TOEIC Listening Comprehension Test and the Qualities of Test Usefulness

Masako Ishikawa

Abstract

The Test of English for International Communication (TOEIC) is widely used in important decision-makings, such as employment and admission for schools. Even though preparation for the TOEIC test is often discussed and materials that prepare potential test-takers to get high scores on the test are abundant, the qualities of the TOEIC test itself do not seem to have gathered as much attention. In this context, this paper investigated one of the qualities of the test: its usefulness, which is considered to be the most important quality of a test. Based on the framework of test usefulness proposed by Bachman and Palmer (1996), the usefulness of the listening section of the TOEIC test is examined in terms of the six qualities: reliability, validity, practicality, authenticity, interactiveness, and impact.

Introduction

The Test of English for International Communication (TOEIC) is an English language proficiency test organized by Educational Testing Service (ETS). According to the *TOEIC User Guide* (ETS, 1999), "It measures the everyday English skills of people *working* [italics added] in an international environment" (p. 4). Unlike the Test of English as a Foreign Language (TOEFL) test, which caters to candidates whose interest is to use English in *academic* settings, the TOEIC is a test of English proficiency in *business* settings. The difference between these two tests appears to originate from time of their initial creation. Though it does not seem to be a well-known fact, the TOEIC test was developed by ETS with a request of the Japanese Ministry of International Trade and Industry (MITI) in 1979. MITI requested ETS to develop an English proficiency test whose primary purpose was "to determine the proficiency levels of

employees, or potential employees, for human resource planning and development in the contexts of *business, industry, and commerce* [italics added]" (p. 2). Although MITI did not seem to have planned to develop a test that would be used worldwide back then, ETS claims that the TOEIC test is now taken by more than five million people every year throughout the world. As the name of the test demonstrates, "TOEIC test scores indicate how well people can *communicate* [italics added] in English with others in the global workplace." (p. 4).

The TOEIC test is a two-hour multiple-choice test that consists of 200 questions divided into 100 questions each in listening comprehension and reading comprehension sections. Test-takers receive sub-scores which range from 5 to 495 for each section and total scores which range from 10 to 990. Some revisions to the test were implemented in 2006; however, basic formats have not changed. The TOEIC listening section consists of four parts: photographs (ten questions), question-response (30 questions), short conversations (30 questions), and short talks (30 questions). Focusing upon the TOEIC listening comprehension section, this paper attempted to investigate if the TOEIC listening section measures the everyday English skills of people working in an international environment as ETS claims. For this purpose, usefulness of the TOEIC test was examined. According to Purpura (2004), a good or useful test enables us "to elicit scorable behaviors from which to make trustworthy and meaningful inferences about an individual's ability" (p. 148). Moreover, Bachman and Palmer (1996) argued that the most important thing to consider in designing and developing a language test is "the use for which it is intended" (p. 17), maintaining that "the most important quality of a test is its usefulness" (p. 17). Following this argument, the TOEIC test listening section will be examined by the framework of test usefulness proposed by Bachman and Palmer. This paper addressed the following research question:

Is the test usefulness achieved in the TOEIC listening section?

Usefulness of the TOEIC Listening Section

Even though a test is one of many educational components, such as teaching materials and learning activities, there is a significant difference between a test and those other components in their purpose. That is, the primary purpose of other educational components is to promote learning, whereas that of a test is to measure learners' ability. Accordingly, qualities which have to be considered in determining the test usefulness are quite different from those of other educational components. Therefore, in order

to determine the overall usefulness of a test, some specific qualities that are unique to a test have to be considered. As stated above, Bachman and Palmer (1996) maintained that the most important of such qualities is test usefuleness, proposing a framework of test usefulness which consists of six qualities: reliability, validity, practicality, authenticity, interactiveness, and impact. The researchers maintained the importance of "the optimal balance among the qualities," (p. 40) for tests to be useful. Based on this framework, each quality will be addressed in the following section in order to examine if an optimal balance is achieved in the TOEIC listening section.

Six Qualities that Construct Test Usefulness

1. Reliability

The first quality, reliability, refers to consistency of measurement. As previously mentioned, all the questions are in the form of multiple-choice. Thus, it can be argued that the TOEIC test is highly reliable because of the nature of its test task which enables objective scoring as opposed to tasks, such as essay writing, which have to depend on subjective scoring to some degree. The *TOEIC User Guide* (ETS, 1999) admits, "no test measures performance with perfect accuracy and consistency," which is in line with Bachman and Palmer (1996). However, it also says that candidates who take several TOEIC tests in a short period of time receive "a number of scores that center around an average value known as the 'true' score" (p. 9) and that the scores are within 25 points of the true score two-thirds of the time, which appears to demonstrate the high reliability of the TOEIC test.

2. Validity

The second quality, validity, pertains to the degree to which a test measures what it is designed to measure. Stressing the significance of validity, Messick (1989, as cited in Powers, 2010), a former vice president for research at ETS, defined it as "an integrated evaluative judgment of the degree to which empirical evidence and the theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores." The *TOEIC User Guide* (ETS, 1999) also highlights the notion of validity as well as reliability, on the ground test scores are strongly related to those of other English proficiency tests. According to the *TOEIC Technical Manual* (ETS), a study conducted in 1999 examined the relationship of scores between the TOEIC and the TOEFL listening sections and found that the scores correlated very highly (r = .88),

"indicating that the Listening section of the TOEIC test is indeed an accurate indicator of a candidate's ability to comprehend spoken English." (p. 16). Referring to this result, the *TOEIC Technical Manual* (ETS) maintains high construct validity of the TOEIC listening test. In terms of content validity, it is stressed that every effort has been made for the test to be unbiased so that it could accommodate all the test-takers with different cultural backgrounds. One of the changes made in 2006 was to use different English accents, including not just American, but also British, Canadian, and Australian, which seems to demonstrate one of these efforts. It is also stressed that the test does not require the test-takers to have special prior knowledge of business terms (Powers, 2010).

3. Practicality

The third quality, practicality, is related to the ways in which tests are implemented. The TOEIC test appears to be highly practical in that all the questions are in the form of multiple-choice, which enables scoring by machines. This leads to the saving of both time and human resources in terms of scoring. As for material resources, the TOEIC is a paper-and-pencil test, and the listening section requires just a CD and a CD player for all the test-takers in each room. Therefore, it appears to be more practical than, for example, the TOEFL Internet-based test (ibt) listening section, which requires a computer for each test-taker. This difference seems to be reflected in the difference in fees. Namely, it costs a little less than 6,000 yen (approximately $60) to take the TOEIC test, whereas the fee for the TOEFL test for test-takers in Japan is $200 in the case of regular registration. (Fees vary depending on the country a test is taken).

4. Authenticity

The fourth quality, authenticity, refers to the correspondence between the characteristics of target language use (TLU, Bachman & Palmer, 1996) tasks and those of the test task. It should be noted that authenticity here does not refer to authenticity of the materials (topic and language), which is often employed to judge the authenticity of educational components, including but not limited to tests. In terms of authenticity of the materials, the TOEIC test seems to be relatively authentic. However, the test task (i.e., a multiple-choice question) does not seem to correspond to test-takers' TLU tasks. As stated in the introduction, the aim of the TOEIC test is to measure test-takers' communicative ability in the workplace. Nonetheless, it would be highly unlikely for test-takers to encounter multiple-choice questions in everyday communication, whether workplace or not. As such, it seems to be hard to infer their communicative ability by

just asking them to choose appropriate choices without giving them any output opportunities. Thus, authenticity of the TOEIC listening test appears to be relatively low in spite of its stated purpose.

5. Interactiveness

The fifth quality, interactiveness, is defined as "the involvement of the test-taker's individual characteristics in accomplishing a test task." (Bachman & Palmer, 1996, p. 25) Interactiveness of the TOEIC listening section does not seem to be as low as its authenticity, because the content of the test appears to be engaging for test-takers whose interests are in business settings. Moreover, test-takers who might not be necessarily interested in business settings might get interested in the content. The characteristics of test-takers (i.e., topical knowledge, and affective schemata) may interact with the test task, namely, multiple-choice questions, while they work on the questions using their language ability.

6. Impact on test constituents

Finally, the sixth quality is impact on test constituents (i.e., test-takers, corporations, schools, society, and educational systems). As stated before, the TOEIC is a large-scale test of proficiency in business settings. As such, the scores are used for decision-makings (e.g., employment, promotion, admission) by a great number of corporations and schools. In that sense, the TOEIC listening test could be said to have a considerable impact on test constituents. The impact on test-takers can be positive and negative; test-takers receive single scores (i.e., summative information) for the whole listening section, but they do not receive breakdown of scores for each subsection, which would provide formative information. Thus, if they are satisfied with their scores, the impact may be positive. In contrast, if they are not satisfied, the impact could be negative, as the test does not give any detailed information for the test-takers to improve their abilities after the test. Moreover, the impact on corporations, potential employers of test-takers', can be also positive and negative. That is, if a test score enables corporations to infer a test-taker's ability accurately, the impact could be positive. On the other hand, it could be negative if it does not, which is closely related to the usefulness of the TOEIC test.

Discussion

The usefulness of the listening section of the TOEIC test was examined based on the framework of test usefulness proposed by Bachman and Palmer (1996). The research question addressed was: Is the test usefulness

achieved in the TOEIC listening section? As investigated above, it seems to be safe to say that the test usefulness is achieved overall in the TOEIC listening section. Nonetheless, the balance of six qualities which are the components of test usefulness may not be necessarily optimal; namely, not all the components appear to have attained a high level of quality. The four qualities (reliability, validity, practicality and impact) were found to be fairly high, whereas the other two qualities (authenticity and interactiveness) do not seem to be as high. Especially, authenticity appears to be relatively low as the test task in the form of multiple-choice questions does not correspond to test-takers' TLU tasks, which would involve communication with their colleagues and clients, in spite of the purpose of the TOIEC test. An assessment of test-takers' communicative ability is aimed at in the TOEIC listening section, however, it would be difficult to elicit their communicative ability from the current test task. Meanwhile, modifying the test task to improve authenticity would affect other qualities, such as practicality, negatively. Though it is not the focus of this paper, an introduction of the TOEIC speaking and writing tests in 2006 might have been an example of the efforts to compensate the low authenticity in the listening test in that they provide test-takers with output opportunities, which is likely to provide more reliable information on test-takers' communicative ability.

Finally, though it is not the focus of this paper either, the test use has to be considered here as well. As stated in the introduction, the TOEIC test is designed to assess test-takers' communicative ability in the global workplace. Therefore, taking these tests for potential employment would be an appropriate use; on the other hand, it would be a misuse if they were taken by test-takers whose main purpose is to study abroad. There are language tests such as the TOEFL and Academic Training of the International English Language Testing System (IELTS) organized by the British Council which cater to these test-takers. (General Training of the IELTS is designed for employment.) A pedagogical implication drawn from this is that considering their use and choosing an appropriate test for one's purpose would be crucial in order to expect maximum effect from language tests. Accordingly, language teachers are expected to be facilitators of their learners in making appropriate choices.

Conclusion

The examination of the qualities of test usefulness revealed that the TOEIC listening section is overall useful. Though the balance among the qualities does not appear to be perfectly optimal, given that there is no test

that measure test-takers' ability with perfect accuracy and consistency (Bachman & Palmer, 1996), it would be safe to say that the TOEIC listening section has attained a reasonably good balance.

References

Bachman, L. F. & Palmer, A. (1996). *Language testing in practice*. Oxford: OUP.

Educational Testing Service *The Technical Manual* (n.d.). Retrieved September 23, 2010, from http://www.toefl.org

Educational Testing Service (1999). *TOEIC User Guide*. Retrieved September 23, 2010, from http://www.toefl.org

Powers, D. E. (2010). Validity: What does it mean for the TOEIC tests? *ETS Research Report*. 1-11.

Purpura, J. (2004). *Assessing grammar*. Cambridge:CUP.