English grammar. The reading comprehension part uses business letters, advertisements, news

and so on, as stimuli.

SCORING PROCEDURE

Scores on the TOEIC test are determined by the number of correct answers. Statistical procedures are used to convert scores to a scale score, ranging from 10 to 990. The ETS does not publicize the procedures or the data in detail, but they claim that they have established a reliable system of scoring and evaluation including "item analysis" practiced twice (before and after marking procedure), and "equating procedure" to standardize the test forms.

According to the TOEIC, the stages of item analysis are as follows: 1) calculate percentage of correct answers to decide difficulty level of each question 2) divide the candidates into five groups in order of test results 3) compare the five groups' percentage of correct answers as to each item. The higher the item's difficulty level, the larger the difference between the correct answer percentage of the top and lowest groups should become, provided that the particular item functions well.

Equating procedure is based mainly on the analysis of common test items that are used in different versions of the test. Comparison is made between the new and the past test forms focusing on these recycled items.

BASIC CONSIDERATIONS IN TEST DESIGN

Reliability

According to the TOEIC official guide for candidates, TOEIC test users demand high reliability of their scores. Two components of test reliability are usually taken into account: the reliability of the scoring and the performance of candidates from occasion to occasion (Hughes, 1989). The ETS asserts that the TOEIC score scale is always kept constant and so is the test reliability. In order to make it a highly objective test, the ETS measures only listening and reading skills in the test. Direct tests of speaking and writing are generally not only less objective but also less reliable in terms of measurement of skills. Scoring reliability is very difficult to obtain in tests of productive skills, considering one needs to construct adequate mark scheme and standardize markers with it every time the test is done. This system is also less cost effective, especially in the case of large-scale testing like the TOEIC.

To achieve consistent performances from candidates, the test is administered under strictly uniform conditions every time. Very explicit test instructions, both recorded and written, are provided for candidates. Having many items (200 questions) also serves to maintain scoring reliability (we must say that it is a pretty inefficient way to test language proficiency).

As stated above in the section on scoring procedure, the test form always includes old (recycled) items. This is standard practice on many norm-referenced tests in the U.S. to standardize the test forms and to keep the testing system reliable. Especially in the

TOEIC Institutional Program (IP test), which is administered to a small group of candidates at a time, the test form consists of only old and good items that have been proved to test well in the past Secure Programs.

Construct Validity

A test is said to have construct validity if it can demonstrate that it measures just the ability that it is supposed to measure (Hughes, 1989). In the case of TOEIC, some difficulties in establishing construct validity may arise out of the attempts to test communication abilities by methods that measure only listening and reading skills. An optional interview test, TOEIC-LPI (Language Proficiency Interview), is available for those who scores above 730, but the normal TOEIC test only provides an indirect measure of productive skills, i.e. assesses the overall English ability by testing only receptive skills. The test seems to be constructed on the basic premise that the productive language abilities are proportional to the receptive abilities. However, in fact, communicative ability of "B level" candidates (about 730) varies a lot, from a fairly fluent speaker to a Japanese high school student level, showing asymmetrical abilities of productive and receptive skills.

Content Validity

Content validity is the representativeness or sampling adequacy of the content of whatever the test was designed to measure (Brown, 1996). Ideally, the sample of tasks and texts included in a test is as representative of the target domain as is possible. The content validity of a test may initially involve defining what it is that the testers wanted to assess (Brown, 1996). In the case of a proficiency test like TOEIC, which is designed to assess the ability to communicate in English, defining the target domain is disputable because the area of language from the sample is taken can be too broad. The behavior domains covered by the TOEIC test items reflect the users' purposes for taking the test (ex: recruiting, interviewing, job training), and as a result, it concerns "business English" and "real life communication" in terms of register and topic. It is crucial for a test to have content validity also because it is closely related to the concern of face validity.

Face Validity

"Test appearance" is a very important consideration in test use. In any language testing the bottom line is whether test takers will take the test seriously enough to try their best, and whether test users will accept the test and find it useful (Bachman, 1990). Face validity refers to what a test appears superficially to measure. If a test does not have face validity, it may not be acceptable to the candidate taking it, or the teachers and institutes who may make use of it (Weir, 1990). Still, it is not easy to find out who will find what type of test acceptable. We have seen the TOEIC becoming such a popular English test among Japanese learners and being commercially quite successful (the question of face validity concerns public relations, too). It certainly obtains surface credibility and public acceptability, being valid (both in terms of content and face) to test-

takers.

Criterion-Related Validity

This approach to test validity is concerned with the extent to which test scores correlate with a suitable external criterion of proficiency (Weir, 1990). Similarly, *predictive validity* concerns the degree to which a test can predict candidates' future performance (Hughes, 1989). The predictive power of the TOEIC interests the users greatly and the scores have been frequently used as alternative criterion at institutions. However, TOEFL, one of the highest-stake language tests, has been said to have very little predictive validity, so the TOEIC may not be much better because their test methods are similar in the most sections. What a test-taker with a certain TOEIC score can and cannot do with English is not identified clearly.

WASHBACK

Washback is the influence or the impact that tests will have on teaching and learning. Among language teachers and researchers there are claims that large-scale EFL tests lead to unnatural teaching, with students being taught inappropriate language learning and taking "test preparation" courses instead of "real" English courses. We have heard over the years that the TOEFL strategies, in particular, are taught in the courses that raise scores without providing students with the English they need in interaction or in the university courses they are entering (Alderson & Hamp-Lyons, 1996).

To look on the bright side, the TOEIC is designed as a "communicative language test" with non-academic English, therefore the washback on the language learning may not be as negative as in the case of TOEFL. Unfortunately though, the situation appears to have become comparable recently. TOEIC preparation courses or seminars are very popular among Japanese colleges. In these courses, test-taking strategies are taught, metalanguage is often used by the instructor, and actual test taking is very frequent. Ironically, it is unlikely that in the TOEIC preparation courses students spend much time on communicative tasks such as pair or group work.

Furthermore, the washback effect the TOEIC has on the test makers themselves is significant. The test results of high-stake EFL tests affect entrance to tertiary institutions, course evaluation and promotion at work. Therefore even minor changes in the test cause strong washback in terms of anxiety, fear, the production of new material and in the teaching and learning behaviors of those involved (Shohamy, Donitsa-Schmidt & Ferman, 1996). TOEIC scores have become influential in society and are beginning to be used as an alternative criterion of different English tests. TOEIC test developers are unable to make frequent changes when constructing tests (doing so would upset the customers), and consequently they use the same format year after year.

DISCUSSION OF TEST METHODS

Listening Comprehension

There are 100 questions on tape in four separate parts, which are typically ordered from easy to difficult. There is a thinking gap of 5-6 seconds after each question in the first two parts, and about 10 seconds in the part where the candidates are required to read multiple-choice questions while listening. Through all parts, note taking is prohibited and candidates are not allowed to write anything in the test booklet. This can be said to inhibit candidates' performance on the test and always seems inauthentic because we do take notes in real life even when we communicate in our mother tongue. The general register of the TOEIC listening subtest is "business", with a high frequency of idioms being spoken and relatively few polysyllabic words (Gilfert, 1996). The sample questions quoted below are not endorsed or approved by ETS or The Chauncey Group International Ltd.

Part 1. Picture Recognition

〈Presented orally〉

(A) The people are watching a yacht race.

(B) The seaside is crowded with people.

(C) People are relaxing at the side of the lake.

(D) The dog is chasing the woman. (ALC, 2003)

In this section, visual stimuli (in the form of photographs) are used. The candidates see a photo in the test book, hear four statements and choose the one that best describes the photo. The statements are relatively short and focus on testing the ability to recognize key vocabulary in the context of the photo. This section is the easiest and favorite part for most test takers probably because the photo prompt is very reassuring and motivating. At the same time, however, photographs often confuse them when they do not understand what the people in the photo are doing or they simply do not notice certain items presented in the photo. Cultural differences also frequently cause misinterpretation since the photos are mostly taken in English speaking countries.

By using a photo, the candidate's communicative competence (other than applying the grammatical rules of English) is effectively tested. They need to be able to employ the background knowledge and form a judgment on the information they interpret in order to get a right choice.

Part 2. Question-Response Evaluation

〈Presented orally〉

Q: What happened to those library books?

(A) Yes, I got them from the Central Library.

(B) I think the library is closed today.
(C) Aren't they in the living room? (ALC, 2003)

The candidates listen to a question and choose the appropriate response to it also spoken in English. Questions and responses are mostly short, yet some students feel that this is the hardest part since both the questions and possible responses are only heard, not printed; i.e. there is no information provided visually.

This section is a discrete structure test rather than a test of macro skills. The items are context-reduced and developed to measure specific components of language ability. There are 30 questions in total and they are not contextually connected at all. This is common practice to design a language proficiency test, but nevertheless seems to be rather an inauthentic way to design a communicative language test. Test takers' minds can hardly activate background knowledge nor make inference, listening to contextually disconnected short questions one after another. This low degree of contextualization is one characteristic that differentiates non-academic interactive language use from academic language use (ex: TOEFL) in the test (Bachman, 1990).

Part 3. Questions on Short Conversations

〈Presented orally〉

a : Do I look a little heavy to you?
b : Not at all. Why do you say that?
a : Well. I'm having trouble fitting into some of my clothes.

〈Printed〉

Q: Why is the man concerned?
(A) Because someone said he was heavy.
(B) Because he seems to have gained weight.
(C) Because he can't find some of his clothes.
(D) Because his friend can't lift him. (ALC, 2003)

In this section, candidates listen to 30 short discrete conversations, and this time, the question and four possible answers are printed in the test book. The written information may provide candidates more contexts to which they refer and fit the conversation they listen. However, not a few of the questions seem to be fragments cut out from a longer conversation. The candidates are required to have the skills to make inference fittingly about the background of the conversation and the relationship between the interlocutors with very little or partial information.

Moreover this method makes considerable processing demands on the candidates, as they have to pay attention to the input but they also have to read and retain four propositions in working memory, then soon they have to be listening for the next item (Brindley, 1998). This part has the reputation for being the most difficult part for many examinees.

Part 4. Questions on Short Talks

〈Presented orally〉

Tourism to Hong Kong is down, and that means hotels are reducing their rates in order to fill empty rooms. Although rates at many luxury hotels remain high, with a little looking some great bargains can be found. For budget travelers, guidebooks provide many listings and telephone numbers of cheap lodgings. If you'd like to stay in more upscale lodgings, the best advice is to go through a travel agent rather than calling a hotel directly. The hotel will usually quote you their "rack rate," which is an undiscounted rate.

〈Printed〉

Q: Why are hotels in Hong Kong reducing their rates?

(A) Because people are complaining that hotel rooms are too expensive.

(B) Because more hotels are being built.

(C) Because they have many empty rooms.

(D) Because people don't want to stay in luxury lodgings. (ALC, 2003)

The last part of the listening comprehension is the longer talk with printed questions and answers. Compared to the similar part of TOEFL listening, test takers may feel that the TOEIC materials are less difficult in terms of the length, content, and the number of questions (2-3 questions per talk).

Each talk is typically given by single speakers and recorded without pauses that should occur in natural speeches. This overly fluent speech makes each talk very one-sided and again, inauthentic. Although authentic samples of naturally occurring speech are unsuitable for use in official tests (Brindley, 1990), the seriously unnatural way of talking on the recording seems problematic especially in this section. Many Japanese examinees are rather weak at listening. Clearly, any test involving the processing of language at even moderate speed is likely to be difficult for candidates whose listening ability is weak. This difficulty and perception of speededness as well as the processing demand to listen, retain and read information simultaneously may lead to guessing, even blindly in this part. This part involves candidates to cope with speededness, which is an artificial condition of the listening task, and with speed of processing of working memory, which is a factor in test performance (Bachman, 1990).

Examples of the text types used here are announcements, directions, lectures, narratives, news broadcasting and advertisements, most likely created for the test. The content of talk is fictitious without exception so that no group of test takers will be favored by prior knowledge of the content.

Reading Comprehension

There are 100 questions in the reading section which are supposed to measure ability to recognize language that is appropriate for standard written English. The test format is practically the same as the comparative sections of the TOEFL, but the topics or genre

are significantly different. Part 5 and 6 take the form of a discrete-point test with short decontextualized items. The reading comprehension section adopts predominantly real-life English and business type topics, as in the listening comprehension section.

Part 5. Incomplete Sentences

I can't remember the last time I ----- such a wonderful play.

(A) saw
(B) seen
(C) had saw
(D) was seeing (ALC, 2003)

This part assesses the examinee's knowledge of English structure, grammar or vocabulary. This type of grammar section is very common in most large-scale language tests because large numbers of items can be answered, i.e. various components of language ability can be tested within a short period of time. In order to provide authentic language use and context in the test, this part could be designed as part of a paragraph (ex: cloze test), which is commonly adopted in many university entrance exams. The provision of a detailed context, however, often limits the range of grammatical features being tested (Heaton, 1988), so the TOEIC method may be suitable for its test purpose. Moreover, in terms of content validity, it is crucial not to limit the number of topics, style and vocabulary presented in a single form of the test.

Part 6. Error Recognition

In this section, each question contains four words (or phrases) underlined to call the candidates' attention. They are to select the one that is incorrect or unacceptable grammatically.

Thank you for providing (A) me with these informations (B) and for giving (C) me such useful advice. (D)

(ALC, 2003)

Students who complain that this part is difficult to answer tend to concentrate only on the underlined words, which disables them from comprehending the meaning of the whole sentence and recognizing the sources of error. This is an unnatural way to process language, especially when testing communicative skills. We read for meaning, not to look for errors. On the other hand, the grammar foundation that is necessary for interaction may be tested very effectively by this method. One needs a strong grammar base to do well in this section, in other words this method discriminates well. Still, this method emphasizes the more negative aspects of language learning, and it is also undesirable for language learners to be exposed too much to incorrect forms (Heaton, 1988). It is quite unlikely that test takers interact with the language input and process

them using communicative skills when they take this type of grammar test.

In terms of washback on English classrooms, this particular method appears to have quite a strong negative effect. Teachers may be using more time for test-taking and teaching "tips" to draw correct answers instead of empowering students to communicate with English.

Part 7. Reading Comprehension

It's easy to earn miles for free travel worldwide with Southeast Airlines' Flightways, the frequent-flier program rated #1 by Air Traveler magazine. For example, just two round-trips between Tokyo and Los Angeles will give you enough miles for a free trip to Hong Kong, Singapore, or Kuala Lumpur. You can earn miles on each Southeast Airlines flight you take, as well on flights with our partner airlines. You can also earn miles for your hotel stays, car rentals, long-distance phone calls, and credit card purchases.

Q: Who said that Southeast Airlines had the best frequent-flier program?

(A) Flightways

(B) Air Traveler magazine

(C) A famous credit card company

(D) A survey of airline passengers (ALC, 2003)

The last part of the test consists of approximately 15 different reading passages followed by 2-3 questions each. The principal purpose of the TOEIC is to test English ability to be used in business settings, so the genres of reading passages typically include advertisements, business letters, brochures, research, news reports, and questionnaires. In this section the tester is concerned with candidates' knowledge about forms of business documents as well as their skills of reading comprehension. One of the difficulties test takers face when they study for the TOEIC, however, is genre unfamiliarity and dullness of texts caused by the text nature. It is not easy to find something interesting or beneficial to know in the materials.

In a large scale language testing situation, the tester often attempts to minimize the possible effects of differences in prior knowledge of context that test takers might have (Bachman, 1990), i.e. reduces the likelihood of test bias, where some test takers might have an unfair advantage by being familiar with particular topics. For the TOEIC, the tester attempts to do so by presenting a fairly large number of originally written texts and asking relatively few questions per passage. As can be seen from the sample question above, however, there are passages in which test takers are required to have basic knowledge about international business (and sometimes American culture and society), which is an important aspect of the TOEIC. For example, candidates who are familiar with "mileage service" in the U.S. might have an advantage over others in this case.

Many TOEIC preparation course texts teach strategies and techniques to read for finding answers quickly and yet correctly. In order to read all the passages in the test format and answer the questions within the testing hour, one needs to have skillful

reading strategies: skimming for the gist and scanning for particular information. Quite a large number of the questions ask for specific information in the text, and they can be spotted without overall understanding of the passage. Learners are not likely to be reading for meaning nor for joy in the TOEIC reading course, but on the other hand the teacher can make them be aware of purposeful reading, which is one of the communicative language abilities the TOEIC seeks to assess.

CONCLUSION

In the development and use of language proficiency tests, reliability and validity are the most important qualities to consider. The TOEIC has been accepted in society as a reliable and valid criterion of English proficiency evaluation, confirming its validity as a large-scale language test. However, we need to call its construct validation into question in using and interpreting the test. The TOEIC tests receptive skills and only estimates abilities of productive language use, so it is possible for some test takers to score very highly on the test, but to be unable to use oral or written English in real life. They may end up becoming experts in getting high scores, and not know how to use English in context. The gap between the learners' language skills measured on the test and the skills shown in real life communication still seems significant.

Nevertheless, the TOEIC score is one of the tangible goals for EFL learners, and a number of learners in Japan study especially for the test and take it several times a year. They often do so at the expense of another opportunity or course to study English. In terms of the washback effect of this "TOEIC preparation" courses and materials, the test affects both what and how teachers teach, and consequently learners involved in such a course are not working communicatively in English.

References

ALC. (2003). *"Hajimeteno hitono TOEIC"*, Space AIC,
〈http://www.alc.co.jp/eng/toeic/first/index.html〉, (20, October 2003).

Alderson, J.C. and Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language testing* 13, 280-297.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Brindley, G. (1998). *Annual Review of Applied Linguistics* 18, 171-191.Cambridge: CUP.

Brown, J.D. (1996). *Testing in language programs.* Upper saddle river, NJ:Prentice Hall Regents.

Gilfert, S. (1996) "A review of TOEIC". July 1996. The Internet TESL Journal, Vol. II, No. 8, 〈http://iteslj.org/Articles/Gilfert-TOEIC.html〉 (20, October 2003).

Heaton, J.B. (1988). *Writing English language tests.* NY: Longman.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: CUP.

Shohamy, E., Donitsa-Schmidt, S., and Ferman, I. (1996). Test impact revisited :washback effect over time. *Language testing* 13, 299-317.

Weir, C.J. (1990). *Communicative Language Testing.* Prentice Hall International.