

**Milan Milanović, Ana Milanović**  
*Filološko-umetnički fakultet, Univerzitet u Kragujevcu*

## **TESTING LISTENING IN THE INTERNET-BASED TEST OF ENGLISH AS A FOREIGN LANGUAGE<sup>1</sup>**

In this paper, the Listening section of the Internet-based Test of English as a Foreign Language is analyzed. Based on theoretical attempts to define a construct of listening and on the statement of purpose provided in the ETS's publications, the intention of this paper is to identify the construct, or in other words the abilities necessary for comprehend and analyze an utterance. TOEFL iBT measures a test taker's ability to understand the English language as is heard in an academic setting, thus providing evidence of their ability to understand and participate in lectures and conversations. Therefore, this paper identifies three purposes for academic listening: listening for basic comprehension, listening for pragmatic understanding, and listening for connecting and synthesizing information. Based on these three purposes, there are identified skills and abilities which are constituent part of the construct, which is then operationalized by input materials and test tasks in a language test. The step of test operationalization involves selecting audio materials and formulating test tasks, which in the case of TOEFL iBT take the form of a selected response. Finally, there are certain limitations identified, which are inherent to assessing language skills, as well as to assessing language through computer technology. However, despite the limitations, it is emphasized that assessing listening represents an important segment of overall language assessment, because listening is essential feature of communication. A wider implication of this paper refers to the insight Serbian test developers may benefit from when attempting to develop a test of listening in entrance or any other type of language exams.

**Key words:** construct, language skills, listening comprehension, test question format, TOEFL iBT

### ***1. Introduction***

Listening comprehension is a very complex process whose relevance in language learning and testing is more than obvious. However, although

---

<sup>1</sup> Kraća verzija ovog rada izložena je na Trećem naučnom skupu mladih filologa, održanom 12.3.2011. godine na Filološko-umetničkom fakultetu u Kragujevcu.

test developers, teachers and test users intuitively share the consensus of what constitutes comprehension of an utterance, they invariably find it difficult to define and measure the construct of listening. The construct-related issues are of central interest in language assessment, and consequently they are equally important in assessing any of the language skills – reading, listening, writing, and speaking.

Language tests are usually intended to measure one or more constructs, so that inferences based on test results may be used for making certain decisions about the candidates' competence to effectively use the language in question. These decisions may vary significantly depending on the test purpose. In high-stakes language tests, decisions made on the basis of test scores tend to be of great influence on test takers' future life and career. For this reason, it is important for test constructs to be well defined and accordingly operationalized. In this paper, the Listening section of the Internet-based Test of English Language (TOEFL iBT) will be analyzed for the purpose of defining the constructs of listening, identifying the input materials and question formats used to operationalize the construct(s).

This analysis may be helpful to language instructors as well as to test developers for several reasons. First of all, the impact that international standardized tests may have on language instruction has been recognized by researchers in language assessment (see Wall and Horak, 2006, and 2008) as well as by language instructors who either prepare candidates for taking some of the standardized high-stakes tests or teach English by assessing language skills in the way these skills are assessed in high-stakes language tests. When it comes to Serbian English language teaching community, philological departments administer entrance exams at B2 levels of the Common European Framework of Reference<sup>2</sup>. These exams include assessing structures, and are often accompanied by tests of listening and writing. The purpose of entrance exams is to ensure that students are capable of attending classes where the language of instruction is English, which coincides to some extent to the purpose of TOEFL iBT. Besides, TOEFL iBT is claimed to be placed at the same level of the CEFR, and for this reason a deeper understanding of construct definition, test operationalization and test administration as applied in this test, may be of assistance to test developers and language teachers in Serbian ELT community.

This paper is based on the information available in the ETS's publications targeting prospective test takers and language teachers

---

2 For language proficiency scales and scale descriptors see North and Schneider, 1998; and Council of Europe, 2001.

with the purpose of reflecting on the type of listening constructs the test is intended to measure as well as on the question formats utilized to assess test takers' listening skills (Jamieson et al., 2000; ETS 2007). The corpus for analysis includes 36 sample subtests (Listening sections only) authored by either ETS or Barron's TOEFL preparation team. The examples provided here are available in free online practice materials (through ETS's website available at <http://www.toefl.org>).

## **2. About TOEFL iBT**

The Internet-based Test of English as a Foreign Language is taken worldwide and is administered in certified test centers via fast, secure and reliable Internet connection. This test measures test takers' ability to communicate effectively in English, in a university setting, which means that input materials are like those which can be found at any given university in North America, and that test tasks correspond to a large extent to the tasks a language user may be facing at a campus. Developers of TOEFL iBT envisage that the test measures not only receptive, but also productive language skills, either separately or integratedly<sup>3</sup>. The integration of skills is achieved by presenting input materials in visual and/or audio formats, or, in other words, presenting input materials in the form of reading and/or listening passages in the Speaking and Writing sections. Apart from being presented as input materials in integrated speaking and writing tasks, reading and listening comprehension (and not only comprehension) are assessed individually in the corresponding sections (for more about the TOEFL iBT see ETS, 2006; ETS, 2007, and Alderson, 2009). In this paper, only the Listening section of the test will be analyzed, although it seems obvious that integrated tasks utilizing listening passages also measure the construct of listening.

## **3. How to define a construct?**

Language tests measure one or more constructs, which need to be defined and operationalized in line with the purpose of assessment. The process of defining a construct is a relatively complex task, because there are many factors which need to be taken into consideration. According to Buck (as cited in Milanović, 2011), there are at least three possible ways to define a construct of listening. The first one of them is taken when test developers identify the competences test takers should possess, and create such tasks which should engage the corresponding competences. An alternative to this so-called competence-based approach is that of

<sup>3</sup> For more about testing skills separately and integratedly, see Powers (2010).

creating test tasks which correlate with real world language use and tasks. This is possible only when the corresponding real world situations are familiar and already identified. The third approach combines the first two (Buck 2001:102) and makes use of both competences deemed relevant for a certain language use situation and tasks whose completion is relevant in a particular situation.

Whatever approach test developers choose to take, it will be based on the following sub-steps: determining the purpose of assessment and identifying target language use situation. When it comes to purpose of assessment, TOEFL iBT is a test which aims at measuring test takers' ability to communicate effectively in a university setting. The inferences based on test results are intended to provide the evidence of test takers' communicative ability in English. Knowing what the purpose of assessment is helps us identify target language use situations. Clearly, these situations will be characteristic of a university campus, making it easier for a test developer to come up with a set of test tasks to address the target language use and the purpose of assessment. In the following chapters, I will discuss the purposes for academic listening as identified and tested in the Listening section of the test, input materials, and question formats used to measure the construct of listening. Listening passages and tasks, and questions related to the passages are the means by which the construct(s) of listening is (are) operationalized.

#### **4. Academic listening**

According to ETS, academic listening is done for the following purposes: Listening for basic comprehension, Listening for pragmatic understanding, and Connecting and synthesizing information (2007:12). These purposes are typical of academic setting and should therefore be addressed in a listening assessment by creating test tasks which measure test takers' ability to participate in the listening process in a target language use situation. Each of the three purposes can be broken down into a number of features which explain the purpose in more detail.

**Listening for basic comprehension** involves comprehending the main idea of an utterance, comprehending major points and important details which are related to the main idea. In this case, a listener may not comprehend all details, but they should be able to "catch" the most important ones.

**Listening for pragmatic understanding** refers to the listener's ability to recognize a speaker's attitude and degree of certainty, or their function or purpose. As the name suggests, the abilities required to meet this purpose are of pragmatic nature.

**Connecting and synthesizing information** include a variety of skills, some of which are of pragmatic nature. For example, a listener should be able to recognize the organization by which a piece of information is presented, which is obviously relevant in all classroom activities, during seminars, and presentations. Making connections among pieces of information is part of this ability, as is the pragmatic skill of making inferences and drawing conclusions based on what is implied in an utterance. This incorporates the ability to understand the relationship between ideas, which may include recognizing steps in a process or cause/effect relationship between the ideas, or it may call for making more subtle distinctions, such as those required in the processes of comparison and contrast. Apart from this, a listener should be capable of recognizing various discourse features, such as those of topic changes, allusions, digressions, aside statements, etc.

Now that listening purposes are identified, it may be argued that the test of listening should address these purposes by presenting test takers with authentic listening materials and tasks/questions to elicit test takers' responses proving that they are able to understand spoken English in an academic setting. The request for authenticity in input materials is part of the concern for a test's overall usefulness and construct validity, but due to the spatial limitations authenticity as well as test usefulness will not be discussed in detail here (for more on test usefulness see Bachman, 1990; Bachman and Palmer, 1996; and Chapelle and Douglas, 2006). However, what I find worth mentioning here, given that TOEFL iBT is a language test delivered via the Internet, is the issue of not exactly measuring the construct which is defined through the process of identifying the purposes for academic listening. Two common traps alluring test developers in their effort to operationalize the construct are known as construct-underrepresentation and construct-irrelevant variance (Messick, 1989 as cited in Buck, 2001). The former refers to incompleteness in addressing all parts of the construct by the input materials and corresponding tasks which should engage relevant skills and abilities defined as the construct of listening. The latter is the case of putting validity at risk by (unwillingly) requiring from test takers to demonstrate skills and abilities which have not been defined as the construct. For example, in computer-assisted language test, computer (il)literacy may interfere with actually responding to test task (in this case with listening to a passage and responding to test questions), which will consequently affect test scores and finally inferences made on the basis of the scores.

## 5. Input

The input in the Listening section may be analyzed by means of test task characteristics framework in computer-assisted language assessment (Milanović, 2010), which relies to a large extent on the framework first developed by Bachman (1990), and then modified by Bachman and Palmer (1996, as well as on the frameworks adapted from these two (see Alderson, 2000; Douglas, 2000; Chapelle and Douglas, 2006). The input in the Listening section refers to the listening passages, and test questions or items which can be described in terms of their format and language. The format of the input includes the characteristics of channel, form, language, length, degree of speededness, and vehicle. Question formats are added here to complement the input suggested in the test task characteristics framework. Language of the input can be analyzed in terms of its organizational, pragmatic, and topical characteristics.

### 5.1 Format of the input in the Listening section of the TOEFL-iBT

In the Listening section of the TOEFL-iBT, listening passages, accompanying instructions and test questions are delivered in both visual channel (on computer screen) and aural channel (through the headset), in target language via the Internet. The form of the input is language, accompanied by still images, graphs and charts whose purpose is to raise the levels of contextualization and interactional and situational authenticity. There are two types of listening passages: academic lectures and conversations, whose number and length vary. In short version of the Listening section, there are 4 lectures, whereas in long version the number of lectures is 6. Each lecture is 3-5 minutes long and contains between 500 and 800 words. Each lecture is followed by a set of 6 questions/items, which can be referred to as a testlet, because it is related to the same listening passage and is used to assess various aspect of the listening construct in the passage. The number of conversations per sub-test, i.e. section, is 2 in the short version and 3 in the long version of the Listening section. The conversations consist of 12-25 exchanges followed by 5 questions related to the conversation that preceded them. Individual lectures and conversations are played only once, although there are questions in the related set, in which particular segments of the lecture or the presentation is replayed for the purpose of assessing test takers' pragmatic knowledge. Test takers are not presented with questions in advance, but after a passage has been played, though note taking is allowed throughout the section. The Listening section of the TOEFL-iBT is speeded, because test takers are supposed to listen to

the passages and respond to tasks within the pre-determined timeframe which is not under their own control. The input of the Listening section is “reproduced”, or, in other words, played on computers in a testing center. Task types in the Listening section of the TOEFL-iBT are mostly questions or items in the Bachman and Palmer’s sense, i.e. they elicit selected responses, and will be discussed within Question formats.

### 5.1.1 Question formats in the Listening section of the TOEFL-iBT

There are four question formats in the Listening section of the TOEFL-iBT:

1. Traditional multiple-choice questions with four answer choices and a single correct answer.

Example:

Who is buying new jackets for the team?

- The coach
- The captain of the team
- A former player
- A group of basketball fans

2. Multiple-choice questions with more than one correct answer. When there is more than one correct answer, test takers are prompted to “click on two answers” (Figure 1).

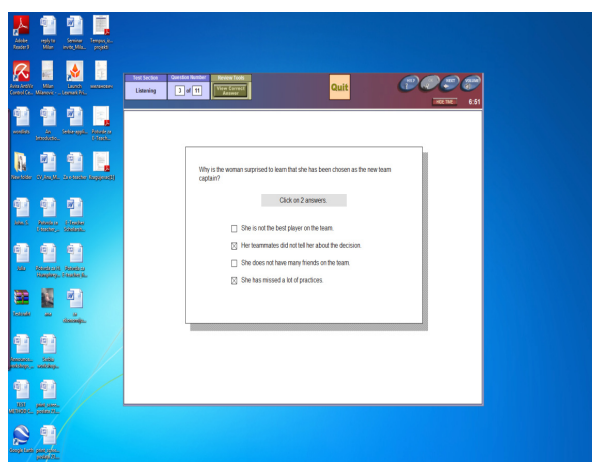


Figure 1: Multiple-choice question with more than one correct answer (ETS, 2007)

3. Sequencing items which require test takers to order events or steps in a process.
4. Questions which require test takers to match objects or text to categories in a chart.

## 5.2 *Language of the input in the Listening section of TOEFL iBT*

The Listening section measures test takers' ability to understand the English language spoken in academic settings. The language heard in passages is predominantly standard American English, with speakers making utterances clearly with only occasional hesitations typical of unplanned spoken language. This section assesses grammatical knowledge, as well as pragmatic and discourse knowledge. Test takers are often required to make inferences with regards to the speaker's attitude and degree of certainty. They also need to focus on details related to the main idea, which they may find difficult to do because they are not allowed to see the questions beforehand. Topical characteristics of the input refer to the type of information that is in the input, which in the case of the Listening section falls into domain of academic discourse. The lectures take the form of monologues, group discussions, or monologues interrupted by a student making a comment or answering a question. The topics covered in this section are related to various academic fields, mostly related to humanities. The second category of materials consists of short interactions between two or more students, a student and his academic advisor, or a student and a campus staff.

## 6. *Limitations to listening assessment in TOEFL iBT*

As is the case with any other language test, TOEFL iBT is not resistant to certain limitations. Language samples in the listening passages of the Listening section are recorded using a standard variation of American English, and as such they do not represent all the possible language varieties that can be encountered at a university in North America. The language of both expository materials and interactions seems deliberate and somewhat unnatural, with no attempts to assess test takers' comprehension of informal language. This was the case, if to a somewhat larger extent, with previous versions of the test which made some researches argue that the test suffered from construct underrepresentation (See Buck, 2001:223). The format of questions used to assess test takers' comprehension is limited to multiple-choice questions which is disputable for at least two reasons. First, it requires that test takers select rather than produce their responses to test tasks. Given that test tasks should correspond to target language tasks in real world situations, it may be argued that although in many situations listeners do no more than just listen, in other situations they are required or prompted to respond to input. This seems particularly true for students attending classes at university or participating in conversations



at a campus, and if we recall the statement of purpose underlying this test, we will see that TOEFL iBT aims at assessing test takers' ability to communicate effectively in English in a university setting. In line with this statement of purpose it may be argued that students at any campus are required to respond to what they have heard, using not only selected but also extended responses in Bachman's sense. Second, the nature of multiple choice question format implies that there is a number of options, which a candidate is prompted to select in order to complete a task by choosing a correct or best answer. It is very difficult to imagine anyone in the real world asking themselves or even being asked such things as test takers are in a listening assessment. Finally, a computer delivery poses limitations of another kind. In target language situations, people seldom use computers to listen to lectures or conversations. However, an exception may be found in e-learning courses where lectures are given online in real time or in the form of audio-video clips, which may be argued to contribute to TOEFL iBT test task authenticity (for more on computer applications in language assessment see Chapelle, 2001, and Chapelle and Douglas, 2006).

## 7. *Conclusion*

The Listening section of the Internet-based TOEFL assesses test taker's ability to understand spoken language. There are identified three purposes for academic listening: listening for basic comprehension, listening for pragmatic understanding, and listening for connecting and synthesizing information. These three purposes form a basis for identifying the constructs of listening which is then operationalized through input materials and test tasks. Input is delivered in audio format, using computer equipment and reliable Internet connection, directly to a test taker's headset. Input can be additionally described in terms of its format and the language, according to the test task characteristics framework, as suggested by a number of researchers. Test tasks take the form of multiple-choice questions, with one or more correct answers, which require from test takers to select them in the manner specified in test instructions. All the responses in the Listening section are selected and to provide them test takers need not possess any background knowledge. Apart from the benefits of testing the skills of listening in a single assessment, this kind of assessment is somewhat limited due to the constraints of computer-assisted language assessment. However, limited as it is, listening assessment is important, not only in high-stakes language assessment, but in language instruction as well, since it goes without say-

ing that a learner's communicative ability would be incomplete without their ability to understand the spoken form of the target language.

## References

- Alderson 2009: C. J. Alderson, Test review: Test of English as a Foreign Language TM: Internet-based Test (TOEFL iBT®). *Language Testing*, 26 (4), 621-631.
- Alderson 2000: C. J. Alderson, *Assessing reading*. Cambridge: Cambridge University Press.
- Bachman 1990: F. L. Bachman, *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, Palmer 1996: F. L. Bachman, S. A. Palmer, *Language testing in practice*. Oxford: Oxford University Press.
- Barron's TOEFL i-BT Test preparation with CD, (2008)
- Buck 2001: G. Buck, *Assessing listening*. Cambridge: Cambridge University Press.
- Chapelle 2001: A. C. Chapelle, *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.
- Chapelle 2006: C. Chapelle, D. Douglas, *Assessing language through computer technology*. Cambridge: Cambridge University Press.
- 2001: Council of Europe, *Common European Framework of Reference: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Douglas 2000: D D. ouglas, *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- ETS 2006: TOEFL iBT: Sample Questions. Available at <http://language-testing.info/articles/artltt.html>. Last accessed January 11, 2011.
- ETS 2007: *TOEFL iBT tips. How to prepare for the TOEFL iBT*. Available at [http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL\\_Tips.pdf](http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_Tips.pdf). Last accessed August 11, 2010.
- Jamieson 2000: J. Jamieson, S. Jones, I. Kirsch, P. Mosenthal, C. Taylor, TOEFL 2000 Framework: A Working Paper Educational Testing Service, Princeton NJ.
- Messick 1989: S. Messick, Validity. In R.L. Linn (ed.), *Educational Measurement*, 3<sup>rd</sup> edn. (pp. 13-103). NY: Macmillan.
- Milanović 2010: M. Milanović, Test Task Characteristics in Computer-assisted (Internet-Based) English Language tests of Reading and Listening: TOEFL i-BT, (unpublished master's thesis)
- North 1998: B. North, G. Schneider, Scaling descriptors for language proficiency scales. *Language Testing*, 15 (2), 217-262.
- Powers 2010: E. D. Powers, The case for a comprehensive, four-skills assessment of English-language proficiency *R&D Connections* 14. Princeton, NJ: Educational Testing Service.

Wall, Horak 2006: D. Wall, T. Horak, The impact of changes in the TOEFL examination on teaching and learning in central and eastern Europe. Phase 1: The baseline study. TOEFL Monograph No. MS-34. Princeton, NJ: Educational Testing Service.

Wall, Horak 2008: D. Wall, T. Horak, The impact of changes in the TOEFL examination on teaching and learning in central and eastern Europe. Phase 2: Coping with change. TOEFL Monograph No. iBT-05. Princeton, NJ: Educational Testing Service.

<http://www.toefl.org>. Accessed on 13 January, 2011.

**Милан Милановић, Ана Милановић**

## **TESTING LISTENING IN THE INTERNET-BASED TEST OF ENGLISH AS A FOREIGN LANGUAGE**

Резиме

У овом раду анализира се део међународног стандардизованог теста енглеског језика „TOEFL iBT“, у коме се тестира разумевање говора. На основу теоријских настојања да се дефинисе конструкт разумевања одслушаног говора и на основу публикација у издању организације која је аутор теста, овај рад ће покушати да одреди и дефинише конструкт, односно вештине неопходне да би се са разумевањем приступило анализи одслушаног исказа у условима језичког теста. TOEFL iBT мери способност кандидата да разумеју говорни енглески језик у академској средини, како би се проверила њихова способност да са разумевањем прате предавања и разговоре, па се у овом раду идентификују околности у којима је слушање неопходна вештина у академском окружењу: слушање за основно разумевање, слушање из прагматичких разлога, и слушање како би се повезали и интегрисали подаци из више извора. На основу околности у којима се одвија слушање у академском окружењу, ближе се одређују и вештине које су неопходне за идентификовање конструкта који се путем различитих задатака операционализује у језичком тесту. Следећи корак је операционализација теста путем одабира аудио материјала који ће послужити као инпут, и формулисање питања, која у случају овог теста подразумевају одабир понуђених одговора. На крају, идентификују се одређени недостаци који су својствени тестирању вештина, као и тестирању путем рачунара, па се констатује да, упркос недостацима, тестирање разумевања говора представља важан сегмент језичког теста, пошто је разумевање говора у циљном језику једна од кључних комуникативних вештина. Шири значај овог рада огледа се у увиду који састављачи језичких тестова у Србији добијају а који им може помоћи у састављању тестова који мере разумевање слушања у оквиру пријемних испита и других језичких тестова.

*Кључне речи:* конструкт, језичке вештине, разумевање слушања, формат испитних питања, TOEFL iBT