Milan D. Milanović*
American University
of the Middle East
Kuwait

371.3::811.111
004.9:81'322
https://doi.org/10.18485/zivjez.2019.39.1.1
Научни рад

OPERATIONALIZATION OF THE SPEAKING CONSTRUCT IN THE INTERNET-BASED TEST OF ENGLISH AS A FOREIGN LANGUAGE

In recent years, tests of speaking are integral parts to standardized language assessments because test users expect candidates to be able to communicate effectively in a foreign language, for a variety of different purposes and in variety of different modes (Powers 2010:1). The Internet-based Test of English as a Foreign Language is designed so as to address a number of language purposes using all four language skills, independently or in integration. In this paper we use a modified version of Bachman and Palmer's test task characteristic framework (1996) to analyze speaking tasks in the Speaking Section of TOEFL iBT, identifying speaking purposes the test addresses through the process of construct operationalization. The findings help us reflect on the authenticity of test tasks, as well as on the correspondence between test tasks and speaking tasks in target language use domains. In addition, task analysis reveals the test's limitations concerning situational/interactional authenticity due to non-live mode of test delivery.

Key words: speaking task, test task, TOEFL iBT, speaking construct, speaking assessment, authenticity

1 Introduction

Assessing speaking skills in large-scale high stakes administrations calls for standardized testing and rating procedures, securing objective and valid test scores based on which test users make

* milan.milanovic@aum.edu.kw

inferences about test takers' knowledge of a foreign language and their ability to speak the language in the real world. Considering the fact that standardized language tests target diverse test taking audience, they are seldom based on a syllabus, as is the case with classroom assessments. Instead, standardized language tests make use of a model of language ability in order to assess individual or integrated language skills, in line with the purpose of assessment and the construct, or the ability being measured.

This paper aims at investigating the construct of speaking in a Speaking Section of a standardized English language test, known as the Test of English as a Foreign Language (TOEFL), which is delivered via the Internet. The test is widely used to make decisions regarding placements in North American and universities worldwide, which implies that this is a high stakes language test, the results of which are used for making important decisions regarding test takers. Consequently, validation projects regarding this test and its sections have been carried out for over a decade in order to ascertain that inferences based on test scores are valid, as well as decisions based on the inferences.

We will make use of the test task characteristics framework, originally developed by Bachman and Palmer (1996), and then modified by other researchers (Douglas 2000; Luoma 2004; and Chapelle and Douglas 2006), in order to study the case of the Speaking Section of the Internet-based TOEFL in order to investigate speaking test tasks and the input used in the tasks with the purpose of identifying the construct of speaking and its operationalization in this test. In the first part of the paper we will identify some important considerations in the process of developing speaking tasks, followed by the introduction of a modified test task characteristics framework which will be used to analyze test tasks in the case study.

The findings will help us identify strengths and weaknesses of the test, particularly in terms of authenticity of test tasks and their resemblance to the real world speaking tasks. These can help instructors and test developers involved in oral assessments, as well as those involved in computer-facilitated language assessments, with identifying various speaking purposes and developing test

tasks to address these purposes. Another point of interest to test developers and classroom test instructors may be the test task characteristics framework we use in this paper to analyze characteristics of test tasks in order to ascertain that they correspond to the real world speaking tasks.

2 Speaking tasks

The purpose of assessment and intended use of the language assessed are key factors to developing test tasks. In the context of speaking assessment, tasks will involve activities taken by speakers who use the language "for the purpose of achieving a particular goal or objective in a particular speaking situation" (Bachman and Palmer 1996 in Luoma 2004: 31). With these considerations, test developers design tasks which, optimally, address the ability being assessed and "cover" the construct as completely as possible in order to provide solid foundations for making inferences on a candidate's ability to use the language in target language use domains. Target language use domains are of particular importance in communicative language testing, as they cover a multitude of situations where candidates are supposed to demonstrate their proficiency in a foreign language. In learning-based assessment, however, tasks are based on syllabi allowing test results to show the progress of students. Whatever the context of speaking assessment, Luoma argues that test designers need to create instructions both to test takers and to interlocutors/examiners, the tasks themselves, including the materials used throughout the test administration, for example pictures, graphs, role play cards, and other types of stimulus materials (Luoma 2004: 29).

Designing test tasks includes decisions regarding a number of key considerations: stand-alone or integrated testing, testing micro and/or macro skills, construct-based or task-based approach, live or tape-based (or recorded) test mode, question format, stimulus materials, etc. (for more on considerations in designing test tasks, see Bachman and Palmer 1996; Luoma 2004, and Brown and

Abeywickrama 2010). Building on the previous research in test task characteristics, more specifically on the work done by Bachman and Palmer, and modifications suggested by Luoma referring to speaking assessment, in this paper we will offer a modified test task characteristics framework in studying the case of the Internet-based Test of English as a Foreign Language, in order to investigate the operationalization of the construct of speaking.

2.1 Test task characteristics framework

The primary purpose of language tests is to make inferences that generalize to those specific domains in which test takers are likely to need to use the target language. In other words, we want to make inferences about test takers' ability to use language in a target language use (hereinafter TLU) domain (for more on target language use domains see Bachman and Palmer 1996:44). For this reason, care should be taken for test tasks to resemble the target language use tasks. Moreover, test tasks should be designed in such a manner that the interaction between the test taker and the task is similar to the interaction between the language user and the target language use situation task (Buck 2001: 108). The way we develop test tasks, i.e. their format, contents, nature, the media in which we present them, the equipment and facilities we use to administer the test, may all influence test takers' performance on the test, and consequently the inferences made on the basis of that performance. This influence is also known as "test method effect" and is documented in the research on second language testing (e.g. Bachman 1990), because it can affect validity of test scores and, consequently, the validity of inferences based on the scores. For this reason it is necessary to take precautionary measures when developing test tasks by carefully analyzing their characteristics.

Task characteristics are worth considering for several reasons. First, they provide us with an insight of what constitutes language tasks and what it is that constitutes test tasks, what links there may exist between these two groups of tasks, enabling us (as test devel-

opers) to develop test tasks which correspond to (target) language tasks. Second, test task characteristics will help determine the extent and ways in which a test taker's language ability is engaged. Third, the degree to which test task characteristics correspond to particular target language use task will determine the authenticity of test task as well as the validity of inferences made on the basis of test performance (Bachman and Palmer 1996:43). The framework of test task characteristics that will be used in its modified form (described below) builds on those originally proposed by Bachman (1990), Bachman and Palmer (1996), and on those developed by modifying the two frameworks in the works of Douglas (2000), and Chapelle and Douglas (2006), as well as on the work of Luoma (2004).

Bachman and Palmer developed a framework of language test tasks, stating that the purpose of their framework was to provide a basis for test development and use. They use the term 'task' to refer to both TLU tasks and test tasks, because they found that the characteristics described in their framework apply to both TLU tasks and language test tasks. There are five aspects of tasks which they set out to describe using the framework: setting, test rubric, input, expected response, and relationship between input and response. The characteristics of test tasks in the framework proposed by Douglas include: rubric, input, the interaction between input and response, and assessment criteria. His main intention was to outline a framework of task characteristics in language use situations and language for specific purposes tests that will allow test developers to analyze TLU situation and to develop test tasks which will reflect the characteristics of the target situation. Comparing his framework to that of Bachman and Palmer's, one can notice that he made some changes in his framework. Among other modifications that Douglas made in his framework, there is one which I find worth mentioning here - the introduction of **construct definition** within the characteristics of assessment (Douglas 2000). The framework developed by Chapelle and Douglas features characteristics of input and expected response under the same set of characteristics, and since the input and expected response feature almost the same set of characteristics they will be discussed as one and the same set.

Finally, in the work of Luoma, the importance is given to instructions both to test takers and to examiners, the tasks themselves, including the considerations of stimulus materials to which test takers respond in generating speech (2004: 29).

The framework of test task characteristics to be used in this paper is based on the considerations recommended by authors listed above, and as such, it features **construct** definition, as recommended by Douglas (2000), as well as a number of characteristics featured in Bachman and Palmer's characteristics of **test rubric**, combining them with characteristics of **the input and expected response**, or more specifically, with those pertaining to the format of the input and expected response: **instructions** to test takers (including their language and channel characteristics, specifications of procedures and tasks), **structure and time allotment** (including number of tasks, sequence of tasks, time allotted to tasks), **evaluation criteria** (including explicitness of criteria and procedures for scoring), **format of the input and expected response** (including channel, form, language, length, degree of speededness, and vehicle). Bachman and Palmer's framework includes other test task characteristics pertaining to test rubrics (i.e. availability of preparation and practice materials) and to the characteristics of the input and expected response (i.e. language characteristics), and the one developed by Milanović (2010) includes the characteristics of the toolbar in computer-based assessments, but these will not be discussed in this paper. Using this, somewhat adapted test task characteristics framework (see Table 1), we will analyse speaking test tasks in the Internet-based TOEFL in order to ascertain how the construct of speaking is operationalized.

Table 1. Test task characteristic framework (adapted from Bachman and Palmer 1996; Douglas 2000; Luoma, 2004; and Chapelle and Douglas 2006).

Test task characteristics framework
Construct definition
Characteristics of test rubric
Instructions
Language
Channel (aural, visual)
Structure and time allotment
Number of tasks
Sequence of tasks
Time allotted to tasks
Evaluation criteria
Explicitness of criteria and procedures for scoring
Characteristics of the input and expected response
Format
Channel
Form
Language
Length
Degree of speededness
Vehicle

2.1.1 On defining a construct of speaking

Defining a construct of speaking in a particular test will be closely related to the purpose of assessment and the intended use of test results based on which inferences will be made about a test taker's ability to use a given language in the real world, i.e. in TLU domains. There are several possibilities for defining constructs: a

construct may be based on a syllabus, which is a way to go in educational settings where there are instructional objectives to be met; the alternative way is to base a construct definition on a theoretical model of language ability (Bachman and Palmer 1996), or in the case of speaking assessment on a theoretical definition of a speaking ability. Another perspective is offered by Buck (2001) who suggests taking either a competence-based approach, or a 'task-based' approach, with the possibility of combining these two when necessary (for more on approaches to defining constructs see Bachman 1990; Bachman and Palmer 1996; and Buck 2001).

Proficiency tests, such as the Internet-based TOEFL, are theory-based rather than syllabus-based, and in this paper, we will rely on the publicly available information in order to define the construct of speaking in this test (for more on defining constructs see ETS 2007; Milanović 2010; Milanović 2011a; Milanović 2011b, and Milanović 2019). The construct is operationalized through test tasks and stimulus materials or the input used to elicit responses, and these have to reflect the ability measured rather than anything else to ensure construct validity of an assessment. In other words, if a test measures other than the ability defined in the construct definition, the validity of results is compromised due to the "construct-irrelevant variance". Similarly, if the selection of test tasks does not cover all of the ability defined as the construct in a test, the validity of results is equally at danger, because of the "construct-underrepresentation". Test bias, is another threat to validity of results, because test tasks may (un)intentionally favour certain groups of test takers, while at the same time they put others at disadvantage (for more on threats to test validity see Bachman 1990; Bachman and Palmer 1996; Elder 1998; Chapelle 2001; Chapelle and Douglas 2006, Milanović and Milanović 2013). Luoma argues that test developers in speaking assessments, have to be aware of the ability they want to assess in a particular context in order to be able to develop tasks and scoring criteria that match the construct definition (Luoma 2004: 28).

2.1.2 Characteristics of test rubric

2.1.2.1 Instructions

As Bachman and Palmer point out, instructions should be explicit, because test users will make inferences based on test performance, and if instructions are obscure and inadequate, one may expect them to possibly affect test takers' performance (1996:190). This characteristic can be described in terms of language in which the instructions are presented, the channel of presentation, and the specification of procedures and tasks.

2.1.2.1.1 Language

Instructions can be delivered in native or target language, or a combination of both may be used. In standardized high-stakes test administrations, with test takers taking the same kind of test throughout the world, the instructions are given in target language.

2.1.2.1.2 Channel

Instructions may be presented on the computer screen, i.e. presented in the visual channel, or they can, additionally be recorded and played into the headphones. The latter is the case when speaking is assessed via the Internet or in any other computer-assisted testing situation. Test takers are prompted to put on headphones, adjust the volume before they proceed to tasks. They are, also, asked to try out the microphone before the test starts, in order to ascertain that equipment is functioning well. Alternatively, in oral assessments, however, there is no equipment and/or technology involved in carrying out speaking tasks. Instead, the examiner presents a test taker with instructions and materials, guiding conversation as envisaged by test task developers. Buck (2001:119) warns that attention must be paid that the language of instructions is easier than the level of the language in stimulus materials, because any misunderstanding related to instructions may lead to construct-irrelevant variance.

2.1.2.2 Specification of procedures and tasks

Specification of procedures and tasks is another important aspect of test development, particularly in computer-assisted language tests, when test takers are facilitated to use computer equipment in order to respond to test tasks. Sometimes tutorials are provided for those test takers who are either not proficient users of computers or who do not have experience in responding to tasks which require relatively complex equipment manipulation. This issue is increasingly addressed by provision of preparation and practice materials, offered either commercially or free of charge (and often available online). The specification may be short or long, and it may refer either to the whole section (in tests where different language skills are assessed) or it can be intended to specifically describe particular tasks or items. In oral assessments, though, the specification of procedures and tasks is provided by the oral examiner.

2.1.2.2.1 Test structure and time allotment

Structure covers the practical considerations of putting a test together including the number of parts and their sequence, whereas time allotment here refers to the amount of time designated for a candidate to produce speech. I suggest that test task structure and time allotment in computer-assisted language tests be considered together as time is often shown on the toolbar throughout the test (with the possibility offered to test takers to hide the clock, should they wish to do so). In computer- and the Internet-based speaking tests, test takers can easily be informed of the number of sections/tasks and their respective salience. This information usually takes the form of the number of the current speaking task in relation to the overall number of tasks per speaking section/test, shown on the screen. In oral assessments, however, it is often duty of an examiner to inform a candidate of these matters.

2.1.2.3 Evaluation criteria

This set of characteristics is derived following the approach suggested by Douglas, although there are other approaches as well

(see Bachman and Palmer 1996; Alderson 2000; Buck 2001). Douglas distinguishes evaluation criteria from assessment criteria, the former being related to the extent to which test takers are informed about the nature of the criteria used to score their responses, while the assessment criteria refer to the same set of criteria and procedures described in more technical terms. These two sets of characteristics target different audience, the evaluation criteria are aimed at familiarizing test takers with what will constitute an acceptable response, whereas the assessment criteria are the tool used by test developers and test raters. Douglas suggests that this set of characteristics should include the explicitness of criteria, and procedures for scoring, but in the framework used here these will be discussed as the same set, because they often appear together in test rubric available to test takers in test preparation materials (ETS 2010).

2.1.2.3.1 Explicitness of criteria and procedures for scoring

Assigning scores to test takers' responses is based on the assumption that certain responses are correct, while others are incorrect, and that they can be scored as such. Making these explicit is important as it helps test takers in allocating time and applying different test-taking strategies. In responding to multiple-choice questions, for example, test takers are usually told to select "the correct answer". Consequently, this implies that there is only one correct answer, whereas all the other answers provided are incorrect. In some other tasks, test takers may be told to sequence sentences in a summary, implying that there will be only one correct sequence, etc. With respect to scoring criteria, test takers may be instructed to use the information provided in the reading/listening passage before they proceed to their speaking task, where their background knowledge is of no relevance to providing the correct response.

The extent to which the scoring criteria are made explicit to test takers is vital to test takers' awareness of what constitutes a sufficient response (Buck 2001:122). When a prompt elicits an open-ended response, test takers should know how much as well as what is considered adequate.

2.1.3 Characteristics of the input and expected response

The input and expected response can be described by a very similar set of characteristics, and for this reason their characteristics will be discussed together. Namely, both the input and the expected response feature the characteristics related to their **format** and **language** (the characteristics of language, however, will not be discussed in this paper due to its space constraints). The only difference between the input and expected response, with regards to format, relates to the type of input/expected response, respectively. When it comes to the type of input, or the material which test takers are expected to process and respond to, Bachman and Palmer make distinction between the item and the prompt (1996: 48). The former is used to elicit short answers, referred to as **limited production**, whereas the latter requires a longer response, referred to as **extended production**. The characteristics pertaining to the type of expected response are divided into three possible categories: selected, limited production and extended production. In standardized computer-delivered tests of speaking, however, limited and extended production is more likely to be encountered.

2.1.3.1 Format of the input and expected response

In both computer-delivered tests of speaking and in live oral assessments, test tasks and the input can presented in visual or aural **channel** in the **form** of language samples, often accompanied by non-language content or visuals, such as illustrations, graphs, tables, diagrams, etc. Depending on the purpose, textual content may be accompanied by recorded sound or multimedia. The **language** of the text and tasks will play a crucial role in tests of speaking, and, as is the case with other test types, the language of the text, listening passages, and other forms of input and tasks may be native, target, or both, depending on the test purpose and the task type. Some computer-delivered tests provide a language bar, where there is a menu out of which a desired language can be selected, so that instructions/prompts, tasks/items can be presented in it. However,

when it comes to standardized language tests, the input material as well as the accompanying rubric is presented in target language, adding to overall test validity. The **length** of input depends on the test developer's intentions, the construct being measured and the test type. In language testing, it ranges from a single word to a couple of pages, and depending on the length test takers take different time to interpret the reading material, listening passages, and other types of input. The length of expected response in speaking assessments will be determined by the test developer's intention, type of task and, consequently type of input. Most often, tests of speaking require a limited or extended production response. Another characteristic under the input format which is worth considering in development of a speaking test refers to **speededness**, or the rate at which the test taker has to process the text they are reading/listening to before they proceed to producing speech. Time allotment, obviously, overlaps with this characteristic as it refers to the amount of time allotted for an item or a section within a test. Although **vehicle** of presentation ("live" and/or "reproduced") seems to be more relevant to computer-based tests of speaking, it is desirable that test designers consider presenting texts as contextualized as possible.

3 The case study of the Internet-based Test of English as a Foreign Language

3.1 Defining the construct of speaking in the Internet-based TOEFL

In the Speaking section of the Internet-based TOEFL, productive language skills are assessed using **independent** and **integrated** speaking tasks. Reading extracts and listening passages are utilized as stimulus (source) materials, allowing test takers to demonstrate their ability to synthesize information from various sources in order to produce spoken responses. When rated, these responses form a basis for making inferences about their language

abilities, and for this reason they may play an important role in test takers' life. This requires that speaking construct is adequately defined and operationalized through stimulus materials/input and test tasks.

The Speaking section measures test takers' ability to speak successfully in and outside the classroom. As is the case with other sections of the test, the Speaking section addresses a number of purposes for speaking in target language situations. Since the target language use domain is identified as a university setting, it seems natural to expect students to speak in order to respond to questions, participate in academic discussions with other students, summarize and synthesize information from various sources, express their views on topics under discussion. These usually refer to classroom activities, but outside the classroom, students must participate in casual conversations, express their opinions, and communicate with people for different reasons and with different purposes (ETS 2010:16). Knowing what speaking purposes are addressed and what skills are required to address the purposes provides us with an insight into the construct definition. The construct is operationalized by test tasks and input (stimulus) materials used in these tasks, and for this reason this paper is aimed at investigating test tasks in the Speaking section of the test.

The Speaking section, which lasts about 20 minutes, contains six tasks, out of which the first two tasks are **independent speaking tasks**, and the remaining four are **integrated speaking tasks**. The former require that test takers speak about some familiar topic, whereas the latter call for speaking based on something they have read and/or heard, implying that they should be able to demonstrate their ability of integrating their language skills in English.

3.2 Independent Speaking Tasks

3.2.1 Construct definition

There are two independent speaking tasks in the Internet-based TOEFL. In the first independent task, test takers are ex-

pected to demonstrate the ability to talk about a **personal preference** when they are given a choice to talk about certain categories – for example, people they find important, events and activities, etc. (see Example 1).

Example 1: Personal preference

Describe an ideal marriage partner. What qualities do you think are most important for a husband or wife? Use specific reasons and details to explain your choices. (Barron's 2006)

In the second task (see Example 2), the idea of a personal **choice** is further developed, because test takers have to make a choice and defend it while choosing between two contrasting courses of action or behaviours.

Example 2: Making a choice

Some people like to watch the news on television. Other people prefer to read the news in a newspaper. Still, others use their computer to get the news. How do you prefer to be informed about the news and why? Use specific reasons and examples to support your choice. (Barron's 2006)

3.2.2 Characteristics of test rubric

The instructions and the questions for both tasks appear on the screen and are read aloud by the narrator in the target language, i.e. in English. Both visual and aural channels are used to provide the instructions as well as the tasks themselves. Given that there are two tasks, they follow one another, and test takers are not at liberty at choosing what task they want to attempt first, because there is a fixed ordering of test tasks. The test task in which test takers talk about a familiar topic comes first (Example 1), and is followed by the task in which they are supposed to express their opinion about a familiar topic (Example 2). Test takers are given 15 seconds to prepare for the task, and 45 second to respond by recording their

answers. The remaining time is shown on the screen, both for preparing and recording the answers.

The evaluation criteria are available in the test rubric as well as in the preparation materials. A test taker is told by the narrator that a satisfactory response will take all of the time provided for recording the response to the test question. Additionally, preparation materials provide complete analytical rubrics, including the scores and description of the performance leading to a particular score (ETS 2006: 44). It becomes evident that test takers' performance is evaluated against the following criteria: general description, delivery, language use, and topic development. In the actual testing situation, test takers are not made aware of all scoring criteria mentioned above, but the whole set of criteria is publicly available beforehand.

3.2.3 Characteristics of the input and expected response

The input in the independent speaking tasks includes a prompt provided in both visual and aural channel, and test takers are given 15 seconds to prepare, and 45 seconds to record their answers to the two respective prompts. The text of the prompt is provided in the form of the target language, and the expected responses should be recorded in the same form and language, i.e. in English. When it comes to the length of the input, it varies accordingly to the length of each question contained within the prompt. The length of the expected response is anything up to 45 seconds, leading to the conclusion that the type of expected response will be classified as extended production. The whole test is speeded, as the time for processing input and recording response is limited. As to the vehicle of presentation, the input is reproduced, because it is provided through computer technology, and the same applies to the expected response which is recorded and sent through the system to test raters. It should be noted that the characteristics of speedednes and vehicle of presentation will prove to be the same throughout the test.

3.3 Integrated Speaking Tasks

When it comes to integrated Speaking tasks, there are four of them. They will be considered and analyzed independently.

3.3.1 Integrated speaking task 1: Campus situation topic: FIT and EXPLAIN

3.3.1.1 Construct definition

The first one refers to a Campus situation topic, asking test takers to read a passage and then to listen to the passage providing comments on the issue in the reading passage. A test taker is then required to **summarize** the speaker's opinion within the context of the reading passage, or in other words to show how the speaker's opinion fits within the discussion contained within the reading passage.

Example 3: Campus situation topic: FIT and EXPLAIN
Reading passage:
Notice concerning proposed changes in language requirements
All international students at Community College are currently required to submit a TOEFL score of 80 in order to be admitted to credit classes. Students who score lower than 80 are referred to the English Language Institute for additional language instruction. The college is considering a proposal that would allow students with a score of 75 to take at least one credit class while they continue to study part time in the English Language Institute. The students would be assigned to an academic advisor who would help them select an appropriate course. This proposal will be discussed at a public meeting in the student union at 7 P.M. on December 1.

Listening passage:

A student expresses her opinion of the policy:
"I think this is a good policy because sometimes international students don't do well on an English test but they're good students in their major field. So, they can probably succeed with lower score on the TOEFL. And besides that, um, it's a good idea to begin with one

credit class instead of a full course load. I mean, studying part time while finishing the language classes will provide more of a ...transition into regular courses. Also, after several semesters in the English Language institute, most international students become impatient... and this opportunity ...to begin regular classes...it would increase their motivation toward the end of their language program."

Question:

The student expresses her opinion of the policy for international students. Report her opinion and explain the reasons that she gives for having that opinion. (Barron's 2006)

3.3.1.2 Characteristics of test rubric

The instructions are provided in target language and in both visual and aural channel. Test takers are familiarized with the number and sequence of tasks at the beginning of the Speaking section of the test, and specific objectives of each respective tasks are repeated to the test taker before the prompt has been displayed. The reading passage is 75-100 words long while the listening passage contains between 150 and 180 words and is 60 to 80 seconds long. A test taker is given 30 seconds to prepare his or her summary, after which time there is 60 seconds for recording their response.

Test takers are familiarized with the requirements of summarizing and rephrasing, and are warned against copying the words and phrases they hear. Although note taking is allowed, test takers are not supposed to read the materials they have prepared, because raters are trained to recognize a statement that is read.

3.3.1.3 Characteristics of the input and expected response

The input is provided in both visual and aural channels, comprising a text, displayed on the computer screen, followed by a recording of a speech accompanied by an appropriate picture of a student discussing a topic with her fellow-students. Both language

samples are in target language, the length of which is already discussed in the Characteristics of test rubric. When it comes to the degree of speededness, the task is speeded, as it allows for limited processing time and recording the response. When it comes to the vehicle of the presentation, the task is reproduced.

Test takers are expected to respond in English, and the type of their response will be classified as extended language production. The expected response, on the other hand, may take up the whole minute or less, although test takers are advised by the narrator reading the instructions to use up all of the time provided for recording the response. The expected response is speeded as well, given the fact that only one minute is provided for recording the response, which will be forwarded to test raters as a reproduced language sample.

3.3.2 Integrated speaking task 2: Academic course topic: GENERAL vs. SPECIFIC

3.3.2.1 Construct definition

Integrated task number 2 is supposed to cover an academic course topic contained within a reading passage, which is between 75 and 100 words long. This passage defines a term, a process, or idea from an academic subject. The listening passage, lasting 60 to 90 seconds, provides an excerpt from the lecture with examples and specific information aimed at illustrating the term, process, or idea from the reading passage.

A test taker is asked to **combine and convey** important information from the reading passage and the lecture excerpt, which means that it is important for a test taker to understand that the listening passage is there to provide some specific information which is related to the more general approach as taken within the reading passage. Finally, test takers use pieces of information they hear or read in order to **synthesize** them, demonstrating their ability to distinguish between general and specific information. There is a 90 second time allocated to this task, out of which 30

seconds go towards the preparation of response, while 60 seconds is all the time provided for recording a response.

Example 4: Academic course topic: GENERAL vs. SPECIFIC

Reading passage:
Although the first inhabitants of Australia have been identified by physical characteristics, culture, language, and locale, none of these attributes truly establishes a person as a member of the Aboriginal People. Because the Aboriginal groups settled in various geographical areas and developed customs and lifestyles that reflected the resources available to them, there is a great diversity among these groups, including more than 200 linguistic varieties. Probably the most striking comparison is that of the Aboriginal People who inhabit the desert terrain of the Australian Outback with those who live along the coast. Clearly, their societies have developed very different cultures. According to the Department of Education, the best way to establish identity as a member of the Aboriginal People is to be identified and accepted as such by the Aboriginal community.

Lecture (listening passage):
According to your textbook, the aboriginal people are very diverse, and I would agree with that; however, there are certain beliefs that unite the groups, and in fact, allow them to identify themselves and others as members of diverse Aboriginal societies.
For one thing, unlike the anthropologists who believe that tribes arrived in eastern Australia from Tasmania about 40,000 years ago, the Aboriginal People believe that they have always been in Australia, and that they have sprung from the land. Evidence for this resides in oral history that has been recorded in stories and passed down for at least fifty generations. This history is referred to as the "Dreaming". The stories teach moral and spiritual values and provide each member of the group with an identity that reflects the landscape where the person's mother first became aware of the unborn baby, or to put it in terms of the "Dreaming", where the spirit enters the mother's body.

So, I am saying that the way the Aboriginal People identify themselves and each other, even across groups, is by their membership in the oral history that they share.

Question:
Explain how the Aboriginal People are identified. Draw upon information in both the reading and the lecture. (Barron's 2006)

3.3.2.2 Characteristics of test rubric

The instructions are provided in English, in both visual and aural channels. The reading passage is between 75 and 100 words long, whereas the listening passage may take between 60 and 90 seconds. A test taker is given additional 90 seconds to prepare their response and record it.

As is the case in the previous task, a test taker is warned against mere copying the words from the prompt. Evaluation criteria stated in test preparation materials provide insight in what criteria for correctness are applied when assessing responses to integrated speaking tasks (ETS 2010:45), while some commercially available materials provide examples of satisfactory responses (see Example 5).

Example 5: A satisfactory response to the Integrated speaking task 2

The Aboriginal People are culturally and linguistically diverse, in part because the geography dictated both limitations and opportunities for their communities. So the establishment of identity as a member of the Aboriginal People because of their appearance, language, culture, or geographical location is considered not accurate. The department of Education suggests that the best means of identification is to be recognized and accepted by other members of the Aboriginal society. Um, according to the lecturer, even diverse groups have certain unifying beliefs that are passed down as oral tradition, called the "Dreaming". The stories associated with this tradition are used to teach ethical principles and spiritual lessons. It would probably be through knowledge of this shared oral history that Aborigines would identify each other. (Barron's 2006)

3.3.2.3 Characteristics of the input and expected response

The input to this task comprises prompts in both visual and aural channels, containing reading and listening passages in the English language. The length of the input and expected response, as mentioned above, amounts to about 5 minutes, indicating that the task is speeded, as well as the response to the prompt in the task. As can be seen in the Example 5, the expected response will take the type of extended language production, recorded in the form of a monologue. Both the input and expected response are delivered via the Internet, so the conclusion can be drawn that the language of the task is reproduced rather than live.

3.3.3 Integrated speaking task 3:
Campus Situation Topic: Problem/Solution

3.3.3.1 Construct definition

In this task, test takers are instructed to demonstrate their understanding of the problem and to **express an opinion** about solving the problem. Given that there is no reading input, test takers are provided with 20 seconds to prepare their answer, and then 60 seconds to record it.

Example 6: Campus situation topic: Problem/Solution

Listening passage:

Student: So, I really like my roommate, I mean, we get along great, and she's a good student. We take a lot of the same classes, and we study together, but...well, I have a problem, and I just don't seem to be able to deal with it.
Advisor: Okay. What's the problem?
Student: You see, she has a boyfriend in Florida, and she calls him on my cell phone.
Advisor: Then you have a bill for the extra minutes at the end of the month.
Student: Exactly, and I really can't afford it.

Advisor: Have you talked with her about it?
Student: Not really. I just don't know what to say.
Advisor: Well, I think you have a couple of options. You can tell her that you can't let her use your cell phone anymore, and you can offer to go with her to buy her own cell phone. That way she'll see that you're trying to help.
Student: That might work, but she'll probably still be upset...
Advisor: Or another possibility is...you could let her use your cell phone if she pays you for the extra minutes that you have to cover for her, and if she agrees to pay her share of all the bills from now on. Just tell her that you can't afford it.

Question:
Describe the woman's problem and the two suggestions that her advisor makes about how to handle it. What do you think the woman should do, and why? (Barron's 2006)

3.3.3.2 Characteristics of test rubric

Instructions are provided in target language in aural and visual channels. Test takers are familiarized with the task and advised to listen carefully in order to be able to respond to the question following the input. The time allocated to listening to the input is between 60 and 90 second, whereas the time for preparing and recording the answer takes up approximately 80 seconds.

Test takers are supposed to demonstrate their understanding of the problem presented in the input, consider possible solutions, and respond to the question using up most of the time allocated to recording the answer (60 seconds).

3.3.3.3 Characteristics of the input and expected response

The input in this task type is presented in aural channel, in target language. The recording is accompanied by visuals, i.e. photos, as can be seen in the Example 6, where the student and her advisor are pictured having the conversation the script of which is provided in this example. The aim of the visuals used in the task is to provide context, given that test takers are not able to

actually be present at the place and time of the conversation. The listening passage lasts between 60 and 90 seconds, and contains between 180 and 220 words. As all other tasks in this test, this task is speeded as well, and provided on the computer, implying that the vehicle of its presentation is reproduced.

The expected response takes the form of a recording of a speech in target language, lasting up to 60 seconds. As is the case in all other tasks, there is no direct interaction between a test taker and an examiner, in this case a rater, because the speech is recorded and submitted through the Internet, implying that the expected response is speeded and reproduced.

3.3.4 Integrated speaking task 4: Academic Course Topic: Summary

3.3.4.1 Construct definition

The test taker is asked to **summarize** the lecture, showing that he or she understands the relationship between the overall topic and examples provided as illustrations. Preparation time designated for this task is 20 seconds, while the response time is the same as in previous tasks, i.e. 60 seconds.

Example 7: Academic Course Topic: Summary

Professor:
Although cities have not generally been associated with wildlife, there are many species that have become so much a part of the urban landscape that they are, for the most part, unnoticed neighbours. For example, in New York's Central Park, almost 300 species of birds have been identified. Urban parks certainly provide some of the world's safest and in many ways, best wildlife habitats, and as the natural habitats shrink, well, these urban parks become more and more important to the conservation of the wildlife, including not only birds... but, uh ... but also freshwater animals, and, even small mammals. So, as you can see, man-made areas are one important type of habitat in cities. But artificial structures in the urban landscape...these can also provide a home for animals that adapt to the life in the city. For in-

stance, chimney swifts are birds that originally lived in hollow trees, but now chimney swifts are commonly found in the long brick chimneys in factories or other vertical shafts in tall buildings. Think about it. A city has more chimneys than there are hollow trees in a forest of equal area. Consequently, these birds flock to the city. Another case of adaptation is the urban drainage system, which is usually made up of concrete ditches, and they naturally attract stream and marsh animals. Again, to use New York City as an example, probably 250 species of fish are found in the harbour, many of which make their ways into pools and ponds and ditches in the New York drainage system. In Boston, the Back Bay was actually designed to create habitats and attract marshland wildlife to the city.

Question:
Using the main points and examples from the lecture, describe the two general types of habitats for wildlife found in urban areas. (Barron's 2006)

3.3.4.2 Characteristics of test rubric

Instructions to test takers are provided in English, both on the computer screen and in the earphones securing increasing the chances of test takers to understand them completely before proceeding to the task itself. In this task, there is no reading input, so test takers are required to listen to a recorded lecture before they proceed to summarize its content. The time allocated to the task is between 90 and 120 seconds.

Instructions make it explicit to test takers that they should summarize the contents of the listening passage without actually copying words and phrases they heard, implying that they should demonstrate their skills of paraphrasing and reformulating other person's utterance. The criteria against which the response is rated include: general description, delivery, language use, and topic development. It should be pointed out that the same set of rating criteria is applied in rating performance on all integrated speaking tasks (ETS 2010: 45).

3.3.4.3 Characteristics of the input and expected response

The input in this task consists of a listening passage, provided through the headset, with the help of visuals (photos of a lecturer, talking about a given topic). The excerpt from a lecture contains between 230 and 280 words, and it takes between 90 and 120 seconds to play it. In this excerpt a narrator explains a term or concept, using examples to illustrate the term or concept (ETS 2010: 18). The language of the lecture can be somewhat technical, but no background knowledge of the subject matter is required to respond to the task, because test takers are expected to demonstrate their ability of summarizing what they have heard. The talk itself is reproduced, and when it comes to degree of speedednes, it may be assumed that the task is speeded.

The expected input should consist of a recorded summary of the input, provided in target language. Test takers are allowed to take notes, but they are not supposed to read them, while recording their answer. The length of the response may vary, but the allocated time for this task is 60 seconds, and when that time expires, candidates are no longer able to make a recording. This also implies that the task itself is speeded, as well as that the expected response is reproduced from the point of view of test raters.

4 Discussion

4.1 Independent speaking tasks

The analysis of independent speaking tasks draws attention to the lack of interaction as is usual in the real world communication, but on the other hand, short monologues on a given topic are to be found in any kind of interpersonal interaction. This leads to the conclusion that though deprived of interactiveness, tasks requiring extended production, taking up to 45 seconds, are just like any other real world tasks where people express and defend their opinions about familiar matters, although in the form of a short mo-

nologue. With this respect, independent speaking tasks used in this test resemble real world speaking tasks requiring that speakers produce answers to questions similar to those used in Examples 1 and 2. For this reason, it may be assumed that independent speaking tasks possess a certain degree of authenticity.

This test is administered over the Internet, using computer technology. This implies that there is no live interaction between a speaker (test taker) and an examiner, leading to the conclusion that non-linguistic elements (mimics, gesticulation, eye contact, etc.), which are present in live interaction, are missing in assessing speech. Another limitation to the authenticity of test tasks in the Internet-based TOEFL refers to recording test takers' answers, which is seldom the case in TLU domains, where people are asked opinion about a familiar topic.

4.2 Integrated speaking tasks

The analysis of the Task One shows that this task represents one of those language situations where people respond to what they have read and heard, summarizing other people's ideas and opinions. No background knowledge is needed for attempting this task, as all the information necessary is provided in the prompt. This is important as it helps enhancing test fairness, although, on the other hand, in real world situations, people tend to know more about the topic they talk about. Lack of interaction is obvious, but the authenticity seems to be enhanced through contextualization by the means of visuals used in the task.

In Task Two, test takers need to demonstrate their ability to relate specific information they hear to a more general matter found in their reading. If we bear in mind that test scores are mainly used to make decisions about placement at North American universities, this task appears to be both relevant to the construct measured and authentic as the skills it is used to measure can be applied outside the testing situation, i.e. in the real world classrooms. The length of both the input and expected response may be found to be disputa-

ble, as in the real world, students process larger texts and materials heard in lectures. However, as testing is based on a sample, it may be assumed that a task like this can be relevant to TLU domains.

Task Three represents a situation which may be encountered in university campuses, where students are supposed to interact and communicate with regards to day-to-day situations and problems. No background knowledge is necessary for responding to the task, as all the material needed is provided in the input. However, test takers are asked to provide opinions and offer solutions to the problem, so they may rely on their own experience when offering solutions and advice. A language task like this can be encountered in everyday life in TLU domains. The problem with this task is the evident lack of interaction and "real" communication, because test taker is not really participating in the situation described in the task. On the other hand, skills demonstrated in this task may be useful in reporting problems and suggesting potential solutions to third parties, in this case to "invisible" examiners.

Making a summary is a necessary skill in the academic setting, both in writing and in speaking, and this justifies Task Four in the Internet-based TOEFL. However, in the real world classrooms, students are often able to ask for repetition and clarification, and this is not the case in this test. Also, the complexity and technicality of the subject matter may be a point of dispute when it comes to responding this task, although it appears that background knowledge is not necessary to successfully respond to it. It should be observed that students who are better at taking notes and instantaneously paraphrasing them are bound to be more successful in attempting the task. Hence, note-taking may be considered a part of the construct, as well as the ability to listen actively and memorize what is being said.

As mentioned earlier, the Internet-based TOEFL, regardless of the fact that it is a well-researched and validated standardized language test, is not resistant to limitations, some of which have already been mentioned. The following limitations, however, deserve to be emphasized here, as they refer to the input to the speaking tasks. 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). 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).

5 Conclusion

Test task characteristics applying to speaking tasks in the Speaking Section of the Internet-based TOEFL are analyzed in this paper, with the purpose of investigating the construct of speaking and its operationalization through tasks and materials utilized to elicit test takers' responses to test questions. The analysis has revealed that there are two types of speaking tasks used to address the construct of speaking and various speaking purposes encountered in a university setting – individual and integrated speaking tasks. These correspond to a number of speaking purposes in the domains of target language use, i.e. in university settings, where students are supposed to talk about familiar topics, give and support opinions, synthesize information from various sources, paraphrase and reformulate what they have heard or read, etc. The stimulus materials used as the input in the tasks are adapted to suit the testing purposes, but are, nevertheless, taken from authentic sources, increasing test task authenticity and correspondence of test tasks to the real

world speaking tasks. However, in some cases the language of the input is not entirely natural, computer delivery of the test reduces situational authenticity of the tasks, but validation studies make it clear that validity of test scores is not jeopardized by these and similar limitations (ETS 2008).

Speaking of limitations, the reader should be aware of some limitations to this case study. First, this case study features only one of a few popular standardized tests of English, and it would be interesting to see how the construct of speaking is operationalized in other large-scale tests of English, for example in IELTS and TOEIC. Second, the case study is carried out using publicly available information, while the fact remains that developers of standardized language tests retain certain pieces of information which are considered to be confidential. Third, not all facets of the test task characteristics framework have been analyzed (see more in Bachman and Palmer 1996), for example, language characteristics in both the input and expected response, mainly because the texts and recordings in the input are diverse and impossible to analyze in a single case study, and because the samples of expected responses are not available for analysis, etc.

The findings presented in this paper are intended to be of use to test developers and language testers interested in assessing speaking skills in both large-scale and small-scale administrations. The construct of speaking as it appears in this test can be defined and assessed in similar fashion in other testing circumstances depending on the purpose of assessment, and the test task characteristics framework could be of use in comparing test tasks to the tasks in target language use domains.

REFERENCES

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Barron's. (2006). TOEFL iBT.

Brown, D. H., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices* (2nd ed.). White Plains, NY: Pearson Education.

Buck, G. (2001). *Assessing Listening.* Cambridge: Cambridge University Press.

Chapelle, C. A. (2001). *Computer Applications in Second Language Acquisition.* Cambridge: Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing Language through Computer Technology.* Cambridge: Cambridge University Press.

Douglas, D. (2000). *Assessing Languages for Specific Purposes.* Cambridge: Cambridge University Press.

Elder, C. (1998). What counts as bias in language testing. *Melbourn Papers in Language Testing, 1* (7).

ETS. (2010). *TOEFL iBT Tips. How to prepare for the TOEFL iBT*. Retrieved August 10, 2010, from ets.org: http://www.ets.org./Media/Tests/TOEFL/pdf/TOEFL_Tips.pdf

ETS. (2008). *Validity Evidence Supporting the Interpretation and Use of TOEFL iBT Scores.* Retrieved December 14, 2013, from Educational Testing Service Web site: www.ets.org/s/toefl/.../toefl_ibt_insight_s1v4.pdf

Luoma, S. (2004). *Assessing Speaking.* Cambridge: Cambridge University Press.

Milanović, M. (2011). Characteristics of the Test Rubric in the Reading and Listening Sections of the TOEFL I-BT. *Nasledje , 17*, 267-282.

Milanović, M. (2010). *Test task characteristics in high-stakes Internet-delivered tests of English: Test task characteristics in the Listening and Reading sections of the TOEFL-iBT.* Neobjavljena master teza. Beograd: Filološki fakultet.

Milanović, M. (2011). The Construct of Reading and its Operationalization in the Internet-based Test of English as a Foreign Language. *Philologia , 9*, 73-82.

Milanović, M., & Milanović, A. (2013). Etički problemi u testiranju jezičkih kompetencija. *Nasleđe , X* (26), 69-87.

Milanović, M. (2019). *Investigating Authentic Forms of Assessment in Testing English for Specific Purpose Speaking Skills.* Unpublished doctoral dissertation. Belgrade.

Powers, D. E. (2010). The case for a comprehensive, four-skills assessment of English-language proficiency. *R&D Connections 14* .

Милан Милановић

ОПЕРАЦИОНАЛИЗОВАЊЕ ГОВОРНОГ КОНСТРУКТА У ТЕСТУ ЕНГЛЕСКОГ КАО СТРАНОГ ЈЕЗИКА КОЈИ СЕ СПРОВОДИ ПУТЕМ ИНТЕРНЕТА

Резиме

Током последње деценије, стандардизовани међународни тестови којима се проверава знање енглеског језика, као свој саставни део укључују и проверу вештине говора будући да крајњи корисници резултата тестирања очекују од кандидата да поседују способност комуникације на страном језику, из различитих разлога и путем различитих медија (Пауерс 2010:1). Тест енглеског као страног језика који се спроводи путем интернета (TOEFL iBT) састављен је тако да узима у обзир различите комуникативне циљеве који се остварују уз учешће све четири језичке вештине (читање, писање, говор и слушање), изоловано или у интеграцији једне са другима. У овом раду, аутор користи модификован Оквир карактеристика задатака (Бакман и Палмер 1996) како би идентификовао различите комуникативне циљеве, односно сврхе у које се користи вештина говора путем операционализације говорног конструкта у самом тесту. Путем анализе тестовних задатака аутор рада испитује њихову аутентичност, као и везу између тестовних задатака и говорних задатака у домену употребе циљног језика. Резултати анализе указују на то да овај тест садржи одређена ограничења у погледу интеракцијске/ситуационе аутентичности настала услед непостојања саговорника коме се кандидат усмено обраћа.

Кључне речи: говорни задатак, тестовни задатак, TOEFL iBT, говорни конструкт, тестирање говора, аутентичност

Данијела Врањеш[*]