Questioning in TOEFL iBT speaking test: a case of washback and construct underrepresentation

Ali Sayyadi[1] and Reza Rezvani[2*]

* Correspondence: rezvanireza@
gmail.com
[2]Department of English Language,
Yasouj University, Yasuj, Iran
Full list of author information is
available at the end of the article

Abstract

Admission in English-medium universities or institutions of higher education depends on the results obtained by candidates in large-scale proficiency tests including Test of English as a Foreign Language (TOEFL) internet-based test (iBT). The structure and administration procedure of the TOEFL iBT speaking test leaves no room for carrying out reciprocal interactions and, consequently, examining applicants' questioning ability. This study highlights the significance of non-native students' skills to ask oral English questions in academic contexts as experienced by non-native graduates from English-medium universities and in the view of Iranian TOEFL iBT instructors. Further, the washback of the absence of the skill in the TOEFL iBT test was investigated in speaking preparation classes. Twelve non-native graduates and nineteen Iranian TOEFL iBT instructors participated in the study. They were all interviewed about the significance of oral questioning. The instructors' views were also sought about the consequences of the disregard of questioning in the test. To triangulate the data, two classes of the instructors were also observed. Classroom observations and interviews were analyzed through content analysis. The results indicated that the participants generally assumed questioning as indispensable in academic interactions. Despite their acknowledgment of its significance, as both the interviews and observations unveiled, the instructors, however, tended not to work on their students' questioning because of the absence of the skill in the test, the students' reluctance, limited preparation time, and dependence of their professional reputation on the test results rather than target situation performance. The study further discusses the implications of the findings for the test construct representation and preparation courses.

Keywords: TOEFL iBT, Higher education, Questioning, Washback, Construct underrepresentation, Preparation courses

Introduction

Historically, critical decisions have been made about English learners' and users' language proficiency based on their performance on high-stakes tests. The results obtained from high-stakes tests might give rise to serious consequences which directly impact the educational and occupational decisions to be made for test-takers. Many

people throughout the world seek to migrate to English-speaking countries (ESC), for instance, to pursue their higher education, and their admission to English-medium universities is highly dependent on their performance on large-scale tests. Accordingly, studying various aspects of such tests with the aim to improve their usefulness has been assumed to be increasingly critical because more reliable and valid results by such tests help universities to identify and select more qualified applicants with adequate communication skills to cope with their academic studies (Chapelle, 2020).

Test of English as a Foreign Language (TOEFL) internet-based test (iBT) speaking test has been widely used to assess applicants' oral communicative capacities and, consequently, predict their readiness to manage the language demands of higher education (Brown & Ducasse, 2019). It is a reportedly reliable and valid speaking test (Bridgeman et al., 2012; Chapelle et al., 2008; Enright & Tyson, 2008; Ockey et al., 2015) since "it seeks to simulate typical authentic academic activities, and the scores are said to extrapolate to performance in real-life academic settings" (Brooks & Swain, 2014, p. 354). In other words, it simulates a life-like situation in which the test-takers are to fulfill the tasks they are going to confront in the target language use situation (TLUS) (Brown & Ducasse, 2019).

Given its structure and administration procedure (for details, see Alderson, 2009), it seems that the TOEFL iBT speaking test assesses examinees' communicative abilities solely based on their responding skills and considers no room for judging their questioning skills (Brooks & Swain, 2014). This is arguably incongruent with the commission intended for TOEFL iBT speaking test to simulate TLUS tasks in English-medium universities since questioning plays an integral part in interactions in academia (Almeida, 2012; Dool et al., 2020; Rezvani & Sayyadi, 2015). In seminars, conferences, laboratories, and classrooms, for instance, questioning might be employed by attendees to obtain information, maintaining the control of academic debates, seeking clarification and further elaboration, exploring the insights of addressees, encouraging further thoughts, and merging prior knowledge and new information in attempts to make sense of ideas, and so forth (Almeida, 2012; Dool et al., 2020; Rezvani & Sayyadi, 2015).

Despite the contribution of previous studies in enhancing the body of the related literature, it appears that the significance and benefits of questioning in academia as perceived by actual language users are under-researched. As one of its objectives, the current study accordingly seeks to investigate the significance and usefulness of non-native students' (NNS) skill to ask oral English questions (OEQ) in academic contexts in English-speaking countries (ESC) as viewed by non-native graduates (NNG) and Iranian TOEFL iBT instructors.

Tests and, more specifically, their defining characteristics might have substantial washback effects on test-takers, teachers, and the preparation courses. Language learners, for example, who are preparing to sit a test with certain features and demands may employ varying study plans and strategies, develop different motivations, and adopt disparate test-taking strategies (Hung & Huang, 2019; Rezvani & Sayyadi, 2014). To teachers, the characteristics and demands of a given test is decisive in opting for the most appropriate teaching methodologies, materials, syllabus design, and assessment strategies (Ali & Hamid, 2020; Rezvani & Sayyadi, 2016). Accordingly, the absence of room for questioning in the test might bring about critical washback effects on TOEFL iBT instructors' pedagogical perception and practice. To date, however, no study has

probed how TOEFL iBT applicants' questioning skill is treated by teachers in speaking preparation courses as a washback effect of the test. Therefore, with this being the case, the current study was conducted to contribute to enhancing the body of knowledge in the related literature.

Literature review

The significance of questioning

To perform a lucid conversation, as a systematic event, conversants need to understand and respond to one another to organize their social activities (Gardner, 2019). This ping-pong transaction involves both responding to messages received and raising questions. Various studies have highlighted the significance of questioning as a communicative performance. Rezvani and Sayyadi (2015), for instance, consider effective questioning as one of the six basic social and psychological needs of human beings, enabling them to accomplish communicative goals in communities. Chudinova (2020) also pointed to questioning as an indispensable element for performing various social speech acts such as requesting, inviting, probing, and so forth. Willis and Willis (2007) also take it as a means for comprehension, confirmation, and clarification checks while verbally interacting with others.

The related literature is replete with studies scrutinizing the benefits of questioning in the academia (e.g., Almeida, 2012; Chudinova, 2020; Cotton, 2004; Rezvani & Sayyadi, 2015). Cotton (2004), for example, argued in his study that learners' questions can generate interest in new subjects, ideas, and challenges and induce them to be reflective about their ideas. Almeida (2012) also discussed the role of questioning in learners' meaningful learning and concluded that questioning reveals learners' quality of conceptual understanding and aids them to merge their prior and new knowledge in order to make sense of ideas. However, few studies, if any, have examined the benefits of questioning in other academic settings (e.g., various higher education contexts) as experienced and perceived by actual language users. Nor have any studies explored the practice and development of questioning in relation to the characteristics and demands of a test. The current study, hence, was aimed to probe the usefulness of students' questioning in different layers of higher education centers in ESCs as directly experienced by NNGs and perceived by Iranian TOEFL iBT instructors.

TOEFL iBT speaking test

The conceptualization of the TOEFL iBT speaking test is in accord with an expanded version of Bachman's (1990) model of communicative language ability and Canale and Swain's (1980) model of communicative competence tapping one's linguistic knowledge (i.e., syntactic, textual, and sociolinguistic knowledge), strategic competence, and the contextual use of language (Brooks & Swain, 2014). The groundwork of TOEFL iBT development was laid by studies (e.g., Rosenfeld et al., 2001; Taylor and Angelis, 2008a, b) which reflect the most essential skills for academic success in English-medium universities including the capacity to summarize, combine, and convey important information in lectures and other academic events accurately and coherently through the contextual use of language (Biber & Gray, 2013; Enright & Tyson, 2008; Frost et al., 2020).

The TOEFL iBt speaking test has been subject to ongoing validation studies. Some studies (e.g., Bridgeman et al., 2012; Cotos & Chung, 2018; Cumming et al., 2005; Enright & Tyson, 2008; Ockey et al., 2015) have examined the alignment between the test and TLUS tasks and reported TOEFL iBT to be a valid measure of communicative ability owing to setting up an actual situation requiring applicants to demonstrate their capacity to perform various TLUS tasks in the test. Cotos and Chung (2018) sought to explore to what extent the language functions elicited by TOEFL iBT speaking tasks are in line with those fulfilled by international teaching assistants in the authentic discourse. They concluded that "TOEFL iBT speaking tasks elicit most of the language functions identified in the discourse, suggesting that this test accounts for the functional language ability necessary for effective instructional performance as a teaching assistant in the target domain of language use" (p.1).

On the other hand, some studies raised concern about what underlying ability it measures (Johnson, 2001; Kyle et al., 2016). For example, Kyle et al. (2016) explored the construct validity of independent and integrated TOEFL iBT speaking tasks and found that "although the independent tasks included in the TOEFL iBT may represent a single construct, responses to integrated tasks vary across task sub-type" (p.1). Such findings would call oral interviews' construct validity into question arguably because its design might hinder score interpretation and extrapolation beyond the testing context (Johnson, 2001). Hence, it seems that no consensus on the authenticity and construct validity of TOEFL iBT speaking tasks has been reached in the related literature, which induced the present researchers to carry out the present study in order to enhance the body of knowledge in this regard.

TOEFL iBT preparation courses

Most prior studies on TOEFL iBT preparation courses are concerned with applicants' and teachers' perceptions of the test structure and usefulness of the preparation courses (e.g., Daneshvar et al., 2020; Masfufah, 2018; Ma & Cheng, 2015), the effects of the preparation courses on the test performance of learners (see Liu, 2014), and the washback effects of the test on the courses (Barnes, 2017; Erfani, 2012). Barnes (2017), for instance, studied how teaching to the test can change the notion of *a good teacher* and found that general English teachers, as a washback effect of the test, need to demonstrate varying qualities when teaching to TOEFL iBT test. Erfani (2012) also sought to compare the washback effects of TOEFL iBT and IELTS on teaching and learning activities and concluded that while the TOEFL iBT preparation courses were found to include a wider range or academic activities, the IELTS ones attempted to tap learners' communicative capacities more noticeably.

Given the review of the related literature, it seems that the washback effect of the absence of assessing applicants' questioning in TOEFL iBT test on the pedagogic and learning behaviors in preparation courses has been unexplored. Accordingly, the present study sought to probe how applicants' questioning skill is treated by teachers in the preparation courses.

Research questions

This study addresses the following questions:

1. What are the perceptions of NNGs from universities in ESCs about the significance of students' questioning skill in English-medium higher education settings?
2. What are the perceptions of Iranian TOEFL iBT instructors about the significance of students' questioning skill in English-medium higher education settings?
3. How is Iranian applicants' questioning skill perceived and treated by the instructor participants in their TOEFL iBT speaking preparation courses?

Method

Given the nature of the research questions, a qualitative research method was used in the present study to probe the participants' perceptions about the significance of NNG's questioning skill as well as the practice and development of applicants' questioning skill in TOEFL iBT speaking preparation courses.

Data collection instruments

The data of the study were collected through in-depth semi-structured interviews with the participants. To develop the interview questions, twelve and six interview questions were generated, in the first step, for the responding iBT instructors and NNGs, respectively. The questions were formulated after a review of the related literature. When the questions were developed, three TEFL experts from Shiraz University, Iran, were requested to review the interview questions in terms of clarity, relevance, and comprehensibility. Based on their comments, some of the questions were revised, merged, or omitted. The final draft of the interview questions, including, respectively, 7 and 3 questions for TOEFL iBT instructors and NNGs, was piloted with a small sample of two TOEFL iBT instructors and two NNGs from Shiraz. An analysis of their answers along with their follow-up comments on the questions leads to the reformulation of one of the interview questions of both groups of the respondents.

In an attempt to triangulate the collected data to investigate TOEFL iBT instructors' practical approaches towards applicants' OEQs in preparation classes, further data were also sought through detailed field notes and voice records on the related in-class practices and interactions.

Sample and sampling procedures

To select a sample of NNGs who had experienced the communicative demands in English-medium academia, a purposive sampling procedure was utilized. Due to inaccessibility of NNGs from English-medium universities in the study context, the researchers looked for participants through social academic networks such as LinkedIn and Academia. The NNGs selected for the study were required to have completed at least one higher-educational level in ESCs. The participants identified with the features wanted were sent an invitation to participate in the study along with a brief description of the study. Those who accepted to participate (16 out of 150 cases) were requested to share their demographic information and leave their telephone number or email address for the interview. Eventually, 12 international NNGs (7 males and 5 females, with varying ages, nationalities, and fields of study) were interviewed before the saturation of data (see Table 1 for more information).

Table 1 Demographic information about the non-native graduates

No	Name[a]	Age	Gender	Nationality	Field of study (level)	University attended (country)
1	Abbas	25	Male	Lebanese	Finance (B. A.)	McGill (Canada)
2	Himmat	34	Male	Indian	Biochemistry (Ph. D.)	Alberta (Canada)
3	Omar	42	Male	Syrian	Oncology (M. Sc.)	Newcastle (England)
4	Aziz	33	Male	Iraqi	Applied Linguistics (Ph. D.)	Newcastle (England)
5	Ata	35	Male	Turkish	Archeology (M. A.)	York (England)
6	Arash	44	Male	Persian	Marketing (M. A. & Ph. D.)	Lancaster (England)
7	Abram	38	Male	Russian	Counseling (M. A.)	Adams State (The USA)
8	Adalinda	29	Female	German	Journalism (M. A.)	Oregon (The USA)
9	Hadis	35	Female	Persian	Marketing (Ph. D.)	Stanford (The USA)
10	Bojing	30	Female	Chinese	Accounting (M. Sc.)	Oakland (The USA)
11	Adeola	30	Female	Nigerian	Microbiology (M. Sc.)	Lincoln (New Zealand)
12	Julia	38	Female	Swedish	Public Policy (M. A.)	Waikato (New Zealand)

[a]Note: The names are fictitious.

As for TOEFL iBT instructors and speaking preparation courses, the study was conducted in Tehran and Shiraz, two of the largest and most populated Iranian metropolises with a comparatively large number of language learning centers offering TOEFL iBT preparation courses. To select the intended TOEFL iBT instructors, with at least 5 years of experience in teaching speaking preparation courses, a snowball or chain sampling method was employed. More specifically, the researchers initially referred to one reputable language institute in each city and interviewed one TOEFL iBT instructor. The initial respondents were then requested to suggest, possibly, some similar TOEFL iBT instructors to be interviewed. Having analyzed the data collected from the first two instructors, the researchers contacted and interviewed the new cases suggested. The next respondents, once interviewed, were also requested to introduce some other analogous cases. This procedure continued up to reaching the state of data saturation and coherence. In total, the viewpoints of 19 TOEFL iBT instructors (13 males and 6 females) with their ages ranging from 32 to 45 were elicited (see Table 2).

To investigate how oral questioning is taught and practiced in real classes, two classes of the respondents (one from each city) were observed. To elicit natural class interactions, the researchers requested instructors for permission to observe their TOEFL iBT preparation classes before they are affected by the interviews. The first two instructors who consented to both class observation and follow-up interviews were Hamed (from Tehran) and Peyman (from Shiraz). Therefore, one class from each instructor was observed. The observed TOEFL iBT speaking preparation classes in Tehran and Shiraz included 10 (6 males and 4 females) and 14 (9 males and 5 females) students, respectively.

Data collection procedure

The demographic information of the interviewees was sought before they participated in the study. The respondents were assured of the confidentiality of their identity and the information exposed. All interviews were conducted in English. The interviews with the NNGs, taking about 10 min, were carried out through Skype and voice calls. The

Table 2 Demographic information about the instructors

No	Name[a]	Age	Gender	Teaching experience	City
1	Mahnaz	45	Female	11 years	Tehran
2	Hossein	43	Male	8 years	Tehran
3	Hamed	42	Male	10 years	Tehran
4	Mehran	39	Male	6 years	Tehran
5	Nader	34	Male	5 years	Tehran
6	Amin	43	Male	6 years	Tehran
7	Mohammad	38	Male	6 years	Tehran
8	Sara	32	Female	5 years	Tehran
9	Salman	44	Male	10 years	Tehran
10	Ahmad	39	male	7 years	Tehran
11	Elahe	36	Female	6 years	Shiraz
12	Maryam	38	Female	5 years	Shiraz
13	Javad	35	Male	5 years	Shiraz
14	Nazanin	35	Female	5 years	Shiraz
15	Payam	36	Male	7 years	Shiraz
16	Peyman	40	Male	6 years	Shiraz
17	Sadegh	43	Male	5 years	Shiraz
18	Ali	39	Male	7 years	Shiraz
19	Soraya	36	Female	6 years	Shiraz

aNote: The names are fictitious.

TOEFL iBT instructors' insights were elicited through face-to-face interviews, each lasting for about 25 min on average.

To ensure the trustworthiness of the data, care was exercised to avoid bias through the application of the strategies suggested by Schumacher and McMillan (2006). More specifically, we employed a persistent fieldwork, accounted for participants' language verbatim accounts, and checked the data informally with the participants for accuracy after the interviews and also after data analysis.

Besides, to explore how the applicants' OEQs is noted and worked on in preparation classes, one of the researchers observed two TOEFL iBT speaking preparation classes (Hamed's and Peyman's) each for 8 sessions. Anything in connection to the instructors' or students' questions and questioning was observed and recorded. Field notes were also taken when an observation needed further attention for later analysis. As recommended by Bogdan and Biklen (1998), care was taken to be unnoticeable in note-taking not to catch the class participants' attention.

Data analysis procedure

To analyze the data accumulated through classroom observations along with interviews, content analysis was employed as suggested by Glaser and Strauss (1967). The analysis was iterative throughout the interviews and after the observations. To carry out a constant content analysis, both data sources were transcribed verbatim, integrated with notes taken. The transcriptions were then read frequently and recursively so that the class oral interactions could be envisaged in detail. This also helped to find connections

among what we observed and the interviews. Both researchers developed open codes independently concerning the research questions, which sometimes entailed going back and forth through the data. The analyses were then compared, and a few areas of disagreement were found. These were resolved through discussion, and subsequent analysis by the first researcher generated emergent recurrent themes. Through more refined cross-referencing between the themes and memos and participants' accounts categories and relationships among the themes were developed.

Results

The non-native graduates' perceptions

The interview questions encouraged the NNGs to discuss the advantages of questioning to deal with various issues in and outside English-medium classrooms. One of the main advantages reported for NNSs' questions in classrooms was that they might enhance the quality students' lesson learning. Specifically, 10 interviewees believed that questioning might dispel students' confusions about lessons and, hence, help them learn the new lessons more efficiently. As a case in point, Hadis experienced questioning as a very common practice in the courses she took because "professors [in her university] did not teach the lessons point by point and asked students to study the books and ask questions about their problems". Students' questions in classrooms, as assumed by Bojing and Himmat, can also foster the classroom interaction between teachers and students and among students, giving rise to "a competitive environment that helps students to grow" (Himmat) and, hence, "improving the quality of learning" (Bojing).

Further, 3 respondents noted that students' questions may enhance learning with and from peers. Adeola, for instance, held that "when students ask questions in classrooms, they not only improve their level of learning but also help their classmates indirectly to get rid of their likely confusions".

Students' questions, as pointed out by 4 respondents, may raise untouched subjects to learn because effective questions, as held by Julia, "can lead to doubt and reveal many unknown facts, and when somebody can unveil new facts, then he is successful in education". Similarly, Omar stated that questioning would "aid students to learn new lessons in classrooms and also help them get more insightful about each lesson".

The interviews also made the respondents recall how they had benefited from their questions outside classrooms in English-medium academic settings. Three respondents pointed to a growing reliance on questioning outside classrooms because, as Abram argued, "NNSs' academic survival and especially settlement were partially dependent upon their questioning". Similarly, the respondents recollected how they had employed questions to tackle their confusions and problems among others about, for example, registration processes (6 respondents), finding locations (12 respondents), class, conference, or workshop times (6 respondents), assignments (7 respondents), and university rules and norms (5 respondents).

When commenting on NNSs' questioning ability, epithets including *important* (by 7 respondents), *significant* (by 5 respondents), *determining* (1 respondent), and *vital* (1 respondent) were recurrently used. These were also suggestive of the considerable criticality attached to questioning skill in universities in ESCs. Table 3 summarizes the major themes and sub-themes of schemes emerging from the perceptions of the NNGs.

Table 3 NNGs' perceptions of non-native students' questioning in academia

Issue	Theme	Sub-themes
Students' questions in the academic context	Significance	• Perceived to be high • Impacts on quality of learning through a. Eliminating students' confusions b. Fostering classroom interaction c. Peer learning
	Benefits in classrooms	• Impacts on the quantity of learning through a. Raising untouched subjects
	Benefits outside classrooms	• Eliminating confusions about a. Registration processes b. Finding locations c. Class, conference, or workshop times d. Assignments e. University rules and norms

The TOEFL iBT instructors' perceptions

The results emerging from the analysis of the TOEFL iBT instructors' views on the significance of NNSs' questions in academia showed multiple practical benefits for the students. NNSs are likely to be confused when new lessons are taught in English, and nothing better than asking questions can reveal and dispel their confusion because, as Soraya argued, "the depth of learners' questions can clearly show their depth of thinking and understanding".

Campus life is full of challenges, particularly for newcomers. Most respondents pointed to unfamiliarity with campus rules, sociocultural norms, and specific curricular and extracurricular programs as some noticeable issues or sometimes headaches NNSs are faced with. All these normally entail "asking for help or information" as commented by one of the respondents.

The respondent instructors also discussed the role of questioning in peer interactions. NNSs' interactions with other students, which might be typically initiated through questioning, might promote their learning, reduce their likely culture shock, and create rapport. Questions for peers were perceived by two instructors to be even more effective than those for teachers because NNSs, especially the less confident ones, normally tend to put their questions to their readily available peers rather than dominant teachers.

Four instructors extended their views beyond the classroom context and highlighted the importance of students' questions in seminars and scientific meetings. They argued that the questions that the students raise in seminars may "involve them in scientific discussions with teachers, researchers, and students" (Payam) and "make them more sophisticated because each question might create several other questions and issues for discussion in students' minds" (Peyman). In short, the analysis of the TOEFL iBT instructors' perceptions of learners' questions in academic discourse suggested the themes and sub-themes summarized in Table 4.

TOEFL iBT applicants' questioning skill

Having been informed of the TOEFL iBT instructors' perspectives on the significance of NNSs' questioning skill in academia, we sought their views on the importance of students' questions and their acquisition in TOEFL iBT speaking preparation courses. One

Table 4 TOEFL iBT instructors' perceptions of non-native students' questions in academia

Issue	Theme	Sub-themes
NNSs' questions in the academic context	Significance	• Perceived to be high
	Benefits	• Eliminating students' conceptual understanding and confusion • Eliminating students' sociocultural confusions • Peer interaction • Enhancing students' insightfulness

main theme emerging from the interviews was that despite their acknowledgment of the criticality of NNSs' questioning skill, the instructors had never attached much weight to TOEFL iBT students' questions and considered no systematic plan to improve their students' questioning skill.

Specifically, 15 instructors argued plainly that there would be no point in developing students' questioning when it is of no use in taking the test. They strongly emphasized that they are mainly supposed to prepare applicants for a test with clear requirements and tasks, which are not designed to estimate applicants' proficiency in posing English questions.

Practice and assessment of questioning were perceived by Peyman, Mahnaz, and Ali to involve reciprocal or "face-to-face" interactions in which "there is a listener only to receive and respond to the questions" (stated by Peyman). The target test was viewed to be far from being reciprocal, for, as Ali argued, it "occurs only between an applicant and a computer" which is solely tasked with "administering the test and recording and transmitting the provided responses to the scoring center" rather than with "carrying out online human-like interactions with applicants". Accordingly, Ali contended that "you should be silly to ask a machine questions when you know that the machine is not going to provide you with no [sic] feedback or response". Ali concluded that it would be "unwise" to involve his students in a scenario that is not going to be performed on the scene.

The instructors stated that they had never thought of planning and implementing classroom tasks for promoting applicants' questioning skills. Their strong emphasis on the futility of focusing on applicants' questioning skill in such preparation classes encouraged us to ask whether neglecting students' questioning skill might make any troubles for them in TLUS. Interestingly, ten instructors believed that this is likely to impact negatively on their future actual communications in social and academic contexts. However, the instructors had to stick to test-oriented classroom tasks and consequently neglect the development of applicants' questioning skills because "both instructors and applicants are evaluated based on the applicants' test performance", Hamed commented. Nazanin also asserted that "candidates take the preparation classes with high expectations because they spend much money on them. In this stage, getting the needed score is the only important thing to them. They just expect us to prepare them for the test".

On the other hand, Elahe, Nader, Sara, and Sadegh (out of the 19 instructors) were the only TOEFL iBT instructors who claimed to be concerned about their students' questioning and made conscious attempts to develop it. All in all, the analysis of the TOEFL iBT instructors' views gave rise to the following themes and sub-themes (see Table 5).

Table 5 Major themes and sub-themes about the TOEFL iBT instructors' perceptions of learners' questions TOEFL iBT speaking preparation courses

Issue	Theme	Sub-themes
Applicants' practice of oral questioning in TOEFL iBT speaking preparation courses	Significance	• Mainly downplayed as a washback effect of the test
	Development	• Mainly neglected for 1. The test's construction 2. Applicants' reluctance 3. Applicants' expectations 4. Instructors' occupational reputation

The observed TOEFL iBT classrooms

To build up a more accurate picture of the applicants' practice of questioning and in an attempt to triangulate the collected data, two TOEFL iBT speaking preparation courses (each for eight sessions) were observed. Table 6 provides descriptive statistics for oral English questions posed by the applicants in each session.

Table 6 illustrates that questioning was not very often practiced by the applicants, in both courses. Specifically, the number of English questions raised by learners in each of the observed courses did not exceed two, on average, per session. Noticeably no English question was asked in C_1 and C_2 for three and five consecutive sessions, respectively. Besides, twelve out of the thirteen questions asked by the applicants in C_2 were phrased in two sessions when they were supposed to perform certain role plays entailing asking questions. The single remaining question was asked by an applicant for the meaning of an unknown word.

A further surprising point observed in C_1 was that nine out of the twelve oral questions asked were raised by only two applicants. Three applicants asked the rest of the questions (each only one), and the other ten applicants never happened to ask any questions. C_2, nonetheless, demonstrated a more balanced case because English questions were raised, though scantly, by various applicants. That was mainly because of the instructor's tendency to assign role plays and group works to various pairs and groups. In both classes, most questions seemed to be ungrammatical or incomprehensible, which made the instructors or partners ask for clarification mainly by gestures, facial expressions, and sometime a single-word question like "Why?". As a case in point, the following dialogue from C_2 reveals the poor questioning skill of an applicant who, together with his peers, was supposed to prepare a detailed summary of an academic speech presented to them as a listening task:

Applicant A: He said the FAO will discuss the effects of food prices, limited resources, and something else that I did not get.

Applicant B: Climate change and increased energy needs.

Applicant A: You say climate change and the other one I didn't understand?

Applicant B: What?

Table 6 Descriptive statistics for the TOEFL iBT applicants' oral questions

Course	Minimum	Maximum	Total in 8 sessions	Mean	Std. deviation
Applicants in *C_1	0	4	12	1.50	1.604
Applicants in **C_2	0	7	13	1.63	2.774

Note: C_1 TOEFL iBT speaking preparation course in Shiraz
C_2 TOEFL iBT speaking preparation course in Tehran

Instructor: What do you mean?

Applicant A: uuuuh, I mean, [code-switching] mænzuræm Ine ke *'climate change'* væ unyekI ke goftæn hæmun t izai bud ke mæn næfæhmIdæm? [I mean, are *climate change* and the other thing he mentioned the points I didn't get?]

Instructor: Ok. Yes, they were.

In the above dialogue, although the instructor had always insisted on using English in class talking, Applicant *A*'s incomprehensible question and his subsequent failure to revise his question left him with no choice but to restate the intended question in Persian.

When interacting with their instructors or peers, the applicants of both classes posed Persian rather than English questions for permission, breaks, further explanations, statement repetition, and class time modification.

In C_2, there were occasions that the students were required to perform role plays with little practice or rehearsal. Such activities appeared to make the students anxious. When it came to moments that normally demanded asking a question to go on, they hesitantly waited for their partners or teachers to help. This led to long and sometimes frustrating delays and eventually structurally poor questions or code switching.

Even a single attempt by the instructors was not observed in these two classes to encourage the applicants to enhance their ability to ask questions. Rather, all too often, they turned a blind eye to their students' frequent use of their mother tongue to pose questions. This is the case though we noted that they were sensitive about their students' use of Persian when responding to questions, which is revealing in the following excerpt from C_1:

Instructor: What were you supposed to do for this session?

Applicant: fekr konæm [I think] [interrupted immediately by the instructor]

Instructor: In English, please.

Discussion

The findings of the study are discussed from multiple perspectives with respect to the significance of questioning skill in academia and accordingly how excluding it from the TOEFL iBT speaking test might impact the validity and washback of the test.

Questioning as a significant target task practice

To answer the first and second research questions, the significance and benefits of NNSs' oral questioning skills in English-medium academic settings were examined based on NNGs' and Iranian TOEFL iBT instructors' perceptions. As the results of the study indicated, both groups attributed a high degree of criticality to questioning and regarded it as an essential communicative skill in academic settings. Specifically, the NNGs reported that they had benefitted from questioning in and outside classrooms in order to enhance the quality and quantity of mastery of a particular skill or course lesson in ESCs through eliminating their confusions, fostering classroom interactions and peer learning, and raising untouched course subjects, among others. These benefits, to a large extent, were also perceived and pointed out by the TOEFL iBT instructors. Such results support several related studies reporting similar findings regarding the benefits of questioning in the academia such as showing the students' level of

language knowledge and proficiency (Chudinova, 2020; Rezvani & Sayyadi, 2015), letting them benefit from their peers' varied explanations (Almeida, 2012), inducing them to be reflective about their ideas (Cotton, 2004), and helping them to comprehend and subsume new knowledge into knowledge already acquired (Robinson & Song, 2019).

Applicants' questioning in preparation courses

The study was also an attempt to examine how applicants' questioning skill (i.e., the third research question) is treated in preparation courses given its absence in TOEFL iBT speaking test. The results showed that the teacher participants of the study saw no point in expending effort in developing the applicants' questioning skill, since it was not conducive to their test results, and the applicants themselves were not normally eager to work on skills not assessed by the test. Accordingly, having been aware of the discrepancy, the instructor participants restricted their mission to teaching to the test tasks rather than TLUS tasks, which is clearly because of the influence of testing on teaching and learning or overt washback (Dong et al., 2021; Prodromou, 1995; Xie & Andrews, 2012). A similar conclusion was reached by Nikolayev (2016) who argued that teachers normally tend to teach solely to the test in test preparation courses because it "would allow the students to get accustomed to the test format and thus be fully aware of what to expect on the test day" (p.97).

Another finding of note is that although the instructors were acutely aware of the significance of questioning as well as the adverse consequences of neglecting its development on applicants' future academic lives, they exclusively taught to the test and did not care about the demands of General English courses because their reputation and income heavily depended on their students' test results. This result ties well with Hawkey's study (2006) in which expectations of students aiming to get favorable test results along with those of institutes seeking reputation were found to constrain their teaching to IELTS as a high-stakes test.

Contrary to the findings of Munoz and Alvarez (2010) who found language testing to be beneficial in developing authentic classroom communication as well as those of Enright and Tyson (2008) who considered TOEFL iBT to be proactive in encouraging communicatively oriented pedagogic classroom activities resembling those in academic situations, the results of the current study unveiled that the interactions in TOEFL iBT preparation classes were not completely consistent with real-life situations since the learners were scantly observed to initiate English interactions through, for example, questioning. This study also suggested that the discourse in preparation courses is not fully in accord with the ideal language learning classrooms envisaged by Powell and Powell (2010) and Stokhof et al. (2017) who argued that classrooms should construct an authentic communicative environment which mirrors the linguistic complexities and ambiguities of real life and guides the learners to extend their language use from classroom to real-life situations.

Concerns about TOEFL iBT speaking test

Neglecting the assessment of applicants' questioning skill may give rise to critical concerns about the validity and authenticity of TOEFL iBT speaking test. An attempt is made to discuss the findings and, more specifically, examine the validity of the test as

long as the assessment of questioning skill is concerned based on Bachman and Palmer's (1996), Messick's (1998), and Kane's (2013) validation approaches.

Bachman and Palmer (1996) in their *checklist approach to evaluating a test* argued that for a test to be valid, it is supposed to tap into the *ideational*, *manipulative*, *heuristic*, and *imaginative* functions of language. They further asserted that a well-constructed test needs to sample tasks consistent with those in TLUS, or otherwise, its authenticity and construct validity might be questionable. On the other hand, as the results of the study indicated, TOEFL iBT speaking test fails to measure its applicants' skill in performing, at the very least, the heuristic function of language which is typically carried out through questioning (Thwaite, 2019). From Bachman and Palmer's perspective, negligence of applicants' questioning skill, in other words, would also point out traces of inconsistency between the test tasks and those in TLUS, threatening the authenticity, content validity, generalizability, construct validity, and, as a result, the usefulness of the tests.

Falling short of assessing questioning skill as an essential component of social and academic interactions would also lead to *construct underrepresentation*, because of "missing something relevant to the focal construct that, if present, would have permitted the affected examinees to display their competence" (Messick, 1998, p.11). Messick (1998) outlines construct underrepresentation along with construct irrelevance as the two general threats to the validity of a test. Accordingly, missing applicants' questioning skills, as an indispensable part of interactions in TLUS, might undermine the construct validity and limit the score interpretations of the TOEFL iBT speaking test.

An alternative perspective to test validation was posited by Kane (2013) who views validation as the process of putting forward a chain of argument-based propositions about scoring, generalization, representativeness, extrapolation, and implications of a test and providing evidence for the plausibility, completeness, and coherence of the propositions. Building on this approach to examine the validity of TOEFL iBT, Enright and Tyson (2008, p.3) put forth the following propositions to be evidenced by reviews of research and empirical studies:

1. The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.
2. Tasks and scoring criteria are appropriate for obtaining evidence of test-takers' academic language abilities.

Regarding the first preposition, this study concluded that TOEFL iBT tasks are not fully reprehensive of TLUS tasks because applicants are not tasked by the test to raise questions. In actual academic settings, however, students may come up with and pose miscellaneous questions while listening to, taking note of, summarizing, or discussing what is presented in lectures. A similar pattern of results was obtained by Brooks and Swain (2014) who compared the oral performance of 30 TOEFL iBT test-takers in the test and real academic situations and documented solely one single question raised by one of the applicants in the testing situation in contrast to a significantly larger number of questions in actual academic settings. The findings, nevertheless, are not in accord with those revealed by Cumming et al. (2005) reporting the speaking test tasks to be realistic and appropriate simulations of how students speak in academic contexts.

Regarding the second proposition, it was unveiled that the computerized design of the test tasks keeps applicants passive with no chance to initiate a conversation for example by asking a question or to have a reciprocal talk with their interlocutors. Missing in the test is a *responding addressee* as a requisite for carrying out a *reciprocal* and *life-like* interaction (Oliver & Azkarai, 2019) involving, by necessity, asking questions. As assessment of applicants' ability to use the English language orally in academic settings is limited to examining their responding abilities, it can be argued that inappropriate or, at least, insufficient scoring criteria have been assumed for TOEFL iBT speaking test.

Conclusion

The current study sought to investigate the significance and usefulness of oral questioning for NNSs in English-medium academia from the viewpoints of NNGs and Iranian TOEFL iBT instructors. Further, it examined how TOEFL iBT instructors treated their students' oral questioning skills in speaking preparation courses. The results indicated that both NNGs and instructors assumed questioning as an indispensable aspect of academic interactions yielding various benefits for NNSs including the elimination of conceptual and sociocultural confusions, new knowledge acquisition, promotion of learning quality, and insightfulness and peer learning enhancement.

The interviews and observations unveiled that the TOEFL iBT instructors despite acknowledging the criticality of posing OEQs in academia refrained from working on their students' questioning skill due to the absence of the skill in the test, unwillingness of students to develop test-irrelevant skills, limited preparation time, and dependency of their professional reputation upon their students' test rather than target situation performance. In short, the instructors were preparing their students solely for the TOEFL iBT tasks and disregarded significant demands in the TLUS requiring their questioning.

Implications of the findings

Students' questioning skill has been documented (e.g., Almeida, 2012; Graesser & Person, 1994; Rezvani & Sayyadi, 2015) as the Cinderella of second/foreign language courses. Therefore, drawing on research on such a crucial skill might suggest fresh directions for those directly or indirectly concerned about it. The results of this study might have important implications not only for language instructors and learners but also for test developers and test users.

More specifically, the findings of the study might raise language instructors' awareness of the significance and key role of questioning in academic life. As language instructors are normally supposed to set up classroom conditions aligned with real-life situations (Gardner, 2019), the findings of this study might encourage them to devise more systematic plans to dedicate a certain part of their class time to getting their students to practice questioning through, for instance, more reciprocal and life-like interactions in classrooms. Further, language teachers might be urged not to evaluate their students' oral proficiency based only on the quality of the students' responses to their questions. Rather, they might require their students, as a part of their oral exams, to ask their teachers and/or peers. In doing so, the teachers might be encouraged to rely on more authentic oral tasks such as role plays which typically demand the application

of reciprocal communicative skills on the side of the examinees. In other words, to provoke positive washback of the classroom tests, it is also advisable to assess their speaking capacity in part based on the efficiency, relevance, and accuracy of their questions.

The misalignment between the test and TLUS tasks, as indicated in this study, might encourage policy-makers and TOEFL iBT speaking test developers to characterize language ability more inclusively and in close correlation with TLUS. More specifically, use might be made of tasks assessing applicants' both responding and questioning skills in TOEFL iBT. This certainly enhances the test authenticity, content validity, and ultimately construct validity or *test usefulness in* Bachman and Palmer's (1996) terminology. Further, the findings might induce researchers to develop new rating scales for speaking assessment tapping test-takers' life-like language skills including questioning. Theoreticians might also develop more comprehensive frameworks of language competence representing the skills required to carry out various language functions including the heuristic one.

Limitations and suggestions for further research

One of the limitations of the current study, though practically formidable and totally common in qualitative research, was that the findings were drawn based on the viewpoints of a relatively small sample of respondents, which prevented the generalizability of the findings. Further, the observation of only a couple of TOEFL iBT speaking preparation courses should be acknowledged as another limitation hindering the study from drawing a more detailed and conclusive picture of the way applicants' questioning skill is actually treated by instructors and applicants in these courses.

Similar studies perhaps with larger samples and with varied characteristics such as age, gender, first language background and language proficiency can provide a more comprehensive perspective on the insights and practical tendencies of TOEFL iBT instructors and applicants towards questioning skill. It is also suggested that the same or similar research problems be addressed through the use of alternative data sources and research methodologies like quantitative and mixed-design methods.

Abbreviations
ESC: English-speaking countries; iBT: Internet-based test; NNS: Non-native student; OEQ: Oral english question; TEFL: Teaching english as a foreign language; TLUS: Target language use situation; TOEFL: Test of english as a foreign language

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]University of Tehran, Tehran, Iran. [2]Department of English Language, Yasouj University, Yasouj, Iran.

References

Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing*, *26*(4), 621–631. https://doi.org/10.1177/0265532209346371.

Ali, M. M., & Hamid, M. O. (2020). Teaching English to the test: Why does negative washback exist within secondary education in Bangladesh? *Language Assessment Quarterly*, *17*(2), 129–146. https://doi.org/10.1080/15434303.2020.1717495.

Almeida, P. A. (2012). Can I ask a question? The importance of classroom questioning. *Procedia-Social and Behavioral Sciences*, *31*(1), 634–638. https://doi.org/10.1016/j.sbspro.2011.12.116.

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

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

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis (TOEFL iBT Research Report-19)*. Educational Testing Service.

Bogdan, R., & Biklen, S. (1998). *Introduction to qualitative research in education*. Pearson.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, *29*(1), 91–108. https://doi.org/10.1177/0265532211411078.

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real life academic speaking activities. *Language Assessment Quarterly*, *11*(4), 353–373. https://doi.org/10.1080/15434303.2014.947532.

Brown, A., & Ducasse, A. M. (2019). An equal challenge? Comparing TOEFL iBT™ speaking tasks with academic speaking tasks. *Language Assessment Quarterly*, *16*(2), 253–270. https://doi.org/10.1080/15434303.2019.1628240.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/1.1.1.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. SAGE Publications.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language (pp.1-25)*. Routledge.

Chudinova, E. V. (2020). On the significance of questioning in learning activity. *Psychological Science and Education*, *25*(4), 60–70. https://doi.org/10.17759/pse.2020250406.

Cotos, E., & Chung, Y. R. (2018). Domain description: Validating the interpretation of the TOEFL iBT® speaking scores for international teaching assistant screening and certification purposes. *ETS Research Report Series*, *2018*(1), 1–24. https://doi.org/10.1002/ets2.12233.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL test (TOEFL Monograph No. MS-26)*. Educational Testing Service.

Daneshvar, A., Bagheri, M. S., Sadighi, F., Yarmohammadi, L., & Yamini, M. (2020). Iranian TOEFL iBT and the IELTS teachers' views on the structure of the TOEFL iBT and IELTS receptive and productive sections in terms of dynamic and static assessment. *Journal of Language and Translation*, *10*(4), 1–16.

Dong, M., Fan, J., & Xu, J. (2021). Differential washback effects of a high-stakes test on students' English learning process: Evidence from a large-scale stratified survey in China. *Asia Pacific Journal of Education*, *41*(1), 1–18.

Dool, M. A., Mughal, S. H., & Rind, A. A. (2020). Usage of effective questioning to enhance students' level of participation in a grade XI chemistry classroom of a government degree college in rural district of Jacobabad. *sjesr*, *3*(3), 222–231.

Enright, M., & Tyson, E. (2008). *Validity evidence supporting the interpretation and use of TOEFL iBT scores*. Educational Testing Service.

Erfani, S. S. (2012). A comparative washback study of IELTS and TOEFL iBT on teaching and learning activities in preparation courses in the Iranian context. *English Language Teaching*, *5*(8), 185–195.

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test-takers' use of stimulus content. *Language Testing*, *37*(1), 133–155. https://doi.org/10.1177/0265532219860750.

Gardner, R. (2019). Classroom interaction research: The state of the art. *Research on language and social interaction*, *52*(3), 212–226. https://doi.org/10.1080/08351813.2019.1631037.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge University Press.

Hung, S. T. A., & Huang, H. T. D. (2019). Standardized proficiency tests in a campus-wide English curriculum: A washback study. *Language Testing in Asia*, *9*(1), 1–17. https://doi.org/10.1186/s40468-019-0096-5.

Kane, M. (2013). Articulating a validity argument. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing*, (pp. 48–61). Routledge.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319–340. https://doi.org/10.1177/0265532215587391.

Liu, O. L. (2014). Investigating the relationship between test preparation and TOEFL iBT® performance. *ETS Research Report Series*, *2014*(2), 1–13.

Ma, J., & Cheng, L. (2015). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth, and significance. *TESL Canada Journal*, *33*(1), 58–79. https://doi.org/10.18806/tesl.v33i1.1227.

Masfufah, S. (2018). Indonesian college students' perceptions of TOEFL preparation class. *EduLite: Journal of English Education, Literature and Culture*, *3*(1), 66–78. https://doi.org/10.30659/e.3.1.66-78.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1-3), 35–44. https://doi.org/10.1023/A:1006964925094.

Munoz, A. P., & Alvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language testing*, *27*(1), 33–49. https://doi.org/10.1177/0265532209347148.

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. https://doi.org/10.1177/0265532214538014.

Oliver, R., & Azkarai, A. (2019). Patterns of interaction and young ESL learners: What is the impact of proficiency and task type? *Language Teaching for Young Learners*, *1*(1), 82–102. https://doi.org/10.1075/ltyl.00006.oli.

Prodromou, L. (1995). The backwash effect: From testing to teaching. *ELT Journal*, *49*(1), 13–25. https://doi.org/10.1093/elt/49.1.13.

Rezvani, R., & Sayyadi, A. (2014). A qualitative study of washback effects of new Iranian TEFL Ph. D. program entrance exam on applicants' study plans and strategies: Insights from instructors and applicants. *Journal of Research in Applied Linguistics*, *5*(2), 113–127.

Rezvani, R., & Sayyadi, A. (2015). Instructors' and learners' questioning: A case of EFL classroom discourse in Iran. *Journal of Teaching Language Skills*, *34*(3), 141–164.

Rezvani, R., & Sayyadi, A. (2016). Washback effects of the new Iranian TEFL Ph. D. program entrance exam on EFL instructors' teaching methodology, class assessment, and syllabus design: A qualitative scrutiny. *Journal of instruction and evaluation*, *9*(33), 159–180.

Robinson, S. E., & Song, J. J. (2019). Student academic performance system: Quantitative approaches to evaluating and monitoring student progress. *International Journal of Quantitative Research in Education*, *4*(4), 332–353. https://doi.org/10.1504/IJQRE.2019.100170.

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*. Educational Testing Service.

Schumacher, S., & Mcmillan, J. (2006). *Research in education evidence–based inquiry*. Pearson Education.

Stokhof, H. J., De Vries, B., Martens, R. L., & Bastiaens, T. J. (2017). How to guide effective student questioning: A review of teacher guidance in primary education. *Review of Education*, *5*(2), 123–165. https://doi.org/10.1002/rev3.3089.

Taylor, C., & Angelis, P. (2008b). The evolution of the TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language*, (pp. 27–54). Routledge.

Taylor, C. A., & Angelis, P. (2008a). The evolution of the TOEFL. Building a validity argument for the Test of English as a Foreign Language, 27-54.

Thwaite, A. (2019). Halliday's view of child language learning: Has it been misinterpreted? *Australian Journal of Teacher Education (Online)*, *44*(5), 42–56. https://doi.org/10.14221/ajte.2018v44n5.3.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University press.

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing*, *30*(1), 49–70. https://doi.org/10.1177/0265532212442634.

Publisher's Note