Iranian Journal of Applied Linguistics (*IJAL*)

Vol. 21, No. 1, March 2018, 215-262

Authenticity Evaluation of TOEFL iBT Speaking Module from the Perspective of Applied Linguistics and General Education

Marzieh Souzandehfar*, *Jahrom University, Iran*

Abstract

For the first time, this study combined models and principles of authentic assessment from two parallel fields of applied linguistics as well as general education to investigate the authenticity of the TOEFL iBT speaking module. The study consisted of two major parts, namely task analysis and task survey. Utilizing Bachman and Palmer's (1996) definition of authenticity, the task analysis examined the degree of the correspondence between the characteristics of the speaking module tasks in the TOEFL iBT test and those of target language use (TLU) tasks. In the task survey, a Likert Scale questionnaire of authenticity was developed by the researcher based on Herrington and Herrington's (1998; 2006) four criteria of authentic assessment. The questionnaire was sent through email to 120 subjects who had already taken the test in order to elicit their attitudes towards the degree of the authenticity of the speaking section tasks. The results of the task analysis revealed a limited correspondence between the characteristics of the test tasks and those of the TLU tasks. However, the results of the task survey indicated that except for one factor (indicators), most of the test takers had a positive view toward the authenticity of the speaking module tasks in terms of the three other factors (context, student factor, task factor).

Keywords: Authentic assessment; Speaking module; TOEFL iBT

Article Information:

Received: 24 October 2017 Revised: 2 December 2017 Accepted: 10 February 2018

Corresponding author: Department of Translation Studies, Jahrom University, Fars, Iran Email address: souzandeh@jahromu.ac.ir

1. Introduction

1.1. Authenticity

It is perhaps no exaggration to say that authentic assessment is the most significant goal of language testing. Ingram (2003) claims that:

> The history of language testing (especially of attempts to measure practical language ability) is, to a large extent, the history of attempts to bridge the gap between tests and real-life language use…it is the history of progress towards more authenticity in language testing. (p. 4)

The notion of authenticity has always been open to debate within the fields of applied linguistics as well as general education. In applied linguistics, the idea emerged in the late 1970s when communicative methodology was gaining importance and there was a growing interest in teaching and testing 'real-life'. In general education, on the other hand, it took more than another decade before the notion was recognized. Since then, there has been much overlap in the definitions in both fields, yet the debates have remained largely independent of each other (Lewkowicz, 2000).

In applied linguistics, there are two major pathways to the discussion, which are confusingly mistaken for each other (Pinner, 2016). The first pathway to the discussions of authenticity in applied linguistics relates mainly to language learning materials, which also includes the tasks utilized to engage learners (Gilmore, 2009, 2011; Malone, 2017; Mishan, 2005; Morrow, 2018). In fact, these discussions which take a more practical view of the 'authenticity debate', argue that authentic materials should be "real

language produced by a real speaker or writer for a real audience and designed to convey a real message" (Morrow, 1977, p. 13). The second pathway is in particular concerned with the process of 'authentication' (Mishan, 2005; van Lier, 1996; Widdowson, 1978, 1994). Here, authenticity is not something absolute, but relative, and is concerned with a process of personal engagement with the language (van Lier, 1996). This is exactly in line with Widdowson's (1978) argument about the distinction between 'genuineness' and 'authenticity' of language. Widdowson (1978) claimed that "genuineness is a characteristic of the passage itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and has to do with appropriate response" (p. 80). In other words, as Pinner (2015) states, "simply taking a newspaper out of an English speaking context quite often means you leave the real reason for interacting with it behind, which seriously impairs its authenticity" (p. 2).

Hung and Victor Chen (2007, p. 149) have also heavily criticized what they call extrapolation techniques, i.e. the act of taking something out of one context and bringing it into another (the classroom) expecting its function and authenticity to remain the same.

In one of his most recent works, Pinner (2016) replaced the 'classic' definition of authenticity with a reconceptualized version, which, as he claims, is more inclusive to other varieties of English. He poses the 'paradox of authenticity' arguing that

> At one end it is too complicated to have a single definition, and at the other end practitioners talk about 'authentic' materials when they generally mean newspapers or other items that have simply been extrapolated from a target language speaking community. (P. 2)

Pinner (2016) believes that authenticity is not something absolute, but "rather relative to the learner and their unique and individual beliefs" (p. 1). He tries to discuss authenticity in light of emergent theories of language acquisition, such as chaos/complexity theory and dynamic systems approaches and consequently, introduces the Authenticity Continuum, which is a framework for treating authenticity as a socially mediated and contextually dependent dynamic process of investment.

Unfortunately, research into authenticity is rather scarce, but the situation is further exacerbated when it comes to authenticity in language testing and authentic assessment. Gilmore (2007) reviews over a century of literature on authenticity, providing a comprehensive and in-depth overview in which he identifies eight different and overlapping definitions, only two of these referring to authenticity in language testing, i.e. authenticity as it relates to assessment and the Target Language Use Domain (Bachman & Palmer, 1996).

With respect to authenticity in language testing, in the early 1990s, Bachman built on the ideas put forward by Widdowson (1978) and Breen (1985). He suggested that there was a need to distinguish between two types of authenticity: situational authenticity, i.e. the perceived match between the characteristics of test tasks to target language use (TLU) tasks, and interactional authenticity, i.e. the interaction between the test taker and the test task (Bachman, 1991). In so doing, he claimed that authenticity involved more than matching test tasks to TLU tasks. In fact, he saw authenticity also as a quality arising from the test takers' involvement in test tasks.

In 1996, Bachman and Palmer put a step forward and separated the notion of authenticity from that of interactiveness, defining authenticity as 'the degree of correspondence of the characteristics of a given language test task to the features of a TLU task' (Bachman and Palmer, 1996, p. 23). This definition corresponds to that of situational authenticity, while interactiveness replaced what was previously called interactional authenticity. The premise behind this change was the recognition that all real-life tasks are by definition situationally authentic, so authenticity can only be an attribute of other tasks, that is, those used for testing or teaching.

On the other hand, in the realm of general education, Herrington and Herrington (1998; 2006), the two leading scholars in the field of authentic assessment, developed the most canonical guidelines for defining authenticity in the field. They categorized their guidelines into four groups, that is, *context, student factors, task factors,* and *indicators*. The first criterion of authentic assessment requires fidelity of context to reflect the conditions under which the performance will occur (rather than contrived, artificial, or decontextualized conditions). Student factor or student's role requires students to be effective performers with acquired knowledge, and to craft polished performances or products. It also requires significant student time and effort in collaboration with others. With respect to authentic activity, or task factors, test items should involve complex, ill structured challenges that require judgment, and a full array of tasks. In addition, this criterion requires the assessment to be seamlessly integrated with the activity. The last factor, i.e. indicators, is concerned with multiple indicators of learning. It also requires achieving validity and reliability with appropriate criteria for scoring varied products.

Similarly, Herrington, Oliver, and Reeves (2002) developed 10 criteria of an authentic task in an online environment. Their work is known as Approach 2, and in many respects, it reflects the features identified by Herrington and Herrington (1998; 2006) with emphasis on relevance beyond the classroom to the real world, diversity of outcomes, complex tasks, and integration with assessment.

Approach 3 consists of a five dimensional framework designed by Gulikers, Bastiaensand Kirschner (2006). These dimensions have already been included by Approaches 1 and 2, and do not indicate any additional features of authentic assessment.

Approach 4 is based on the work of Frey and Schmidt (2007) that recognized the following features of authentic assessment: nature of the stimuli, complexity, conditions, resources, consequences, and whether tasks are determined by an assessor or student.

Another approach which indicates the features of authentic assessment has been adopted by Keyser and Howell (2008). Although they use some different terminology, their approach isolates the features highlighted in the earlier approaches.

The last approach is introduced by Burkill, Dunne, Filer, and Zandstra (2009). Approach 6 places emphasis on the product as well as the process, the development of real world and higher order cognitive skills (analysis, synthesis and evaluation), the integration of a range of skills into a whole project, and the construction of new ideas and responses. These features largely coincide with the features identified in the earlier approaches.

As the above overview suggests, there have been two parallel debates on authenticity which have remained largely ignorant of each other; one in the field of applied linguistics and the other in the realm of general education. Lewkowicz (2000) suggested that discussions within the fields of applied linguistics and general education need to come closer together in order to provide a more insightful understanding of the notion of authenticity and authentic assessment. Furthermore, he emphasized that such discussions need to be empirically based to inform what has been still a predominantly theoretical debate.

No study has so far combined the models and principles of authentic assessment from the field of applied linguistics with those of general education. For the first time, this study made the two parallel fields across each other to investigate authenticity in language testing. In applied linguistics, due to scarcity of research on authenticity in language testing, Bachman and Palmer's (1996) model of test usefulness could surprisingly be considered as the last development in this respect and consequently utilized for the purpose of the present study.

1.2. Research on the Authenticity of TOEFL iBT Speaking Module

Since the time when the Test of English as a Foreign Language (TOEFL) underwent major revisions, particularly the introduction of speaking as a mandatory section on the TOEFL Internet-based test (iBT), the problem of validity and authenticity of the test has been frequently discussed. In 2012, an announcement was made by TOEFL COE research program to address the topic of validation and more specifically the problem of candidates' performance on the TOEFL iBT test speaking and/or writing sections and its correspondence to their performance on real-life academic tasks; an issue

which is at the heart of authentic assessment. However, the attempts which have been made so far by the researchers in this respect do not seem to be satisfactory with respect to both the number of the studies and the specific topic of authenticity. In fact, most of the prominent studies that have been carried out on TOEFL iBT speaking section (Farnsworth, 2013; Sawaki, Stricker, & Oranje, 2009; Xi, 2008; Zahedi & Shamsaee, 2012) have specifically focused on the evaluation of the construct validity and predictive validity rather than the correspondence between the candidates' performance on the speaking section of the test and their performance on real-life academic tasks.

Only in a few cases (Meng-li, 2010; Ockey, Koyama, Setoguchi, & Sun, 2015), the authenticity of the speaking module has been investigated. In his study, Meng-li (2010) analyzed the authenticity of TOEFL iBT oral test, including the authenticity of text, setting and tasks, interaction between test takers and test tasks, and scoring criteria and process. Meng-li found the test authentic, but at the same concluded that the authenticity of the oral test depends on its definition and the interaction between test takers and test tasks. In a quantitative study, Ockey et al. (2015) made an attempt to determine the extent to which performance on the TOEFL iBT speaking section is associated with the other indicators of Japanese university students' abilities to communicate orally in an academic English environment and to determine which components of oral ability for these tasks are best assessed by TOEFL iBT. The results of the correlations revealed that TOEFL iBT speaking scores were good overall indicators of academic oral ability and that they were better measures of pronunciation, fluency, and vocabulary/grammar than they were of interactional competence, descriptive skill, and presentation delivery skill.

On the other hand, although the importance of research on stakeholder beliefs and attitudes about tests is widely recognized, according to Malone & Montee (2014), little research has examined student test-takers' perceptions of the items on the TOEFL iBT test. In their study, Malone & Montee (2014) explored stakeholders' beliefs (administrators, instructors, and students) about the TOEFL iBT test as a measure of academic language ability. The results indicated that students showed mixed attitudes considering the four skills and their nationality. For example, the German students were the only participants who agreed that the test questions felt natural. German students agreed with all items about the TOEFL iBT's ability to show how well they could perform in English, except on the speaking section. Students from all countries believed that the listening section showed how well they could listen in English. Students did not believe that the TOEFL iBT allowed them to show their ability of speaking English. On the other hand, Saudi and South Korean student responses indicated some disagreement with the TOEFL iBT's capacity to show their abilities in English.

Rosenfeld, Leung, and Oltman (2001) conducted comprehensive studies on listening, speaking, reading, and writing tasks which are necessary for academic success. Also, Stricker, Wilder, and Rock (2004) investigated students' attitudes towards computer-based TOEFL, and another group of research on students (Powers & O'Neill, 1993; Schmitt, Gilliland, Landis, & Devine, 1993; Schmidt, Urry, & Gugel, 1978) has focused on test takers' attitudes toward computer-based testing. However, although their research is illuminating, it does not particularly focus the content of the TOEFL iBT and in particular its speaking module.

As a result, this study built upon Herrington and Herrington's (1998, 2006) criteria of authentic assessment in general education to conduct a task survey, eliciting student test takers' attitude towards the authenticity of the speaking section of the TOEFL iBT test, which is one of the most widely accepted English language assessments around the world.

Furthermore, while most of the aforementioned studies have utilized either qualitative or quantitative research methods, this study, in addition to combining models from the two fields of applied linguistics and general education, utilized a mixed method to investigate the authenticity of TOEFL iBT speaking module through a task analysis and a task survey.

2. Theoretical framework

In 1996, Bachman and Palmer proposed a model of test usefulness that includes six test qualities – reliability, construct validity, authenticity, interactiveness, impact, and practicality. Unlike Bachman (1990, 1991) who distinguishes between two types of authenticity, situational authenticity (i.e., the perceived match between the characteristics of test tasks to target language use (TLU) tasks), and interactional authenticity (i.e., the interaction between the test taker and the test task), Bachman and Palmer (1996) put a step forward and separated the notion of authenticity from that of interactiveness, defining authenticity as 'the degree of correspondence of the characteristics of a given language test task to the features of a TLU task' (p. 23). This definition corresponds to that of situational authenticity, while interactiveness replaced what was previously termed interactional authenticity. To find the degree of correspondence between test and TLU tasks – that is, to determine the authenticity of test tasks – Bachman and Palmer proposed a framework of task characteristics. This framework

provides a systematic way of matching tasks in terms of their setting, the test rubrics, test input, the outcome the tasks are expected to give rise to, and the relationship between input and response (See Table 1).

Furthermore, for the second part of the study, i.e. the task survey, Herrington & Herrington's (1998, 2006) list of the essential characteristics of authentic assessment was drawn upon from the field of general education to develop a questionnaire of authenticity to elicit test takers' attitudes towards the authenticity of the speaking section tasks in the TOEFL iBT test. The list consists of four categories: *context, the student's role, authentic activity,* and *indicators*. Using these guidelines, assessment is most likely to be authentic if it satisfies the following criteria:

Context:

· Requires fidelity of context to reflect the conditions under which the performance will occur (rather than contrived, artificial, or decontextualized conditions) (Meyer, 1992; Reeves & Okey, 1996; Wiggins, 1993)

Student's role

· Requires students to be effective performers with acquired knowledge, and to craft polished performances or products (Wiggins, 1989,1990, 1993,)

· Requires significant student time and effort in collaboration with others (Kroll, Masingila, & Mau, 1992; Linn, Baker, & Dunbar, 1991)

Authentic activity

· Involves complex, ill structured challenges that require judgment, and a full array of tasks (Linn, et al., 1991; Torrance, 1995; Wiggins, 1990, 1993, 1989)

· Requires the assessment to be seamlessly integrated with the activity (Reeves & Okey, 1996; Young, 1995)

Indicators

· Provides multiple indicators of learning (Lajoie, 1991; Linn, et al., 1991)
· Achieves validity and reliability with appropriate criteria for scoring varied products (Lajoie, 1991; Resnick & Resnick, 1992; Wiggins, 1990).

3. Research Questions

In its announcement in 2012, TOEFL COE proposed a set of research topics, the first and most urgent of which was that of validation. The first problem being addressed under this topic was concerned with relating candidates' performance on the TOEFL iBT test speaking and/or writing sections to their performance on real-life academic tasks. Consequently, this study is considered as a response to this announcement with its focus on the authenticity of the tasks in the speaking section of the TOEFL iBT test.

Utilizing a mixed method, and drawing upon Bachman and Palmer's (1996) model of test usefulness from the field of applied linguistics and Herrington and Herrington's (1998; 2006) essential elements of authentic assessment from general education, this study was intended to investigate the authenticity of the speaking module tasks in the TOEFL iBT test in two ways, task analysis, and task survey. More specifically, it tries to answer the following questions:

- To what extent do the characteristics of the TOEFL iBT speaking section tasks correspond to those of TLU tasks?
- To what extent do test takers believe that the TOEFL iBT speaking section tasks are authentic?

Table 1
Task characteristics

Characteristics of the setting
Physical characteristics
Participants
Time of task

Characteristics of the test rubrics
Instructions
Language (native, target)
Channel
Specification of procedures and tasks
Structure
Number of parts/tasks
Salience of parts/tasks
Sequence of parts/tasks
Relative importance of parts/tasks
Number of tasks/items per part
Time allotment
Scoring method
Criteria for correctness
Procedures for scoring the response
Explicitness of criteria and procedures

Characteristics of the input
Format
Channel (aural, visual)
Form (language, non-language, both)
Language (native, target, both)
Length
Type (item, prompt)
Degree of speededness
Vehicle (live, reproduced, both)
Language of input
Language characteristics
Organizational characteristics
Grammatical (vocabulary, syntax, phonology, graphology)
Textual (cohesion, rhetorical/conversational organization)
Pragmatic characteristics
Functional (ideational, manipulative, heuristic, imaginative)
Sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language)
Topical characteristics

Characteristics of the expected response
Format
Channel (aural, visual)
Form (language, non-language, both)
Language (native, target, both)
Length
Type (item, prompt)
Degree of speededness
Vehicle (live, reproduced, both)

Language of expected response
Language characteristics
Organizational characteristics
Grammatical (vocabulary, syntax, phonology, graphology)
Textual (cohesion, rhetorical/conversational organization)
Pragmatic characteristics
Functional (ideational, manipulative, heuristic, imaginative)
Sociolinguistic (dialect/variety, register, naturalness, cultural
references and figurative language)
Topical characteristics

Relationship between input and response
Reactivity (reciprocal, non-reciprocal, adaptive)
Scope of relationship (broad, narrow)
Directness of relationship (direct, indirect)

4. Design of the Study

This study utilized a mixed method to investigate the authenticity of the speaking module tasks of the TOEFL iBT both qualitatively and quantitatively through a 'task analysis' and a 'task survey'. The details in each approach are described in the following sections.

4.1. Task Analysis

To find the degree of the correspondence between the characteristics of the TOEFL iBT speaking section tasks and those of the TLU tasks – that is, to determine the authenticity of the test tasks – Bachman and Palmer's (1996) framework of task characteristics was utilized. This framework provides a systematic way of matching tasks in terms of their setting, the test rubrics, test input, the expected response, and the relationship between input and response. Table 1 shows the complete list of the characteristics.

The test rubric may be a characteristic for which there is relatively little correspondence between language use tasks and test tasks. This is because "in language use this characteristic is generally implicit, while in a test task this needs to be made as explicit and clear as possible" (Bachman and Palmer, 1996, p.50). As a result, in task analyses, including the one in this study, test rubric is omitted from the list of task characteristics.

4.2. Task Survey

In the second part of the study, a task survey was conducted to elicit the attitudes of the test takers towards the degree of the authenticity of the

TOEFL iBT test speaking section tasks. The method in this part was as follows:

4.2.1. Participants

This part of the study consisted of two phases: 1) Validation, and 2) Application of the TOEFL iBT speaking section authenticity questionnaire. The validation phase included a pilot study at two stages, initial piloting and final piloting. At the initial stage, a pool of items consisting of 45 items, was given to two experts and one Ph.D. student for external feedback and revision. Then at the second stage of the validation, the Persian version of the resulting questionnaire, including 34 items, was sent to 247 Iranian subjects through email. These participants had already taken the TOEFL iBT test and were all familiar with the speaking section tasks.

In the second phase of the study, i.e. the application of the questionnaire, a sample of 120 participants, from the same group of subjects in the first phase, participated again in the research work.

4.2.2. Instrument

TOEFL iBT speaking section authenticity questionnaire was first constructed and then validated to be used as an instrument in this study and for conducting further research in the field of foreign language learning.

4.2.3. Procedures

Following Herrington and Herrington's (1998, 2006) essential elements of authentic assessment, the researcher constructed the related questionnaire

adopting a straightforward procedure including three steps: 1) Designing the test, 2) Doing a pilot study, and 3) Administering the test.

Drawing upon the literature in the field of general education, the researcher designed the questionnaire with 45 items in the 5 scale Likert type. Then two piloting stages were conducted, initial piloting and final piloting. At the initial stage, the 45-item questionnaire was given to two experts and one PhD student for external feedback and revision. As a result, the items were reduced to 34. Since the participants were Iranian, in order to prevent any language barrier and avoid any kinds of misinterpretations on the part of the participants, the 34-item questionnaire was translated into Persian and then back-translated into English by a PhD student in TEFL. The congruency between the two texts was 83.20%.

Then, the final piloting was carried out. During this stage of the pilot study, the Persian version of the 34-item questionnaire was sent to 247 participants through email. Later, to apply the validated questionnaire to the subjects, it was administered again to 120 participants from the same group in the pilot study.

4.2.4. Data Analysis

The internal consistency of the questionnaire was assessed with the Cronbach Alpha reliability estimate.

The validity of the TOEFL iBT speaking section authenticity questionnaire was examined through exploratory factor analysis. First, principal axis factoring identified the underlying factors by calculating the eigenvalues of the matrix greater than 1.0. Because of the subjectivity of the

criterion for selecting absolute value, the researcher decided to consider only factor loadings with an absolute value 0.45 or greater. To decide about the number of factors to retain for rotation, the Scree test was used. Since interpretation of the factors can be very difficult, a solution for this difficulty is factor rotation. As a result, Varimax (orthogonal rotation) with Kaiser Criterion was used. This resulted in a rotated component matrix and a transformation matrix. The rotated component matrix illustrated the variables loaded on each factor so that the researchers came up with four factors.

5. Results

5.1. Task Analysis

In instructional contexts, students are supposed to have certain academic speaking skills. Students should be able to speak successfully in and outside the classroom. For example, in classrooms, students must be able to respond to questions, participate in academic discussions with other students, synthesize and summarize what they have read in their textbooks and heard in class, and express their views on topics under discussion. Outside the classroom, students must have the ability to participate in casual conversations, express their opinions, and communicate with people in such places as the bookstore, the library, and the housing office.

Regarding these types of tasks in the target language academic context, and based on the list of task characteristics in Table 1, the TOEFL iBT speaking section tasks are analyzed here to find out the extent of correspondence between the characteristics of these tasks and those of the TLU tasks.

5.1.1. Characteristics of the setting

The first test method facet, the setting, includes physical characteristics, participants, and time of task. With regard to the physical characteristics, location, physical conditions, materials and equipment, and degree of familiarity are important components. The location of the test is usually a laboratory with a set of computer cabins. The physical condition is usually quiet and well lit. For all speaking tasks, test takers use headsets with a microphone, all of which are familiar to the test takers. With respect to the participants, the only participant is the test taker. Finally, time of task varies, but it is determined in advance and the test is usually administered in daytime. Compared to the physical characteristics of the target language academic contexts, there are various types of settings like classrooms, professor's office, department, the campus, library, or bookstore, in each of which varying types of physical conditions, materials, and equipment are available with different degrees of familiarity on the part of language users. In almost all of these situations, instead of a mechanical interaction between the test taker and a computer, a live conversation between the language user and another participant, like a professor, a teaching assistant, a librarian, a book seller, or peers is required. Furthermore, unlike the testing situation, in all of these environments, at least two participants are involved in the conversations. Finally, although most of the target language use (TLU) tasks take place during daytime, there are some cases, like those in a library or a bookstore that might occur at nights. As a result, with respect to the setting dimension of test method facets, there is little correspondence between the speaking module tasks of the TOEFL iBT and those of TLU.

5.1.2. Characteristics of the input

The second dimension along which speaking tasks of TOEFL iBT and those of TLU can be compared is the input. There are three major components under the test method facet of input: format, language characteristics, and topical characteristics. With regard to the format of the speaking module tasks of the TOEFL iBT, the speaking section is approximately 20 minutes long and includes six tasks. The first two tasks are independent speaking tasks on topics familiar to test takers. They ask test takers to draw upon their own ideas, opinions, and experiences when responding. However, test takers can respond with any idea, opinion, or experience relevant to completing the task. The remaining four tasks are integrated tasks, where test takers must use more than one skill when responding. Test takers first read and listen, and then speak in response. They can take notes and use those notes when responding to the speaking tasks. At least one requires test takers to relate the information from the reading and the listening material. Timing and content areas are fixed for all test takers.

With respect to the Independent Speaking tasks, a single question that appears on the screen is read aloud by the narrator. As a result, the input is presented in this section through both aural and visual channels, while this is not usually the case in TLU tasks. In TLU tasks, whether in the campus situations or academic courses, a question about the language user's own ideas, opinions, and experiences is usually asked aurally and sometimes visually. However, it rarely happens that both channels are used simultaneously.

Timing and content areas are fixed for all test takers. In each of the independent speaking tasks, test takers are informed that they have 15

seconds to prepare an answer, and 45 seconds to respond to each of the independent speaking tasks. A clock shows the remaining time for preparation and response. This format of input in this testing situation has little correspondence with that of TLU tasks which usually occur outside of the classroom in the TL setting. In fact, in TL settings, except for examination situations in an academic context, this rarely happens that language users be given a topic related to their own ideas, opinions, and experiences and before starting to speak be informed that they can prepare themselves within a very short period of time (e.g. 15 seconds) and deliver their speech within only 45 seconds.

With regard to the speededness of the input, the speed of the narrator who reads the question on the screen aloud is fixed for all test takers and cannot be slowed down in case of lower proficiency. In addition, the question cannot be repeated in case of misunderstanding on the part of the test taker. However, in TLU tasks, native speakers adjust the speed of their speech to the proficiency level of the foreigners if needed, and in case of misunderstanding, they can repeat the question for them. All these problems in the testing situation arise from the vehicle, which is not live, but reproduced via computers. As a result, there is not enough correspondence between the format of the tasks of the TOEFL iBT speaking module and those of TLU tasks.

Language characteristic is another aspect of the input. This aspect, in turn, includes two major components, i.e. organizational characteristics, and pragmatic characteristics. The organizational and pragmatic characteristics of the input are fixed for all the participants. Since the first two independent speaking tasks of the TOEFL iBT are more concerned with out-of-the-classroom conversations, the vocabularies are more general; however, more

specialized and academic vocabularies can be seen if the topics are about academic courses or the like. Furthermore, Standard English is used with regard to the morphology and syntax. However, the integrated tasks are more concerned with academic issues which necessitate the use of more specialized and academic vocabularies and again the morphology and syntax are based on Standard English. With respect to the correspondence between the organizational characteristics of the input in test task and those in TLU tasks, we might expect more standard English-like morphology and syntax in classrooms; however, this is not the case in out-of-the-classroom situations, where peers make more use of informal and casual varieties and even sometimes slangs. As a result, correspondence between test tasks and TLU tasks in this respect is weak, too.

With regard to the pragmatic characteristics of the input, it should be noted that a wide range of functions, like ideational and manipulative (describing, justifying, proposing, arguing, comparing, contrasting) ones are elicited by the six tasks of the speaking module in the TOEFL iBT. The sociolinguistic aspects of the input, such as dialect/variety, register, naturalness, cultural references, and figurative language are rather fixed and have a tendency more toward the standard, formal, and academic language. As a result, the correspondence between the speaking tasks of the test and those of the TLU is not strong.

With regard to the topical characteristics, it can be said that there is a relatively good correspondence between the topics of the speaking tasks of the TOEFL iBT and those of TLU tasks. The six tasks usually cover a wide range of topics both in classrooms and outside of the classroom in TL situation. In classrooms, students must respond to questions, participate in academic discussions with other students, synthesize, and summarize what

they have read in their textbooks and heard in class, and express their views on topics under discussion. Outside the classroom, students need to participate in casual conversations, express their opinions, and communicate with people in such places as the bookstore, the library, and the housing office. Most of these topics are presented through the six tasks of the speaking module of the TOEFL iBT test. In fact, the topics of the first two independent tasks are more concerned with out-of-the-classroom situations and those of the integrated four tasks are more concerned with the academic, in-class settings or campus environments.

5.1.3. Characteristics of the expected response

The third test method facet in Bachman and Palmer's (1996) model is the expected response. Most of the problems relating to the characteristics of the input, are observed in the expected response of the speaking section tasks of the TOEFL iBT. In order to answer the questions, all test takers need to use headsets with a microphone. Test takers speak into the microphone to record their responses. Responses are digitally recorded and sent to ETS's Online Scoring Network, where they are scored by certified raters. As a result, the channel is merely aural. The length of the responses is limited to the allowed time determined in advance for each task. Although there is time limitation for responses in TLU tasks, there is more flexibility and it is not so mechanically pre-determined.

In the testing situation, good responses are fluid and clear with good pronunciation, natural pacing, and natural-sounding intonation patterns. Raters determine the test taker's ability to control both basic and more complex language structures, and use appropriate vocabulary. Test takers are expected to answer the questions coherently in the presentation of their

ideas. Test taker should be able to synthesize and summarize the information in the integrated tasks. Good responses generally use all or most of the allotted time, and the relationship between ideas and the progression from one idea to the next is clear and easy to follow. However, it is important to note that raters do not expect test takers' responses to be perfect. Even high-scoring responses may contain occasional errors and minor problems in any of the three areas described above. The major problem with this type of expected response in the testing situation is the raters' lack of concern with the sociolinguistic aspects of the language used by the test takers. In the TLU context, language users are expected to be familiar with pragmalinguistic and sociopragmatic aspects of the language when speaking to different members of the TL community in different social positions and different social distances from each other. Lack of knowledge relating these issues can cause various social problems for language users. In addition, the fact that the examinee's body language is not considered in their response can exacerbate the situation.

5.1.4. Relationship between input and response

The final test method facet to be discussed here is the relationship between input and response. In the TOEFL iBT, this task characteristic could be considered as the most problematic one among others. This facet includes adaptability and reciprocity of the setting among others. On the speaking section of the TOEFL iBT, the computerized format of the input precludes adaptability and reciprocity. In non-reciprocal language use like the TOEFL iBT testing situation, there is neither feedback nor interaction between language users. The standardized format remains the same regardless of the nature of response by the test taker. The test taker is unable to ask for clarification of directions or further explanations of tasks; in turn, the rater

cannot ask the test taker to further explain a point or a word that is unclear. Furthermore, negotiation and discussion which is an important part of the academic context of the TL classrooms is not possible in this format.

With respect to the scope of relationship, TOEFL iBT speaking section tasks, especially the integrated tasks are considered to have a broad scope. In these tasks, the range of input that must be processed in order for the test taker to respond as expected is broad. Test takers should listen to a conversation or read a text and try to answer a related question. The same tasks exist in TL academic settings, especially in the classroom. Consequently, there is an acceptable amount of correspondence between the test tasks and TLU tasks regarding the scope of relationship.

Finally, the directness of relationship is concerned with the degree to which the expected response can be based primarily on information in the input, or whether the test taker or language user must also rely on information in the context or in his/her own topical knowledge. According to Bachman and Palmer (1996), "many, if not most, TLU tasks involve an indirect relationship between input and response" (p. 56). In a conversation, for example, the language users expect each other to respond with new, rather than given information, the new information being supplied by the language users. The speaking section tasks of the TOEFL iBT test are more indirect than direct. As a result, the correspondence between test tasks and TLU ones is good enough.

In sum, the computerized format of the TOEFL iBT results in a nonreciprocal test setting which precludes live interaction between the test giver and the test taker. The questions are prerecorded and the speaker is left to respond. However, in TLU tasks, a live face-to-face interaction allows

communication breakdown to be questioned and repaired, and questions of meaning can be clarified. The classroom setting involves several participants, whereas the TOEFL iBT typically involves only one participant, i.e., the test taker.

As a result, based on the comparison between the characteristics of test tasks and those of TLU tasks, it was revealed that the correspondence between them and consequently the authenticity of the speaking module of the TOEFL iBT test is limited.

5.2. Task Survey

5.2.1. Reliability of the Authenticity Questionnaire

To estimate the reliability of the final version of the authenticity questionnaire which included 30 items, Cronbach Alpha was run. The results of the analysis revealed a good reliability index of 0.81.

5.2.2. Validity of the Questionnaire

The authenticity questionnaire, which was reduced to 34 items after the initial piloting, was administered to 247 subjects by email to examine the construct validity of its factor structure through exploratory factor analysis. PCA extracted 10 factors with eigenvalues greater than 1.0 which accounted for about 62% of the variance. Out of 34 items, 30 items had loadings of 0.45 or greater on any factor. The results of the Scree Test indicated that a four-factor solution might provide a more parsimonious grouping of the items in the questionnaire.

Then, orthogonal rotation was run. Varimax with Kaiser Normalization resulted in a rotated component matrix which represented the underlying factor structure. The first factor consisted of 9 items. The second factor consisted of 11 items. Factor 3 consisted of 6 items and items 4, 23, 27, and 30 made up the fourth factor. The total number of items was 30.

After analyzing items comprising each factor, the researcher came to the four original factors of context, student factor, task factor, and indicators. Items representing each factor are displayed in Appendix A, and the validated questionnaire is given in Appendix B.

5.2.3. Application of the TOEFL iBT speaking section authenticity questionnaire

After the pilot study and validation of the questionnaire which was developed by the researcher, the final version of the questionnaire which consisted of 30 items was sent to the same participants through email. 120 subjects responded to the questions by selecting from five options, i.e. 'strongly agree', 'agree', 'undecided', 'disagree' or 'strongly disagree'. The items of the authenticity questionnaires were examined in terms of their percentage so as to see what the subjects' general attitude is toward the factors representing the authenticity of the speaking module of the TOEFL iBT test. To better illustrate the pattern of the respondents' answers to the questionnaire, the first two alternatives (strongly agree and agree) and the last two (disagree and strongly disagree) were combined (see Table 2).

Table 2

Test takers' attitude in terms of frequency (F) and percentage (P)

Items	**SA + A**		**U**		**D + SD**	
	F	**P**	**F**	**P**	**F**	**P**
1. The tasks are the kinds of tasks the examinee might be required to perform in real academic life situation.	102	85%	_	_	18	15%
2. The tasks require the examinee to spend a significant amount of time on the task in collaborative groups.	54	45%	48	20%	42	35%
3. Both the final answer and the route(s) that the examinee takes to come to that answer are considered.	96	80%	12	10%	12	10%
4. In addition to the test, there are other indicators to assess the examinee's speaking ability.	30	25%	42	35%	48	40%
5. The tasks address real-world public problems.	30	25%	48	40%	42	35%
6. The assessment condition is similar to the real-world context in which the task might be performed.	66	55%	_	_	54	45%
7. Collaboration is integral to the task, rather than achievable by an individual learner.	30	25%	18	15%	72	60%

8. The examinee has choice and freedom to show his/her oral proficiency in different ways.	78	65%	18	15%	24	20%
9. The tasks engage the examinee in a variety of tasks, like writing, revising, discussing, providing an engaging oral analysis of an event, collaborating with others on a debate, etc.	96	80%	12	10%	12	10%
10. The tasks have clear connection to issues or experience beyond the assessment context.	48	40%	42	35%	30	25%
11. In doing the tasks, there is an adequate opportunity to plan, revise and substantiate responses.	18	15%	18	15%	84	70%
12. The tasks ask students to create new meaning via a complex process, rather than only recall facts and ideas.	78	65%	18	15%	24	20%
13. The tasks show the process the examinee goes through to reach the correct answer.	42	35%	24	20%	54	45%
14. There is a connection between the tasks and the larger social context within which the examinee will live.	72	60%	18	15%	30	25%
15. The tasks ask examinees to demonstrate understanding by performing a set of complex tasks, like recognition and asking questions.	78	65%	6	5%	36	30%
16. The tasks afford learners the opportunity to examine the problem from a variety of theoretical and practical perspectives.	78	65%	18	15%	24	20%
17. There are multiple acceptable routes towards performing the task rather than only one predetermined and carefully structured answer or performance.	78	65%	6	5%	36	30%
18. The tasks are meaningful in such a way that it replicates real world challenges to see if students are capable of doing so.	54	45%	24	20%	42	35%
19. The tasks cannot be completed by short answers.	90	75%	18	15%	12	10%

20. The tasks primarily support the needs of examinees; i.e. they are enabling and forward-looking, not just reflective of prior teaching.	48	40%	6	5%	66	55%
21. The tasks attend to whether the examinee can craft justifiable answers, rather than typically only asking the examinee to select or write correct responses--irrespective of reasons.	96	80%	18	15%	6	5%
22. The tasks provide the opportunity for students to examine it from different perspectives, using a variety of resources.	96	80%	12	10%	12	10%
23. The test considers other types of performance, like the students' portfolio, special projects, etc.	24	20%	24	20%	72	60%
24. The tasks have value and meaning beyond the assessment context; i.e. activities are not deemed important for success only in the assessment environment.	60	50%	_	_	60	50%
25. The tasks require the examinees to manipulate information to discover new meanings and understandings rather than just to recite factual information.	90	75%	6	5%	24	20%
26. The tasks ask students to analyze, synthesize and apply what they have learned in a substantial manner.	78	65%	24	20%	18	15%
27. The test permits observation of patterns of strength and weakness over a sustained period.	12	10%	30	25%	78	65%
28. The tasks have the examinees to use personal experiences as a context for applying knowledge.	72	60%	12	10%	36	30%
29. In the tasks, the examinees are asked to demonstrate proficiency by doing something rather than selecting from four alternatives to indicate their proficiency.	72	60%	42	35%	6	5%
30. The test provides multiple indicators of success.	24	20%	24	20%	72	60%

As Table 1 reveals, the majority of the subjects agree with the authentic characteristics of the speaking module of the TOEFL iBT test. In fact, out of 30 items, most of the respondents agree with 20 items (1, 2, 3, 6, 8, 9, 10, 12, 14, 15, 16, 17, 18, 19, 21, 22, 25, 26, 28, 29) and disagree with 9 items (4, 5, 7, 11, 13, 20, 23, 27, 30). The condition of item 24 is 50-50. For example, most of the respondents believe that "the tasks are the kinds of tasks the examinee might be required to perform in real academic life

situation" (Item 1), or most of them agree with the fact that "the tasks require justifiable answers, rather than typically only asking the examinee to select or write correct responses - irrespective of reasons" (Item 22). However, they do not believe that "to evaluate the speaking ability of the examinee, the test provides indicators other than the test itself" (Item 4), or they do not agree that "collaboration is integral to the task" (Item 7). Overall, regarding the 30 items of the questionnaire, the participants expressed positive attitudes toward the authenticity of the speaking module of the TOEFL iBT with a mean of 104.71 and a standard deviation of 11.57. Since the overall mean is more than one standard deviation above the neutral point (90), it can be concluded that the subjects, who had already taken the TOEFL iBT test, had a positive attitude towards the authenticity of the test. To present a more vivid picture of the findings, the items of the questionnaire are categorized and summarized based on the four factors underlying the questionnaire items (see Table 3).

Table 3

Test takers' attitude regarding the four factors of authenticity

Factor	SA + A	U	D + SD
Context (Items 1, 5, 6, 10, 14, 18, 20, 24, 28)	51.1%	13.8%	35%
Student factor (Items 2, 7, 11, 12, 15, 16, 19, 21, 25, 26, 29)	57.7%	14%	30.4%
Task factor (Items 3, 8, 9, 13, 17, 22)	67.5%	11.6%	20.8%
Indicators (4, 23, 27, 30)	18.7%	16.6%	56.2%

As Table 2 illustrates, more than 50% of the test takers, agree with the three criteria of context, student factor, and task factor. However, with regard to the last factor, indicators, it is revealed that test takers mostly disagree.

Furthermore, out of the first three factors, task factor gains the highest degree of agreement. In other words, test takers in this study think that the speaking section of the TOEFL iBT involves complex, ill-structured challenges that require judgment, and a full array of tasks (Linn et al., 1991; Torrance, 1995; Wiggins, 1990, 1993, 1989). With regard to the last criterion, indicators, most of the examinees disagree with the fact that the test provides multiple indicators of learning (Lajoie 1991; Linn et al., 1991). This attitude is quite correct because except for the test itself, there is no other indicator of the examinee's speaking ability. It is only through the six tasks in the speaking section of the test that the speaking ability of test takers is determined.

6. Discussion

Comparing the results of the task analysis with those of the survey, we observed a contradiction in terms of the authenticity of the speaking module of the TOEFL iBT test. That is, unlike the task analysis which did not show adequate authenticity regarding the speaking module, the candidates believed that the test is in a satisfactory level of authenticity. The fact that the task analysis did not confirm the authenticity of the test is against the findings of Meng-li (2010) and Ockey et al. (2015) who found the test fairly authentic. One simple and clear explanation for the results of the task analysis is the fact that, except for some rare situations, students hardly communicate orally with a machine in a real academic context. Consequently, most of the factors, in Bachman and Palmer's (1996) model, that recognize a language test as an authentic one do not match the mechanical and unreal conditions of the speaking module of the TOEFL iBT test, and more particularly, as Ockey et al. (2015) also found in their studies, this test does not correspond to face-to-face student-student and

teacher-student conversations that consist quite a large part of the interactions in an academic context. This limitation can simply question the authenticity of the setting, the expected response, the input, and the relationship between input and response.

Another explanation for the results of the task analysis could be the fact that, as Meng-li (2010) concluded in his study, the authenticity of the tests depends on the definition of concept of 'authenticity' itself and the interaction between test takers and test tasks. As Pinner (2015; 2016) argues, authenticity is a dynamic and multidimensional concept which depends on a variety of factors including the learner's motivation, needs, social context, and so many other factors which might not be considered in a static model of authenticity that implies 'one size fits all'. Therefore, the TOEFL iBT speaking module might be authentic according to one model of authenticity but might not be authentic enough based on the other like what Bachman and Palmer's (1996) model revealed in the present study. This shows the urging need for developing a comprehensive model of authentic language testing which can take different factors and variants into consideration and have enough flexibility and dynamicity in dealing with a variety of learners in different contexts.

Regarding the results of the task survey which confirmed the authenticity of the speaking module, three explanations could be raised. The first explanation, which at the same time could be considered as one of the limitations of the study, is that the authenticity questionnaire was given to those who had already taken the TOEFL iBT test. In this respect, these participants are the appropriate ones for this study due to their familiarity with the test and the tasks. However, the problem arises when it is not clear whether all of these test takers have already experienced the TL academic

context. That is, if these subjects had not ever been to an English language university, they could not have made a good judgment, especially when they try to express their attitudes regarding the relationship between the types of test tasks with those in the real context. As a result, their answers might not be a good indication of the authenticity of the speaking module tasks. The solution to this problem could be giving the questionnaire to those subjects who both have already taken the test and also had the experience of studying in the TL academic context.

The second explanation for the results of the task survey could be due to what Malone & Montee (2014) found as mixed attitudes considering the four skills and the candidate's nationality on the TOEFL iBT test. In case of Iranian candidates, their positive view towards the authenticity of the test could be justified based on the fact that for a long time, Iranian students' general proficiency of English was assessed based on TOEFL PBT which doesn't have any speaking module. As a result, the emergence of the TOEFL iBT with a speaking section is considered more authentic by students than the TOEFL PBT which did not test their oral communicative competence at all.

The last explanation which is probably the most thought-provoking justification is concerned with the limitation of the existing authenticity models in the field of language testing, the last one of which is that of Bachman and Palmer's (1996). The contradiction between the results of the task analysis and the task survey could be due to the different models based on which the analysis and the survey were carried out. The task analysis was based on Bachman and Palmer's model of language testing, while the task survey built upon Herrington and Herrington's (1999, 2006) authenticity criteria. In the former, the correspondence between the features of the test

tasks and those of the TLU tasks is central. However, in the latter, this concern is represented in only one of the four authenticity criteria, i.e. context. Other factors concentrate more on the test takers' performance or the characteristics of the tasks, and indicators. Therefore, it is probable that what has been proved inauthentic based on a model considering only one factor (here, the factor of context in Bachman and Palmer's model) might turn out at least fairly authentic based on a model including more criteria (here, Herrington and Herrington's model). Although Bachman and Palmer's (1996) definition of authenticity appropriately accentuates the correspondence between the characteristics of the test tasks and those of the TLU tasks, it ignores, to a large extent, other crucial factors, like student factor, task factor, and indicators that are paid special attention to in the realm of general education. For example, Bachman and Palmer's (1996) definition of authenticity pays less attention to such facts that in an authentic assessment, students should be effective performers with acquired knowledge, and are expected to craft polished performances or products. It also ignores the necessity of significant student time and effort in collaboration with others, complex and ill structured challenges that require judgment, and a full array of tasks. Seamless integration with the activity and multiple indicators of learning are other factors that should be taken into consideration in an authentic assessment. Finally, the fact that an authentic assessment should achieve validity and reliability with appropriate criteria for scoring varied products is of great importance.

7. Conclusion

In conclusion, the contradictory results of the task analysis and task survey, which were based on authenticity models from the fields of applied linguistics and general education respectively, cast doubts on the adequacy

of the existing models of authenticity in language testing and illuminated some ignored and valuable aspects which should be considered in the models. Following Lewkowicz's (2000) suggestion, the present study appropriately revealed the very advantage of combining models and principles of authenticity from the two fields of applied linguistics and general education, showing the fact that the two fields can mutually benefit each other to provide a more comprehensive and inclusive model of authenticity, especially in the field of language testing which suffers from scarcity of research regarding authentic assessment (Pinner, 2016).

The results of the present study can also be considered as a start point for further empirical research to provide more evidence supporting the advantages of the interaction between the two fields. This can open new horizons towards novel ideas and concepts in the realm of authenticity and authentic assessment and bring about more insightful understandings in this respect.

8. References

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

Bachman, L. (1991). What does language testing have to offer? *TESOL Quarterly, 25* (4), 671–704.

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

Breen, M. (1985). Authenticity in the language classroom. *Applied Linguistics, 6*, 60–70.

Broughton, G. (1965). *A technical reader for advanced students*. London: Macmillan.

Burkill, S., Dunne, L., Filer, T. and Zandstra, R. (2009). *Authentic voices: Collaborating with students in refining assessment practices*, Presentation at ATN Assessment Conference, RMIT University.

Dunne, L., Filer, T., & Zandstra, R. (2009, November). Authentic voices: collaborating with students in refining assessment practices. In *ATN Assessment Conference 2009: Assessment in Different Dimensions* (p. 84).

Close, R.A. (1965). *The English we use for science*. London: Longman.

Farnsworth, T. (2013). Assessing the oral English abilities of international teaching assistants in the USA. *The Companion to Language Assessment, 1*, 471-483.

Frey, B., & Schmitt, V. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics, 18*, 402-423.

Gilmore, A. (2007). Authentic materials & authenticity in foreign language learning. *Language Teaching, 40*(2), 97-118.

Gilmore, A. (2009). The times they are a-changin': Strategies for exploiting authentic materials in the language classroom. In Rilling, S. & Dantas Whitney, M. (eds.), *TESOL classroom practice series: Authenticity in adult classrooms and beyond.* Virginia: TESOL Publications, 155-168.

Gilmore, A. (2011). "I prefer not text": Developing Japanese learners' communicative competence with authentic materials. *Language Learning, 61*(3), 786-819.

Gulikers, J., Bastiaens, T., & Kirschner, P. (2006). Authentic assessment, student teacher perceptions: The practical value of the five-dimensional framework. *Journal of Vocational Education and Training, 58*, 337-357.

Herrington, J., & Herrington, A. (1998). Authentic assessment and multimedia: How university students respond to a model of authentic assessment. *Higher Education Research and Development, 17*(3), 305-322.

Herrington, J., & Herrington, A. (2006). Authentic conditions for authentic assessment: Aligning task and assessment. In A. Bunker, & I. Vardi (Eds.). *Research and development in higher education* Volume 29, (pp. 146-151). Milperra, NSW: HERDSA.

Herrington, J., Oliver, R., & Reeves, T.C. (2003). Patterns of engagement in authentic online learning environments. *Australian Journal of Educational Technology, 19*(1), 59-71.

Hung, D., & Victor Chen, D. T. (2007). Context–process authenticity in learning: implications for identity enculturation and boundary crossing. *Educational Technology Research and Development, 55*(2), 147-167.

Ingram, D.E. (2003). Methodology in the new millennium: *Towards more authenticity in language learning and assessment.* Paper to the First International Conference on pedagogies and learning, New meanings for the new millennium, University of Southern Queensland, Toowoomba, 1-4

October, 2003.

Keyser, S., & Howell, S. (2008). *The state of authentic assessment.* Education Resources Information Center (ERIC) database. (ERIC Document No ED503679) (10 March 2010).

Kroll, D.L., Masingila, J.O., & Mau, S.T. (1992). Grading cooperative problem solving. *Mathematics Teacher*, *85*(8), 619-627.

Lajoie, S. (1991). A framework for authentic assessment in mathematics. *NCRMSE Research Review: The Teaching and Learning of Mathematics, 1*(1), 6-12.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for languageassessment literacy. *Language Testing, 32*(2), 169-197.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, 17: 43-64.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21.

Malone M.E. (2017). Training in Language Assessment. In E. Shohamy, & S. May, (eds). *Language testing and assessment. Encyclopedia of language and education* (3rd ed.). (pp. 225-239). Springer, Cham.

Malone, M. E., & Montee, M. (2014). Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability. *ETS Research Report Series, (2)*, 1-51.

Meng-li, L. I. (2010). On the authenticity of iBT TOEFL oral test [J]. *Journal of Chongqing University of Posts and Telecommunications (Social Science Edition), 4*, 1-27.

Meyer, C.A. (1992). What's the difference between authentic and performance assessment? *Educational Leadership, 49*(8), 39-40.

Mishan, Freda. (2005). *Designing Authenticity into Language Learning Materials*. Bibliovault OAI Repository, the University of Chicago Press.

Morrow, K. (1977). Authentic Texts in ESP. In S. Holden (Ed.). *English for specific purposes*. (pp. 13-17). London: Modern English Publications.

Morrow, C. K. (2018). Communicative language testing. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-7). New Jersey: Wiley-Blackwell

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19* (4), 337–346.

Norris, J. M. (2009). Task-based teaching and testing. In M. Long, & C., Doughty (Eds.), *Handbook of language teaching* (pp. 578–594). Cambridge, MA: Blackwell

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39-62.

Pinner, R. (2015). *What we talk about when we talk about Authenticity* [Invited guest post for popular EFL blog]. Retrieved from http://malingual.blogspot.jp/2015/02/what-we-talk-about-when-we-talk-about.html

Pinner, R. (2016). The nature of authenticity in English as a foreign language: A comparison of eight inter-related definitions. *ELTWO Journal, 9*(1), 78-93.

Powers, D., & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1*(2), 153–173.

Reeves, T.C., & Okey, J.R. (1996). Alternative assessment for constructivist learning environments. In B.G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 191-202). Englewood Cliffs, NJ: Educational Technology Publications.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford, & M.C. O'Connor (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph No. MS-21). Princeton, NJ: Educational Testing Service.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 005-30.

Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement, 38*(2), 265–273.

Schmitt, N., Gilliland, S. W., Landis, R. S., & Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology, 46*(1), 149–165.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2* (l), 31–40.

Stevenson, D. (1985). Authenticity, validity, and a tea party. *Language Testing, 2* (1), 41–47.

Stricker, L.,Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based Test of English as a Foreign Language. *Computers in Human Behavior, 20*, 37–54.

Torrance, H. (1995). Introduction. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment* (pp. 1-8). Buckingham: Open University Press.

van Lier, L. (1996). Interaction in the language curriculum: Awareness, autonomy and authenticity. London: Longman.

Widdowson, H. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70(9), 703-713.

Wiggins, G. (1990). *The case for authentic assessment*. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation. (ERIC Document Reproduction Service No. ED 328 606).

Wiggins, Grant (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation, 2*(2). Retrieved March 21, 2014 from http://PAREonline.net/getvn.asp?v=2&n=2 .

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass.

Xi, X. (2008). *Validating the use of TOEFL iBT speaking section scores for ITA screening and setting standards for international teaching assistants*. (Research Spotlight. No. 1). Princeton, NJ: Educational Testing Service.

Young, M.F. (1995). Assessment of situated learning using computer environments. *Journal of Science Education and Technology, 4*(1), 89-96.

Zahedi, K., & Shamsaee, S. (2012). Viability of construct validity of the

speaking modules of international language examinations (IELTS vs. TOEFL iBT): Evidence from Iranian test-takers. *Educational Assessment, Evaluation and Accountability, 24*(3), 263-277.

Appendix A: The Factors of TOEFL iBT Speaking Section Authenticity Questionnaire

Factor 1: Context

1. The task is the kind of task the examinee might be required to perform in real academic life situation.
2. The task addresses a real-world public problem.
3. The assessment condition is similar to the real-world context in which the task might be performed.
4. The tasks have the examinees to use personal experiences as a context for applying knowledge
5. The task has clear connection to issues or experience beyond the assessment context.
6. The task is meaningful in such a way that it replicates real world challenges to see if students are capable of doing so.
7. The task has value and meaning beyond the assessment context; i.e. activities are not deemed important for success only in the assessment environment.
8. The tasks primarily support the needs of examinees; i.e. they are enabling and forward-looking, not just reflective of prior teaching.
9. There is a connection between the task and the larger social context within which the examinee will live.

Factor 2: Student Factor

1. The task requires the examinee to spend a significant amount of time on the task in collaborative groups.
2. Collaboration is integral to the task, rather than achievable by an individual learner
3. In doing the tasks, there is an adequate opportunity to plan, revise and substantiate responses.
4. In the task, the examinees are asked to demonstrate proficiency by doing something rather than selecting from four alternatives to indicate their proficiency.
5. The task asks examinees to demonstrate understanding by performing a set of complex tasks, like recognition and asking questions.
6. The task asks students to analyze, synthesize and apply what they have learned in a substantial manner
7. The tasks ask students to create new meaning via a complex process, rather than only recall facts and ideas.
8. The task requires the examinees to manipulate information to discover new meanings and understandings rather than just to recite factual information.
9. Tasks cannot be completed by short answers
10. The tasks attend to whether the examinee can craft justifiable answers, rather than typically only asking the examinee to select or write correct responses--irrespective of reasons.
11. The task affords learners the opportunity to examine the problem from a variety of theoretical and practical perspectives .

Factor 3: Task Factor 1. Both the final answer and the route(s) that the examinee takes to come to that answer are considered 2. The examinee has choice and freedom to show his/her oral proficiency in different ways 3. The tasks engage the examinee in variet of tasks, like writing, revising, discussing, providing an engaging oral analysis of an event, collaborating with others on a debate, etc. . 4. The tasks show the process the examinee goes through to reach the correct answer 5. The tasks provide the opportunity for students to examine it from different perspectives, using a variety of resources. 6. There are multiple acceptable routes towards performing the task rather than only one predetermined and carefully structured answer or performance.
Factor 4: indicators 1. In addition to the test, there are other indicators to assess the examinee's speaking ability. 2. The test permits observation of patterns of strength and weakness over a sustained period. 3. The test considers other types of performance, like the students' portfolio, special projects, etc. 4. The test provides multiple indicators of success.

Appendix B: TOEFL iBT Speaking Section Authenticity Questionnaire

Please tick the boxes below which best describes your attitude towards the authenticity of the TOEFL iBT speaking section tasks.

Items	SA	A	U	D	SD
1. The task is the kind of task the examinee might be required to perform in real academic life situation.					
2. The task addresses a real-world public problem.					
3. The assessment condition is similar to the real-world context in which the task might be performed.					
4. The tasks have the examinees to use personal experiences as a context for applying knowledge.					
5. The task has clear connection to issues or experience beyond the assessment context.					
6. The task is meaningful in such a way that it replicates real world challenges to see if students are capable of doing so.					
7. The task has value and meaning beyond the assessment context; i.e. activities are not deemed important for success only in the assessment environment.					
8. The tasks primarily support the needs of examinees; i.e. they are enabling and forward-looking, not just reflective of prior teaching.					
9. There is a connection between the task and the larger social context within which the examinee will live.					

10. The task requires the examinee to spend a significant amount of time on the task in collaborative groups.					
11. Collaboration is integral to the task, rather than achievable by an individual learner					
12. In doing the tasks, there is an adequate opportunity to plan, revise and substantiate responses.					
13. In the task, the examinees are asked to demonstrate proficiency by doing something rather than selecting from four alternatives to indicate their proficiency.					
14. The task asks examinees to demonstrate understanding by performing a set of complex tasks, like recognition and asking questions.					
15. The task asks students to analyze, synthesize and apply what they have learned in a substantial manner					
16. The tasks ask students to create new meaning via a complex process, rather than only recall facts and ideas.					
17. The task requires the examinees to manipulate information to discover new meanings and understandings rather than just to recite factual information.					
18. Tasks cannot be completed by short answers.					
19. The tasks attend to whether the examinee can craft justifiable answers, rather than typically only asking the examinee to select or write correct responses--irrespective of reasons.					
20. The task affords learners the opportunity to examine the problem from a variety of theoretical and practical perspectives.					
21. Both the final answer and the route(s) that the examinee takes to come to that answer are considered.					
22. The examinee has choice and freedom to show his/her oral proficiency in different ways.					
23. The tasks engage the examinee in a variety of tasks,					

like writing, revising, discussing, providing an engaging oral analysis of an event, collaborating with others on a debate, etc.					
24. The tasks show the process the examinee goes through to reach the correct answer					
25. The tasks provide the opportunity for students to examine it from different perspectives, using a variety of resources.					
26. There are multiple acceptable routes towards performing the task rather than only one predetermined and carefully structured answer or performance.					
27. In addition to the test, there are other indicators to assess the examinee's speaking ability.					
28. The test permits observation of patterns of strength and weakness over a sustained period.					
29. The test considers other types of performance, like the students' portfolio, special projects, etc.					
30. The test provides multiple indicators of success.					

SA: Strongly Agree, A: Agree, U: Undecided, D: Disagree, SD: Strongly Disagree

Note on Contributor:

Marzieh Souzandehfar is an assistant professor in TEFL. She obtained her PhD degree from Shiraz University. She is now affiliated with the Department of Translation Studies at Jahrom University, Jahrom, Fars, Iran. She teaches Teaching Methods, Testing, Research Methods, and Contrastive Analysis at undergraduate levels. Her research interests include Testing, CDA, Multiliteracies, and Second Language Speaking. She has published more than 10 articles in scholarly journals and has presented papers at national conferences.