12-2018

Domain Description: Validating the Interpretation of the TOEFL iBT® Speaking Scores for International Teaching Assistant Screening and Certification Purposes

Elena Cotos
Iowa State University, ecotos@iastate.edu

Yoo-Ree Chung
Yonsei University

Domain Description: Validating the Interpretation of the TOEFL iBT® Speaking Scores for International Teaching Assistant Screening and Certification Purposes

Abstract
In the past 2 decades, there has been an increasing tendency to use scores from the TOEFL iBT® Speaking test for decisions regarding the certification of international graduate students as teaching assistants at North American universities. To obtain validity evidence in support of the usefulness of the speaking scores for this secondary use of the test, this study adopted the argument-based approach to validation. Focusing on the domain description inference in the TOEFL interpretive argument, the study investigated whether the language functions elicited by TOEFL iBT Speaking tasks can be identified in authentic discourse produced by international teaching assistants (ITAs) with different instructional roles. We compiled and analyzed 2 corpora—a TOEFL iBT speech corpus and an ITA speech corpus. The TOEFL corpus contained 2,738 responses to integrated and independent tasks. The ITA corpus of 119 spoken texts included multiple disciplines and 3 instructional genres: lab, recitation, and lecture. The 2 corpora were manually annotated using the knowledge framework (Mohan, 1986), which is a heuristic in systemic functional linguistics (SFL) used to identify knowledge structures (KSs) and language functions based on how linguistic choices function in the discourse. Then, the following types of data were quantitatively analyzed: discourse units annotated per KS category, discourse units annotated per language function, KS categories occurring in each text, language functions occurring in each text, and KSs and functions from each component of the spoken corpora. The corpus data revealed how the language functions were realized and how they varied. Overall, the results indicated that TOEFL iBT Speaking tasks elicit most of the language functions identified in ITA 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. The discrepancy detected in the use of some functions pertaining to 2 KSs warrants further examination of the extent to which it may impact secondary test use and score interpretation.

Keywords
TOEFL iBT Speaking Section, Domain Description, Language Functions, Curriculum Genres, International Teaching Assistants

Disciplines
Bilingual, Multilingual, and Multicultural Education | Higher Education | Vocational Education | Vocational Rehabilitation Counseling

Comments
This accepted report is published as 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* (TOEFL Research ReportNo. RR-85). Princeton,NJ: Educational Testing Service. Doi: 10.1002/ets2.12233. Posted with permission.

Title: Domain Description: Validating the Interpretation of the *TOEFL iBT*® Test Speaking Scores for International Teaching Assistant Screening and Certification Purposes

Elena Cotos & Yoo-Ree Chung

ABSTRACT

In the past two decades, there has been an increasing tendency to use scores from the *TOEFL iBT*® speaking test for decisions regarding the certification of international graduate students as teaching assistants at North American universities. To obtain validity evidence in support of the usefulness of the speaking scores for this secondary use of the test, this study adopted the argument-based approach to validation. Focusing on the domain description inference in the TOEFL interpretive argument (Chapelle, Enright, & Jamieson, 2008), the study investigated whether the language functions elicited by TOEFL iBT speaking tasks can be identified in authentic discourse produced by international teaching assistants (ITAs) with different instructional roles. Two corpora were compiled and analyzed—a TOEFL iBT speech corpus and an ITA speech corpus. The TOEFL corpus contained 2,738 responses to integrated and independent tasks. The ITA corpus of 119 spoken texts included multiple disciplines and three instructional genres: lab, recitation, and lecture. The two corpora were manually annotated using the knowledge framework (Mohan, 1986), which is a heuristic in systemic functional linguistics (SFL) used to identify knowledge structures (KSs) and language functions based on how linguistic choices function in the discourse. Then, the following types of data were quantitatively analyzed: discourse units annotated per KS category, discourse units annotated per language function, KS categories occurring in each text, language functions occurring in each text, and KSs and functions from each component of the spoken corpora. The corpus data revealed how the language functions were realized and how they varied. Overall, the results indicate that TOEFL iBT speaking tasks elicit most of the language functions identified in ITA 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. The discrepancy detected in the use of some functions pertaining to two KSs warrants further examination of the extent to which it may impact secondary test use and score interpretation.

Keywords: *TOEFL iBT*® Speaking section; domain description; language functions; curriculum genres; international teaching assistants

The Internet-based *TOEFL iBT*® test is one of the most globally accepted English-language proficiency tests that aids in making high-stakes decisions for nonnative speakers of English. As stated in the TOEFL Framework document, the purpose of the test is to "measure examinees' English-language proficiency in situations and tasks reflective of university life" (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000, pp. 10–11). TOEFL iBT measures proficiency through test tasks that require the integration of receptive and productive skills, and the results

are interpreted as indicative of the test takers' ability to combine their listening, reading, speaking, and writing skills to perform academic tasks.

Institutions in more than 130 countries use TOEFL iBT scores primarily for admission purposes. Secondary uses of TOEFL iBT scores have also been suggested, for the "information derived from the proficiency levels may also be used in guiding English-language instruction, placement decisions, and awarding of certification" (Jamieson et al., 2000, p. 10). In the United States, there has been an increasing tendency to use TOEFL iBT speaking scores for decisions regarding the language certification of international teaching assistants (ITAs). The need to assess ITAs' oral language performance emerged in the 1970s when the growing number of ITAs raised serious concerns among undergraduate students and their parents about how the ITAs' insufficient English-language ability impacted the quality of undergraduate education (Bailey, 1983, 1984; Ruderman, 2000). In response to these concerns, many states developed legislative mandates requiring higher education institutions to assess ITAs' oral proficiency and to certify them before they enter the classroom (Brown, Fishman, & Jones, 1990; Dick & Robinson, 1994). At the same time, the law left it to the discretion of institutions as to how certification should take place.

Individual institutions, which use teaching assistants for different purposes and in different instructional contexts, have adopted different commercial or in-house assessments (T. L. Farnsworth, 2014). The most common assessments for ITA certification have been the Spoken Proficiency English Assessment Kit (SPEAK, an institutional version of the retired Test of Spoken English developed by the Educational Testing Service), oral interviews, and teaching simulation tests (Plakans & Abraham, 1990). Generally, the Test of Spoken English served to screen ITA candidates' speaking proficiency prior to their arrival on campus, and SPEAK was administered locally for on-site screening as a tentative judgment of suitability for ITA assignment.

Because ETS discontinued the SPEAK test, ITA programs nationwide have expressed interest in using the TOEFL iBT speaking scores in addition to in-house ITA performance assessments,[1] as this test now contains an integrated speaking performance component. Most commonly, TOEFL iBT speaking scores are used as a screening measure, where the cutoff scores range between 23 and 28 and a score of 27 or higher is considered to be a relatively reliable indicator of ITA proficiency (T. Farnsworth, 2012). While in-house assessments and TOEFL iBT speaking generally serve complementary functions in ITA testing, some universities are contemplating using TOEFL iBT speaking scores for ITA certification,[2] that is, for permission to teach or assist instruction based on the level of oral English-language proficiency. There is certainly a practical and cost-saving advantage to such an application of the speaking scores, as a considerable number of universities already use TOEFL iBT scores for admissions. However, the validity of the speaking scores for ITA assessment purposes has yet to be empirically substantiated. Although a few studies provide evidence supporting the validity of the interpretation and use of the TOEFL iBT speaking scores for ITA certification (T. L. Farnsworth, 2013; Lim et al., 2012; Wagner, 2016; Wylie & Tannenbaum, 2006), "additional evidence needs to be established to support it as a measure of speaking ability in *instructional* settings and the use of the scores for making decisions about teaching assistantship (TA) assignments at institutions in the United States" (Xi, 2007, p. 319). Adhering to Xi's call for more research on this secondary use of the test, we sought to gather evidence needed for interpreting the TOEFL iBT speaking scores as indicators of language ability necessary for effective instructional performance as a teaching assistant in the target domain of language use.

To ensure congruency with the TOEFL iBT validity research, our study employed the argument-based approach to validation presented in Chapelle, Enright, and Jamieson (2008). This is a validation model that consists of a chain of inferences about the interpretations and uses of the test scores, propositional warrants associated with the inferences, and specific assumptions underlying the respective warrants. Each inference is established when the warrant associated with it is sustained by a collection of backing evidence that supports the assumptions underlying the warrant. Of the series of inferences Chapelle et al. (2008) outlined, we aimed to investigate the *domain description* inference, which bears particular relevance to the TOEFL research needs (articulated in the 2014–2023 TOEFL COE Research Program request for proposals). According to Chapelle et al., a warrant supporting the domain description inference is that observations of performance on the test reveal relevant knowledge, skills, and abilities in contexts representative of those in the target domain. Some assumptions underlying this warrant are that representative academic tasks are identified and that critical academic English-language skills needed for study are identified. The assumption we put forth is related to an aspect of academic language skills—functional language, essential in performing teaching tasks in different settings of the target ITA domain. Specifically, we assumed that the language functions elicited by TOEFL iBT speaking tasks are identified in authentic ITA discourse.

Given this focus on language functions, and also considering that the TOEFL iBT test design is grounded in communicative competence theories (Bachman & Palmer, 1996; Canale & Swain, 1980), this study was conducted in the tradition of systemic functional linguistics (SFL). SFL theory treats language as social semiotics, as a resource used for communication (Halliday, 1978). In other words, "speakers bring repertoires of meaning-making resources to an interaction," the meaning-making resources being functional language used "to achieve social goals as the interaction unfolds" (Rose, 2017, p. 4). To analyze functional language in our target domain, we employed the knowledge framework heuristic (Mohan, 1986, 1989), examining the linguistic resources in a corpus of instructional discourse produced by ITAs in authentic lecture, recitation, and lab settings. To further determine if the functional language identified in the ITA domain was elicited by TOEFL iBT speaking tasks, we applied the same heuristic to the analysis of a corpus of test taker responses, then we juxtaposed the language functions in ITA discourse with those identified in the spoken performance of TOEFL iBT test takers. The data revealed how the functions were realized in both corpora, enhancing our understanding of functional language use in the ITA domain and in the test responses to TOEFL iBT speaking tasks. Additionally, the comparison indicated that TOEFL iBT speaking tasks elicited most of the language functions identified in ITA discourse, thus providing backing evidence for the stated assumption. This supporting evidence, as well as a discrepancy detected in the use of some functions, bears implications for interpreting the meaning of scores for the secondary use of this test.

TOEFL IBT SPEAKING AND DOMAIN DESCRIPTION

Validity of TOEFL iBT Speaking

In university settings, TOEFL iBT, including the speaking component, has been adopted for both undergraduate- and graduate-level decisions, as no strong evidence was found to support differences between academic levels and disciplines in terms of tasks relevant for course completion (Jamieson et al., 2000). Numerous studies validated such uses of the TOEFL iBT test with respect to various aspects of assessment, including test task design, relationships with other criteria in academic contexts, and factor structure (Bridgeman, Powers, Stone, & Mollaun,

2012; Ockey, Koyama, Setoguchi, & Sun, 2014; Sawaki, Stricker, & Oranje, 2009). A research agenda has also been initiated for ITA-related uses of scores. T. L. Farnsworth (2013) conducted a construct validity study of the TOEFL iBT speaking test for ITA certification using factor analysis. His study indicated that this test and a local assessment of ITA oral proficiency measured the same underlying construct. Xi (2007, 2008) obtained criterion-related validity evidence for the potential use of TOEFL iBT speaking scores for ITA screening by comparing those scores with local ITA measures administered in three U.S. universities. Lim et al. (2012) used the SPEAK test as a criterion measure, finding moderate correlations between the two tests. An earlier standard-setting study by Wylie and Tannenbaum (2006) established a minimum recommended TOEFL iBT speaking cut score of 23 for ITA screening and the score of 26 as comparable to the Test of Spoken English score of 50. These studies have begun to provide a foundation of evidence supporting the use of TOEFL iBT speaking scores for ITA screening and/or certification purposes.

Bachman and Palmer (1996) maintained that "to justify the use of language tests, we need to be able to demonstrate that performance on language tests corresponds to language use in specific domains other than the language test itself" (p. 23). Kane (2004) concurred, stating that "if the test is intended to be interpreted as a measure of competence in some domain, then efforts to describe the domain carefully and to develop items that reflect the domain . . . tend to support the intended interpretation" (p. 141). From this perspective, uses of TOEFL iBT scores to make decisions regarding admissions and appropriate curricula for test takers can be considered warranted, for situational variables are integrated in task design as part of task characteristics to reflect a variety of academic domains that applicants may encounter at U.S. universities (Butler, Eignor, Jones, McNamara, & Suomi, 2000; Jamieson et al., 2000). Two studies, in particular, inquired whether TOEFL iBT tasks satisfactorily reflect the characteristics of the target academic domain tasks. Cumming, Grant, Mulcahy-Ernt, and Powers (2005) interviewed seven experienced English as a second language (ESL) instructors, obtaining positive evidence for the domain representation of prototype TOEFL tasks as well as for the consistency of student performance on those tasks in ESL classes. In a similar vein, Rosenfeld, Leung, and Oltman (2001) conducted a survey investigation with faculty, undergraduate students, and graduate students from 21 universities in the United States and Canada to evaluate task statements developed based on the TOEFL 2000 Framework (Jamieson et al., 2000), finding that the tasks were considered relevant and important for academic success. A limitation of these studies is that they both relied only on stakeholders' perceptions about the nature of the target domain tasks for the justification of test task design.

Chapelle et al. (2008) explained that "domain description links performances in the target domain to the observation of performance in the test domain" (p. 14), referring to performance as "language skills, knowledge, and processes needed *for study* in English-medium colleges and universities" (p. 19, emphasis added). However, it is reasonable to assume that the types of tasks international applicants have to engage in as graduate students *for study* and the types of tasks ITAs need to engage in *for teaching* are different, and that imposes constraints on the generalization made from their performance on the test to their performance in teaching contexts, that is, in the target domain of language use.

Target Domain Considerations

According to the TOEFL 2000 Speaking Framework (Butler et al., 2000), context variables believed to affect language use include setting and participant roles. Oral communication in academic settings (e.g., one on one, small groups, or large classes) is characterized as "primarily

directed towards acquiring, transmitting, and demonstrating knowledge" as well as intended for "organization, management, and regulation of learning activities" (Butler et al., 2000, p. 2). Another important variable considered in view of task purpose and situational characteristics is discourse features, which can be generic/pragmatic and structural (Butler et al., 2000). The latter include such elements as accomplishment of the task, sufficiency of response length and complexity, grammatical adequacy, and lexical precision. The generic/pragmatic discourse features function at micro and macro levels. The micro level is defined by short interactive turns in an exchange or a series of exchanges (e.g., directives, declaratives, and suasion); the macro level is characterized in terms of extended discourse (e.g., patterns of exposition, organization, and rhetorical properties of text types).

The types of exchanges in ITAs' pedagogic discourse arguably vary depending on the tasks and purposes they have to accomplish to fulfill teaching roles in lab, recitation, or lecture settings (Axelson & Madden, 1994). Presenting content, leading discussions, clarifying questions, and so on, means that the English communication demands in the target ITA domain include functional language ability (Lazaraton & Wagner, 1996)—in other words, the ability to use such language functions as explaining, narrating, describing, defining, comparing, evaluating, and concluding to effectively convey content. Functional language plays a critical role in the assessment of speaking ability in academic contexts (Butler et al., 2000). However, despite being reportedly problematic in ITA–student communication (Madden & Myers, 1994; Tyler, 1992; Williams, 1992), functional language use in different types of exchanges at different levels of interactivity has been underinvestigated in both pedagogic discourse research and in TOEFL and ITA testing research. Existing pedagogic and ITA discourse studies mainly focus on textual features and varying functions of individual lexicogrammatical items, particularly metadiscursive devices (Liao, 2009; Tyler, 1992; Williams, 1992). To our knowledge, Levis, Levis, and Slater (2012) is the only study that presented an exemplary analysis of how ITAs use functional language to organize material, build background knowledge, make connections, and reconstruct knowledge. Clearly there is a gap in knowledge that needs to be addressed to comprehensively describe functional language use in the ITA target domain.

METHODOLOGY

This study was motivated by the need to validate a secondary use of the TOEFL iBT speaking test scores for the purpose of certification of ITAs in English-medium universities. In line with this need, we investigated the domain description inference in the TOEFL interpretive argument (Chapelle et al., 2008). We aimed to conduct domain analysis and determine whether the language functions elicited by TOEFL iBT speaking tasks are identified in authentic ITA discourse. For that purpose, we employed the SFL knowledge framework heuristic and identified the functional-semantic discourse units in two corpora: a corpus of authentic ITA discourse and a corpus of TOEFL iBT speaking responses.

Analytic Framework

SFL is a descriptive and interpretive theory of language as a strategic meaning-making resource. Concerned with functional modeling of language in context, systemic functionalists describe the relation between the language used in discourse and the social practice in which it is situated. According to Mohan (1986), the situation can be considered a semiotic structure. As such, a typical situation includes certain "abstract categories of the field of situation typically realised in discourse by logical meanings of the semantic system" (p. 103). These abstract categories,

termed *knowledge structures* (KSs), constitute the core of the knowledge framework (Mohan, 1986, 1989).

The knowledge framework was initially designed to connect language and thinking skills in content-area teaching, the KSs being intended to represent both the linguistic structures of discourse and the structures in the mind that people use to process discourse (Mohan, 1998). Scholars have also adopted it as a heuristic for analyzing "doing" in the discourse in terms of semantic structure and its linguistic realizations to "illustrate how expert content-area teachers use language to teach and promote critical thinking in their disciplines" (Slater & Gleason, 2011, p. 7). Therefore this heuristic is a particularly relevant analytic framework for our investigation of functional language use in the pedagogic discourse of ITAs.

Underlying the knowledge framework is the concept of activity, defined as "a combination of action and theoretical understanding" (Mohan, 1986, p. 42). The distinction between the theoretical understanding that guides the action (i.e., knowing, or background knowledge) and the action itself (i.e., doing, or action knowledge) is rendered through dyadic theory–practice relationships between three pairs of KSs: CLASSIFICATION–DESCRIPTION, PRINCIPLES–SEQUENCE, and EVALUATION–CHOICE (see Appendix A). The KSs in each theory–practice pair are interrelated. The following utterances are an example for CLASSIFICATION–DESCRIPTION from Mohan (2007). Here the theoretical understanding of classification according to car types (fire engine) is necessary for the practice of describing the object.

1. Mother: What cars have you got there?
2. Stephen: There's a fire engine one with a ladder on.

Overall at theory level (background knowledge), CLASSIFICATION addresses concepts, PRINCIPLES concerns rules and cause–effect relationships, and EVALUATION regards assessments or judgments. At practice level (action knowledge), DESCRIPTION details particulars, SEQUENCE explains procedures, and CHOICE includes descriptions of reason–action relationships. Each of the three pairs of KSs is characterized by linguistic features that indicate the kind of knowledge being constructed. To this end, KSs can be considered macro-level discourse features; a particular KS at the macro level accounts for a variety of linguistic features at the micro level. For example, PRINCIPLES with such language functions as explanation, prediction, causes, and effects can be marked by micro-level linguistic means for general reference, action verbs expressing material processes, conjunctions and adverbials expressing consequences, and lexis expressing cause–effect relationships (see more examples in Mohan (1986)).

Context is another central concept in this framework, as it determines what language choices express particular functional meanings indicative of particular KSs. Context is also important from the perspective of validation and construct definition in language assessment. Xi (2015), for example, viewed context as part of the speaking construct, treated key language use contexts as integral to the domain model, and specified language use domains and speech genres as contextual factors of speaking tasks. As stated earlier, the domain we aimed to study was represented by three instructional settings: lab, recitation, and lecture. The pedagogic discourse produced in these settings can be attributed to respectively named curriculum genres, defined as "goal-driven classroom activities, devoted to the accomplishment of significant educational ends" (Christie, 2002, p. 22). In the following section, we describe how text samples of lab, recitation, and lecture curriculum genres were collected and compiled into our ITA speech

corpus. The knowledge framework analysis was conducted on this corpus as well as on a corpus of TOEFL iBT speaking responses described thereafter.

Compilation and Preparation of the Corpora

International Teaching Assistant Speech Corpus

The ITA discourse data were obtained from 52 international graduate students working as teaching assistants at Iowa State University. The participants were recruited via e-mail and personal communication with the ITAs who had taken the institutional Oral English Certification Test (OECT) between July 2011 and August 2015. Based on OECT results, students are either fully certified (Level 1), conditionally certified (Level 2), certified with restrictions (Level 3), or not certified (Level 4). Level 1 test takers are generally assigned a lecturing instructor role by their departments, while Level 2 and Level 3 test takers are assigned recitation leader and lab assistant roles, respectively. Test takers with Level 4 are still given an assistantship, but with teaching-related responsibilities such as grading or setting up and maintaining equipment. Because we intended to collect data from settings in which ITAs have to engage in direct oral interactions with students, we recruited participants with OECT scores at Levels 1, 2, and 3.

Participating ITAs signed a consent form and agreed to audio-record their speech in class. A research assistant helped the ITAs set up a small, portable audio-recording device. To avoid any interference with the ITA–student interaction, the research assistant waited outside the classroom and collected the recorded samples at the end of each session. Of the 52 ITAs who agreed to participate in the study, 21 led lab sessions only, 9 held recitations, and 9 taught both lab and recitation sessions. The remaining 13 ITAs taught a course as an independent instructor. The ITAs' first-language backgrounds included Indo-Aryan (n = 18), Chinese (n = 15), Persian (n = 5), Korean (n = 4), Spanish (n = 4), Turkish (n = 2), Vietnamese (n = 2), and Slavic (n = 1; plus one unknown). In terms of gender, 31 of our participants were men, and 21 were women.

Most participants (n = 49) provided two speech samples. Table 1 describes the composition of the ITA speech corpus, which contains 119 spoken texts with a total word count of 638,233 words. The lecture and recitation audio files were 50–80 min long, whereas the lab session audio files were 1–3 hours long. The average word counts of a speech sample per curriculum genre were as follows: lab, 8,677 words; lecture, 4,090 words; and recitation, 3,747 words.

Table 1 Composition of the International Teaching Assistant Speech Corpus

Curriculum genre	Speech samples	Participants	Disciplines
Lab	59	30	aerospace engineering, biology, chemistry, computer science, computer engineering, construction engineering, economics, engineering mechanics, mechanical engineering, physics, speech communication, statistics
Recitation	35	18	chemistry, computer science, economics, mathematics, physics
Lecture	25	13	apparel, events and hospitality management, English, food science

Total	119	52	16

All ITA audio files were first processed to remove unnecessary noise or dialogue that occurred before or after the session. Then the files were deidentified, each file name containing the following information: genre index, participant ID number, ordinal number of the speech sample, and month/year when the sample was collected (e.g., lab-01-2-nov2015). For consistency, we adapted ETS transcription conventions. A minor modification was introduced for fillers such as "uh" and "mm," the multiple consequent recurrences of which were transcribed as one instance. The transcribers, native-English-speaking undergraduate research assistants, received training in the use of transcribing conventions and continually consulted one of the investigators when clarification was needed. They transcribed only ITA talk (student talk was not included) and also added a time stamp in the transcription for unclear words or utterances so that the researchers could listen for audio clues, if needed, when later annotating the ITA texts. The transcription files were then formatted as plain text files needed for further annotation.

TOEFL iBT Speech Corpus

Our TOEFL iBT data set consisted of 2,879 audio files and transcriptions provided by ETS. These were speech samples of independent and integrated responses from the TOEFL iBT speaking test. Although the TOEFL iBT speech corpus could not be fully compatible with the ITA speech corpus in terms of proficiency level, we attempted to account for comparability to the degree possible. Considering that the lowest converted final TOEFL iBT speaking score of ITAs who took the OECT between July 2011 and August 2015 was 15, we included responses with 2, 3, and 4 points in the raw score for individual speaking tasks because this score range may be included in converted final scores of 15 and higher. We excluded responses that obtained 1 point in the raw score because the converted final score of their composite scores over the six TOEFL iBT speaking tasks (i.e., a total of 6 points in the raw score) was below 15. Overall, the corpus contained 2,738 speech samples of responses to Tasks 1–6 and scores 2–4 (Table 2). The size of this corpus amounted to 311,570 words (Tasks 1–2 responses, 88,589 words; Tasks 3–6 responses, 222,981 words). The average number of words per text was 114.

Table 2 Composition of the TOEFL iBT Speech Corpus

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6		
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	Total
Score 2	105	88	91	84	92	84	95	106	95	107	87	98	1,132
Score 3	98	118	105	119	109	112	93	93	105	105	108	101	1,266
Score 4	23	28	33	32	23	32	26	32	28	22	31	30	340
Subtotal	226	234	229	235	224	228	214	231	228	234	226	229	
Total	460		464		452		445		462		455		2,738

The TOEFL iBT speech corpus was more diverse than the ITA speech corpus given the greater number of speakers (n = 481) as well as the lack of geographical restrictions. The test takers contributing to the corpus were native speakers of a wider variety of languages (n = 42): 70 speakers of Korean, 69 speakers of Chinese, 59 speakers of Indo-Aryan, 45 speakers of

Spanish, 32 speakers to Arabic, 32 speakers of Japanese, 21 speakers of German, 14 speakers of Turkish, 13 speakers of French, 13 speakers of Tagalog, 11 speakers of Russian, 10 speakers of Thai, and 91 speakers of other languages. Of 481 examinees, 206 identified their gender as female, whereas 217 identified as male, and 58 did not specify their gender.

Refinement of the Knowledge Framework

To better understand how the knowledge framework categories would be realized in our corpora, we first conducted an analysis of a randomly selected set of TOEFL iBT and ITA speech samples. This preliminary application of the framework revealed that existing KS definitions and descriptions did not capture the complexity of the target curriculum genres and were not precise enough for our intended two-tier (i.e., KS and language function) corpus annotation task. Consequently, our initial analysis was extended to refine the framework. With our research assistants (graduate students in applied linguistics), we jointly analyzed texts from both corpora, discussing each sentence and clause vis-à-vis the KSs in Mohan's (1989, 1998) knowledge framework, particularly addressing ambiguous, unclear, ill-defined, and missing functional categories. As a result, we refined the framework and added new functions (italicized in Table 3).

Table 3 Knowledge Structure and Added Language Functions

Knowledge structure	Language functions
KS1 CLASSIFICATION	classifying, defining
KS2 DESCRIPTION	describing, comparing, *exemplifying*, quantifying, spatial positioning
KS3 PRINCIPLES	explaining, predicting, concluding, hypothesizing, demonstrating cause–effect, setting rules, specifying means, specifying ends
KS4 SEQUENCE	reporting, indicating order, indicating process, instructing, narrating
KS5 EVALUATION	evaluating, *conceiving ideas*, making judgments
KS6 CHOICE	*making choices*, *presenting options*, *expressing desire*, *advising*, expressing opinions, *presenting arguments*

Note. Italicized functions are new functions.

Additionally, we developed specific descriptors for the language functions of each KS. Unlike Mohan's (1986) description of KS pairs, our descriptors differed in the degree of specificity. For instance, Mohan listed some of the same linguistic features to describe more than one KS (e.g., general reference and additive conjunction for both KS1 CLASSIFICATION and KS2 DESCRIPTION). Our descriptors referred to individual functions pertaining to an individual KS, as in the example of the *classifying* function in KS1 CLASSIFICATION:

- arranging concrete or abstract concepts in sets/classes/categories pertaining to a broader group
  - e.g., *classify (as/into), divide (into), organize (into), categorize (as/into), sort (into)*
  - e.g., *Alcoholism is classified as a substance abuse disorder in the Diagnostic and Statistical Manual of Mental Disorders (DSM–III).*
- placing in an aggregate/set/group/class/category of concrete or abstract concepts
  - e.g., *include, incorporate (in), be a kind/type of*

    - e.g., *Additional promotional activities included organizing the dedication program for operation Turnkey, the new automated post office, and a conference with representatives of the universities in the area.*
- distinguishing from an aggregate/set/group/class/category of concrete or abstract concepts
    - e.g., *exclude, except (for), other (than), apart from, aside from, leaving out, besides*
    - e.g., *They were all there except me.*
- having as contents or part of the contents and/or indicating whole/part relations
    - e.g., *comprise, constitute, made up of, contain, have*
    - e.g., *An experiment might consist of five rounds.*
- specifying or referring to a type of category/class/group/set
    - e.g., *Assignment 3, Question 2*
    - e.g., *Unit 6 in the textbook has many examples of this.*
- indicating a part or whole of an equation/formula/chemical compound and the like
    - e.g., *equal, K is Y, 2 + 2 = 4*
    - e.g., X cross Y equals Z.
- indicating belonging or possession
    - e.g., *possessives (my, your/s, his, their/s), own, have, possess, belong, be in possession of*
    - e.g., *This mansion belongs to the Adamses. This is the house of my father.*

Annotation of the Corpora

The development of detailed KS descriptors partially overlapped with pilot annotation through concurrent segmentation and classification. These two strands of work informed the development of detailed annotation guidelines. Considering the insights gained from our preliminary analysis and framework refinement, the guidelines defined the clause as the unit of analysis and stipulated rules and exceptions (particularly because both corpora contained learner language that was often inaccurate, incomplete, and prone to different interpretations). Involving the research assistants in the refinement of the knowledge framework helped them better understand the KS categories needed for corpus annotation. Additionally, they went through substantial practice using the annotation guidelines for a series of exercise annotation and adjudication sessions, for which they annotated randomly selected texts independently and then discussed agreements and disagreements with the principal investigator as a group.

Three annotators annotated corpus texts independently, each being assigned different sets of TOEFL iBT and ITA texts. Simultaneous annotation of texts from both corpora was intended to prevent the annotators from developing a bias toward a particular type of discourse. The new KS descriptors and the annotation guidelines were continually consulted during corpus annotation. Examples of linguistic realizations also continued to be confirmed and added to the descriptors of the language functions. Additionally, the annotators consulted with each other and the principal investigator via e-mail regarding excerpts whose functional meaning was unclear. Such issues were also resolved at weekly working group meetings by identifying the most functionally prominent linguistic cues, clarifying causes of possible misunderstanding, and sharing analytic strategies.

For corpus annotation, we used the Callisto software because it allowed for multilayered tagging. Each clause in each text was tagged with a KS as well as with a respective language

function. As illustrated in Figure 1, different colors indicate the KSs of individual clauses, and the functions under each KS are shown in the lower section of the screen. For example, the first clause is annotated as KS4 SEQUENCE (knowledge structure) and *narrating* (language function).

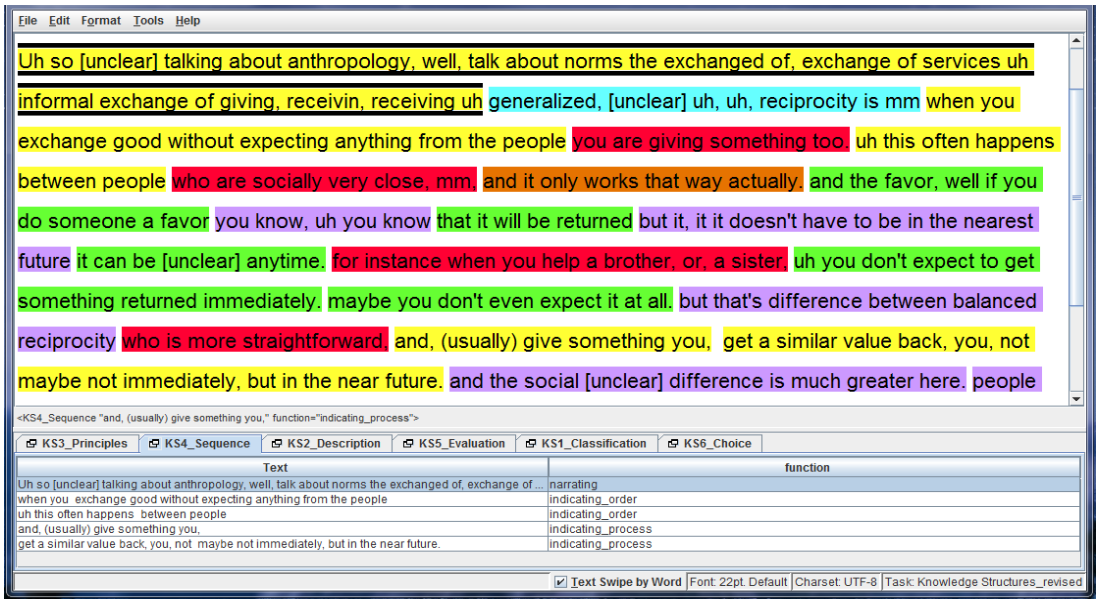

Figure 1 A screenshot of text annotation in Callisto. Text is from the international teaching assistant speech corpus.

Given that the ITA speech samples were significantly longer than the TOEFL responses, annotating the full texts was not feasible for our project's timeline. Therefore we reduced the size of the ITA data for annotation to 311,613 words, which was comparable to 311,570 words in the TOEFL iBT speech corpus. In this attempt for comparability, we did not remove texts from the ITA speech corpus; rather, we annotated the first 2,700–2,800 words from the beginning of all ITA texts in the corpus. The average word count of annotated ITA texts is 2,619.

To verify, improve, and maintain the quality of annotation, weekly reliability check sessions were held throughout the annotation process. For these sessions, the annotators were assigned a total of 26 sets[3] of the same texts; each set contained three to four texts from the TOEFL iBT speech corpus and an excerpt from an ITA text, with word counts ranging between 393 and 439, respectively. Acceptable levels of interannotator reliability were achieved. Cohen's kappa for all the sets was .75. Each annotator's individual agreement with the final adjudications was higher than the agreements between them across all the reliability check sessions (.91 and .81). To assess the annotators' performance in agreement with the final adjudications, Light's kappa coefficients were calculated by averaging pairwise Cohen's kappa coefficients within each set and across the entire sets (Gwet, 2014). The overall agreement among the annotators in relation to the final adjudications was .82. Conger's kappa coefficients were also calculated to estimate the percentage chance agreement among individual annotators because, in addition to the texts prepared for adjudication, they independently annotated individually assigned texts. The overall Conger's kappa estimate was also .82.

Analysis of the Annotated Corpora

The annotated texts were saved as .sgml files. To extract the data from the annotated TOEFL iBT and ITA Speech Corpora, a Python code was written, tested, and applied. This allowed us to extract the annotated units into data types as follows:

- annotated units per KS category
- annotated units per language function
- frequencies of the KS categories occurring in each text
- frequencies of the language functions occurring in each text
- frequencies KS categories and language functions from four subsets of the spoken corpora
  - responses to TOEFL iBT Speaking Tasks 1–2
  - responses to TOEFL iBT Speaking Tasks 3–6
  - ITA lab discourse
  - ITA recitation and lecture discourse

Extracting the data this way enabled us to create four subsets (lab, recitation/lecture, TOEFL iBT Tasks 1–2, and TOEFL iBT Tasks 3–6) and to analyze frequencies at the level of KSs and at the level of language functions. Because ITAs in lab settings tend to draw more upon practical knowledge, while ITAs in recitations and lectures draw more upon the focal content, we treated lab discourse as relatively comparable with responses to TOEFL independent Tasks 1–2, and their recitation and lecture discourse with responses to TOEFL integrated Tasks 3–6. Although we realize that such a correspondence is rather approximate, we considered it to be useful given the background knowledge (theory) and action knowledge (practice) dyad underlying the knowledge framework. Overall, the annotated data afforded descriptive analyses of the two corpora as well as comparative analyses of the four subsets.

Correspondence analysis (CA) was employed for further quantitative analysis of the annotated data using the IBM SPSS Statistics 24.0.[4] This is a multivariate analytic method used in exploratory investigations of associational relationships among multiple categorical data (Clausen, 1998; Hair, Black, Babin, & Anderson, 2009; Sourial et al., 2010; Yelland, 2010). As an exploratory approach, CA does not require independent observations because no statistical inference is to be made. Rather, relative associations between categorical variables are examined through data visualization. In short, our data were organized in contingency tables where the row categories were the four subsets and the column categories were the six KSs or the language functions subordinate to each KS. CA calculated chi-square (χ^2) distances between the row and column categories, and their normalized profiles were plotted as points in low-dimensional graphs, called perceptual maps. The associations between the subsets and KSs, and between subsets and language functions, were examined in view of the distances among them as plotted on perceptual maps.

RESULTS

Focusing on the domain description inference, we aimed to investigate the assumption that the language functions elicited by TOEFL iBT speaking tasks are identified in authentic ITA discourse. To accomplish that, we needed to identify the language functions used to construct knowledge in authentic ITA discourse and the language functions elicited by TOEFL iBT speaking tasks and to determine whether TOEFL iBT speaking tasks elicit the language functions identified in English-medium ITA instructional contexts.

The use of KSs was first examined descriptively based on their occurrence in the TOEFL iBT speech corpus and the ITA speech corpus. The KS frequencies revealed similarities and differences between the two corpora. All six KSs were identified in both corpora (Table 4). KS1 CLASSIFICATION was rarely used both by ITAs and TOEFL test takers. KS2 DESCRIPTION, KS5 EVALUATION, and KS6 CHOICE were relatively equally distributed. A detectable difference, however, surfaced in the use of KS3 PRINCIPLES and KS4 SEQUENCE, the former being more prominent in the TOEFL iBT speech corpus (28.51%) and the latter in the ITA speech corpus (28.94%).

Table 4 Knowledge Structures in the International Teaching Assistant and TOEFL iBT Speech Corpora

Knowledge structure	TOEFL iBT speech corpus		ITA speech corpus	
	Frequency	Percentage	Frequency	Percentage
KS1 CLASSIFICATION	446	1.27	972	1.86
KS2 DESCRIPTION	7,449	21.24	10,265	20.32
KS3 PRINCIPLES	9,997	28.51	9,960	19.72
KS4 SEQUENCE	6,740	19.22	14,617	28.94
KS5 EVALUATION	5,154	14.70	6,680	13.22
KS6 CHOICE	5,284	15.07	8,047	15.93
Total	**35,070**	**100**	**50,511**	**100**

Note. ITA = international teaching assistant.

Another observation worth noting is that, despite the fact that the size of both corpora was almost identical (roughly 311,000 words), the raw frequency of the annotated units in the ITA speech corpus was much higher. The 15,441 KS unit difference between the two corpora reflects the nature of the language produced by the speakers in the two contexts, one of which was interactive, while the other was not. Because we manually annotated the data, it became apparent to us that the ITAs produced many short utterances during their interactions with their students, such as confirmation checks (e.g., *Isn't it? Right?*), solicitation of answers through short questions (e.g., *How? Sorry?*), and quick evaluation of students' input (e.g., *yes, good, bad, excellent*).

Figure 2 further distinguishes the KSs per subset within each corpus, showing that their frequencies were comparable. The corpus-level difference between KS3 PRINCIPLES and KS4 SEQUENCE applies to the subsets as well.

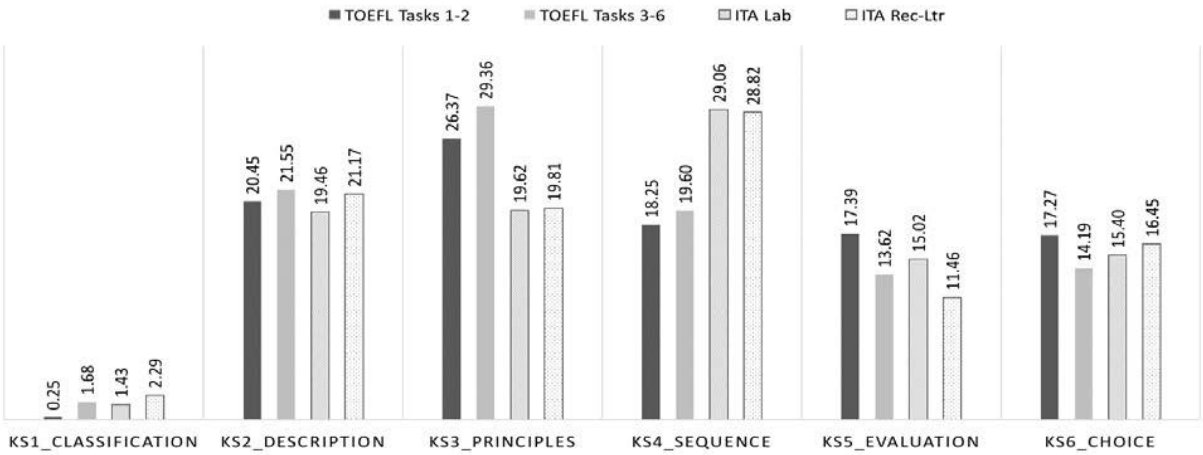

Figure 2 Knowledge structures in the subsets of the international teaching assistant and TOEFL iBT speech corpora.

Within each KS that occurred in the ITA speech corpus, the language functions were distributed as detailed in Appendix B. Notably, except KS1 CLASSIFICATION, each KS was primarily realized through one most frequent function in both lab and recitation/lecture subsets: *describing* in KS2 DESCRIPTION, *predicting* in KS3 PRINCIPLES, *instructing* in KS4 SEQUENCE, *evaluating* in KS5 EVALUATION, and *expressing opinion*s in KS6 CHOICE. These functions accounted for 9%–15% in the lab data and 4%–16% in the recitation/lecture data. The frequency of the remaining functions in the two subsets was similar (lab, 1%–8%; recitation/lecture, 1%–6%). In both subsets, the frequency of the following functions was below 1%: *defining* in KS1 CLASSIFICATION; *setting rules*, *specifying ends*, and *specifying means* in KS3 PRINCIPLES; *narrating* and *reporting* in KS34 SEQUENCE; and *making choices* and *presenting arguments* in KS6 CHOICE. *Making judgments* in KS5 EVALUATION did not occur in the recitation/lecture data and had only two instances in the lab data.

Appendix C further presents parallel findings from the TOEFL iBT speech corpus. Here KS2 DESCRIPTION *describing*, KS5 EVALUATION *evaluating*, and KS6 CHOICE *expressing opinions* were prevalent in both types of tasks (Tasks 1–2, 9%–16%; Tasks 3–6, 7%–16%). In KS4 SEQUENCE, the function that emerged as more frequent was *indicating order* (Tasks 1–2, 10.91%; Tasks 3–6, 9.20%). KS3 PRINCIPLES *predicting* was somewhat richer in Tasks 3–6 (10.89%) than in Tasks 1–2 (7.68%); the opposite can be said about *explaining* (Tasks 1–2, 9.14%; Tasks 3–6, 7.34%). The frequencies of the remaining functions were very comparable, the middle percentage stratum ranging from 1% to 3% for Tasks 1–2 and from 1% to 4% for Tasks 3–6. The functions with frequencies below 1% in both data sets were KS1 CLASSIFICATION *classifying*; KS2 DESCRIPTION *spatial positioning*; KS3 PRINCIPLES *setting rules*, *specifying ends*, and *specifying means*; KS4 SEQUENCE *instructing*, *narrating*, and *reporting*; and KS6 CHOICE *making choices* and *presenting arguments*. The only function that did not occur in Tasks 1–2 was KS5 EVALUATION *making judgments*.

When juxtaposing the frequencies of functions in the TOEFL iBT and ITA speech corpora, corresponding patterns were discernable at the level of functions (Appendix D). KS1 DESCRIPTION *describing*, KS3 PRINCIPLES *predicting*, KS5 EVALUATION *evaluating*, KS6 CHOICE *expressing opinions*, and the two functions of KS1 CLASSIFICATION were similarly distributed in both corpora. KS3 PRINCIPLES *predicting* (9.98%) and *explaining* (7.86%) together differentiate TOEFL test takers' functional language use from ITAs.' A

glaring difference, however, is in the use of the *instructing* function of KS4 SEQUENCE, which was the most common in ITA discourse (15.36%) and extremely rare in TOEFL responses (0.38%). Figure 3 displays what visually appears to be a pattern of occurrence decreasing from approximately 30% to less than 1%. Overall, this pattern applies to both corpora, indicating a comparable distribution of functions. The proportions of individual functions, however, vary to some extent, and this variation is further addressed by comparing the subsets of each corpus.

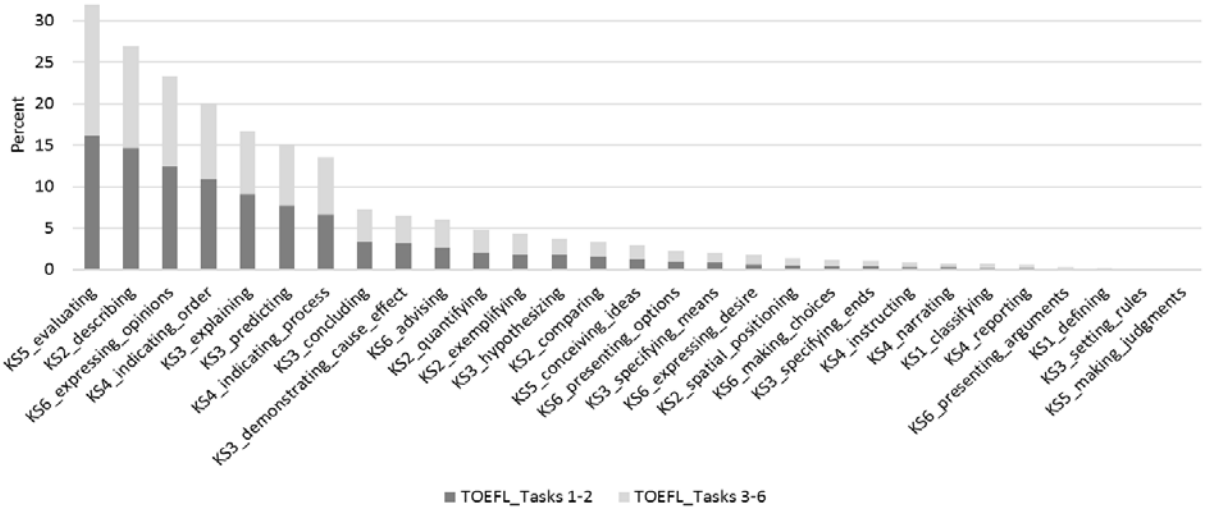

Figure 3 Frequency of functions in the TOEFL iBT and international teaching assistant speech corpora.

Sorting by high to low frequency rate divided the data into three groupings (Table 5). Eight functions in the top group (KS5 *evaluating*, KS4 *instructing*, KS2 *describing*, KS6 *expressing opinions*, *KS3 predicting*, *KS4 indicating order*, *KS3 explaining*, and *KS4 indicating process*) dominated more than 70% of the functional language use in all four subsets, so they can be considered characteristic of the text types at hand. The second grouping of more than 16% contains seven functions (KS6 *advising*, KS3 *hypothesizing*, KS2 *quantifying*, KS3 *concluding*, KS2 *spatial positioning*, KS3 *cause–effect*, KS2 *exemplifying*, KS6 *presenting options*). The remaining 13 functions in the third grouping occurred at less than 9%.

Table 5 Frequency-Based Grouping of Language Functions

Function	Tasks 1–2	Tasks 3–6	Lab	Recitation/lecture
KS5 *evaluating*	16.16%	12.33%	13.84%	10.24%
KS4 *instructing*	0.29%	0.42%	14.96%	15.75%
KS2 *describing*	14.61%	15.80%	12.78%	14.57%
KS6 *expressing opinions*	12.51%	6.92%	9.36%	10.18%
KS3 *predicting*	7.68%	10.89%	8.20%	7.46%
KS4 *indicating order*	10.91%	9.20%	7.96%	6.06%
KS3 *explaining*	9.14%	7.34%	3.16%	3.65%
KS4 *indicating process*	6.64%	7.60%	5.70%	5.48%
Subtotal	*77.94%*	*70.5%*	*75.96%*	*73.39%*

KS6 *advising*	2.69%	3.36%	3.81%	3.73%
KS3 *hypothesizing*	1.76%	2.80%	2.95%	3.52%
KS2 *quantifying*	2.03%	1.74%	2.84%	3.16%
KS3 *concluding*	3.40%	3.39%	2.40%	2.61%
KS2 *spatial positioning*	0.44%	0.42%	2.19%	1.42%
KS3 *demonstrating cause–effect*	3.14%	3.92%	1.38%	1.23%
KS2 *exemplifying*	1.77%	2.61%	0.68%	0.99%
KS6 *presenting options*	0.98%	2.01%	0.87%	1.12%
Subtotal	*16.21%*	*20.25%*	*17.12%*	*17.78%*
KS6 *expressing desire*	0.63%	1.10%	1.31%	1.29%
KS5 *conceiving ideas*	1.24%	1.27%	1.18%	1.23%
KS2 *comparing*	1.61%	0.99%	0.98%	1.02%
KS1 *classifying*	0.23%	0.43%	0.95%	1.80%
KS1 *defining*	0.02%	1.25%	0.48%	0.49%
KS4 *reporting*	0.15%	1.76%	0.28%	0.68%
KS3 *specifying ends*	0.41%	0.26%	0.89%	0.63%
KS3 *specifying means*	0.84%	0.73%	0.57%	0.61%
KS4 *narrating*	0.27%	0.62%	0.17%	0.86%
KS3 *setting rules*	0.01%	0.03%	0.08%	0.10%
KS6 *making choices*	0.43%	0.66%	0.05%	0.07%
KS6 *presenting arguments*	0.04%	0.14%	0.02%	0.05%
KS5 *making judgments*	0.00%	0.02%	0.01%	0.00%
Subtotal	*5.88%*	*9.26%*	*6.97%*	*8.83%*
Total	**100%**	**100%**	**100%**	**100%**

Importantly, comparing the graphs in Figures 4 and 5 horizontally, it becomes apparent that, except for KS4 *instructing*, the frequency-based order of functions in the first grouping (KS5 *evaluating*, KS2 *describing*, KS6 *expressing opinions*, KS3 *predicting*, KS4 *indicating order*, KS3 *explaining*, and KS4 *indicating process*) is similar across the four subsets of the two corpora. This observation is different for the functions in the second and third groupings. For instance, KS2 *exemplifying* placed 12th in TOEFL responses but 21st in ITA discourse; KS2 *special positioning* placed 19th and 13th; KS3 *demonstrating cause–effect* placed 9th and 14th; and KS6 *making choices* placed 20th and 27th, respectively.

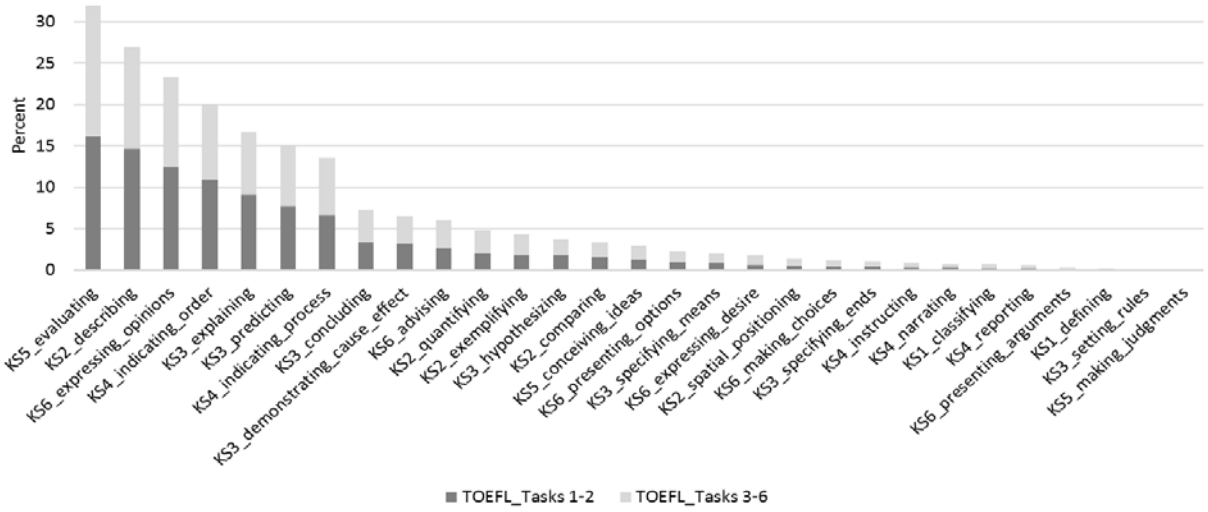

Figure 4 Frequency percentages of functions in the TOEFL iBT corpus.

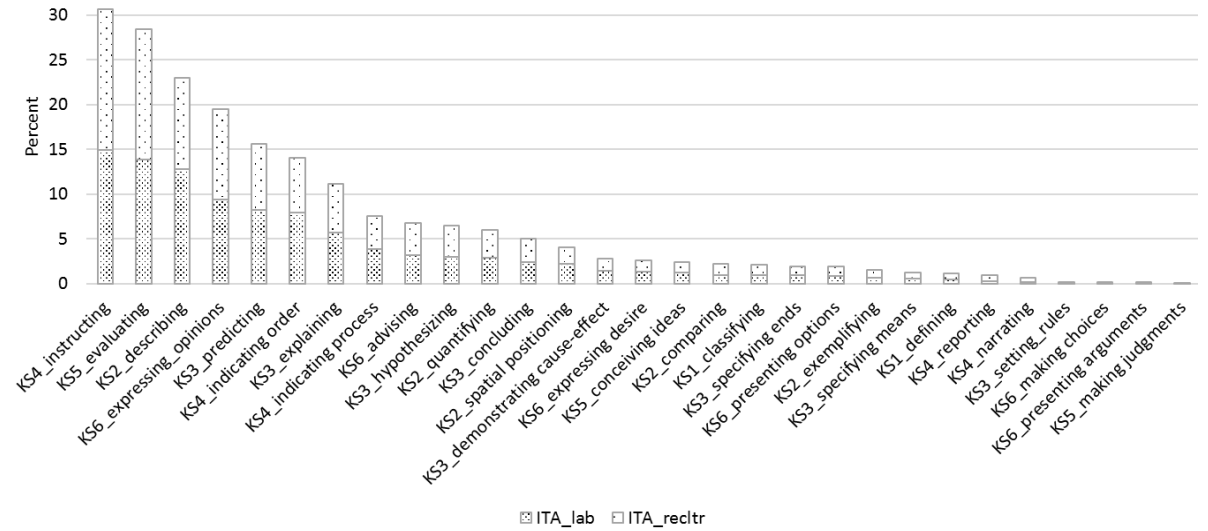

Figure 5 Frequency percentages of functions in the international teaching assistant corpus.

Additionally, we explored the associations among the six KSs and the four subsets included in our spoken corpora through CA. The chi-square value, which in CA represents a weighted Euclidean distance between categories, was significant (χ^2 = 2009.051, *df* = 15), indicating that the measure of association between the KSs and categories is significant. In Table 6, the eigenvalues, called inertia, reflect the relative importance of each dimension; the cumulative proportion of inertia here indicates that about 97% of the entire variance could be explained in terms of two standardized dimensions.

Table 6 Correspondence Analysis Summary for the Relations Between Knowledge Structures and Subsets

				Proportion of inertia		Confidence singular value	
Dimension	Singular value	Inertia	Chi-square	Accounted for	Cumulative	*SD*	Correlation 2
1	.135	.018		.781	.781	.003	−.016

2	.066	.004		.186	**.968**	.003
3	.027	.001		.032	1.000	
Total		.023	2,009.051[a]	1.000	1.000	

[a]$p = .000$; $df = 15$.

Thus we used two standardized dimensions to perceptually map the normalized statistical profiles of our categorical variables (i.e., the subsets in rows; the KSs and the language functions in columns). Figure 6 shows the perceptual map of the associations among the KSs and the subsets. The dashed lines in the figure were added to mark the origin (0,0) of the coordinates. The straight lines connect the origin and the four profile points of genre subsets; the dotted lines connect the origin and the KS categories. The further the row and column categories in the data set are located from the origin, the more discriminating they are. Also, the smaller the angle is between two lines drawn to the profile points of the row and column variables from the origin, the stronger association their corresponding row and column categories have. If the lines from the profile points to the origin are perpendicular, the categories have no relationship to each other. An obtuse angle between the lines suggests that the corresponding categories are in a negative relationship (analogous to a negative correlation).[5] We used different colors to indicate the KS and subset pairs that are closely related to each other. Pink indicates a relative association for KS1 CLASSIFICATION and recitation/lecture (ITA Rec-Ltr in Figure 6), green for KS4 SEQUENCE and Lab, orange for KS5 EVALUATION and TOEFL Tasks 1–2, and blue for KS3 PRINCIPLES and TOEFL Tasks 3–6. There are no lines connecting KS2 DESCRIPTION and KS6 CHOICE because they are very close to the origin (0,0), which means that they do not show a particular relationship with any of the subsets.

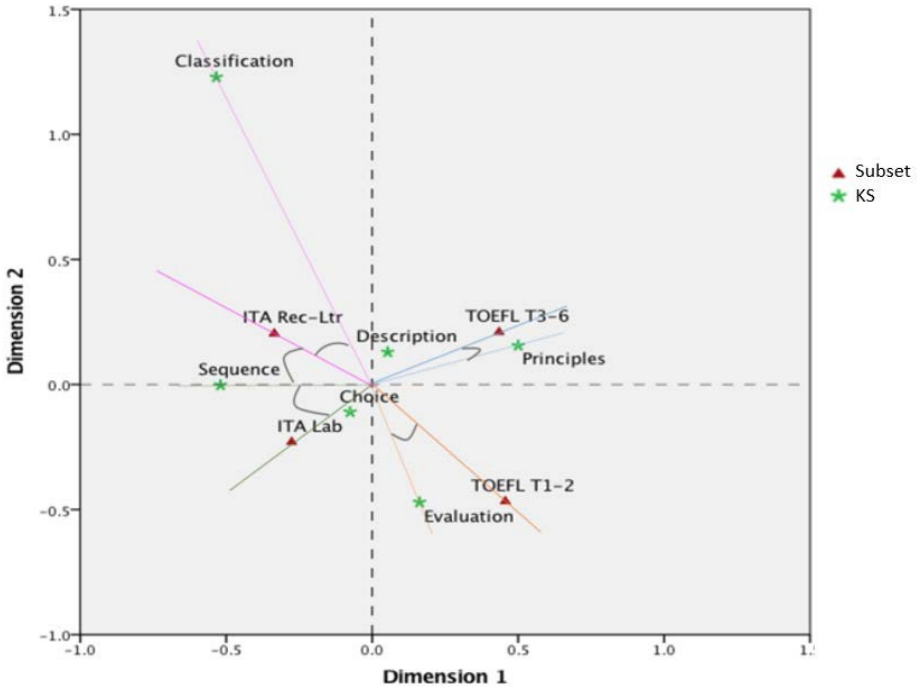

Figure 6 Perceptual map of the relationships among knowledge structures and subsets.

In view of these considerations, based on the location of the profile points on Dimension 1 in Figure 6, it can be inferred that ITA subsets are different from the TOEFL responses in terms of KS use; the profile points of ITA genres are located on the left-hand side of the origin on Dimension 1, whereas the profile points of TOEFL responses are all located on the right-hand side of the origin. With respect to Dimension 2, ITA lab has an association with the responses to TOEFL Tasks 1–2, as both are below the delimiting line. Likewise, ITA recitation/lecture shares commonality with the responses to TOEFL Tasks 3–6, both appearing above the delimiting line of the dimension.

Judging by the lines between the origin and profile points in Figure 6, it becomes clear that KS1 CLASSIFICATION was the most frequently occurring KS in ITA recitation/lecture, as it is located far in the same dimension and forms an acute angle between the two corresponding profile points. KS3 PRINCIPLES is most closely related to TOEFL examinees' responses to integrated Tasks 3–6 and KS5 EVALUATION to the responses to independent Tasks 1–2. KS4 SEQUENCE appears to be related to both subsets representing ITA curriculum genres. The profile point of KS6 CHOICE is very close to the origin, which means that its association is very weak, perhaps because it was relatively common in the other subsets. Similarly, although KS2 DESCRIPTION appears to be related to TOEFL Tasks 3–6, their relationship can be interpreted as weak because it is located close to the origin. The same can be inferred about the relationship of KS2 DESCRIPTION to the other genre subsets.

Using the same statistical technique, we explored the relationships among the functions and the subsets. Again, two normalized dimensions could explain about 95% of the total variance (inertia) of the function-centered data set (Table 7).

Table 7 Correspondence Analysis Summary for the Relations Between Functions and Subsets

				Proportion of inertia		Confidence singular value	
Dimension	Singular value	Inertia	Chi-square	Accounted for	Cumulative	*SD*	Correlation 2
1	.318	.101		.836	.836	.002	.052
2	.119	.014		.118	**.954**	.003	
3	.075	.006		.046	1.000		
Total		.121	10,325.993[a]	1.000	1.000		

[a]$p = .000$; $df = 84$.

Figure 7 presents the perceptual map of the relationships between the individual functions and the subsets. Because many of the functions did not occur with a substantial degree of frequency in any of the corpora, the figure presents 18 rather than all 29 functions. We relied on Gries's (2008) and Lijffijt and Gries's (2012) normalized deviation of proportions (DP) as a measure of dispersion[6] and ran an additional CA by including only the functions with normalized DP indices greater than .25. This threshold index was chosen in order not to oversimplify the perceptual map while trying to more clearly depict the distinctions. In doing so, we also verified the positions of the selected 18 functions on the perceptual map against the perceptual map of all 29 functions; the positions were similar.

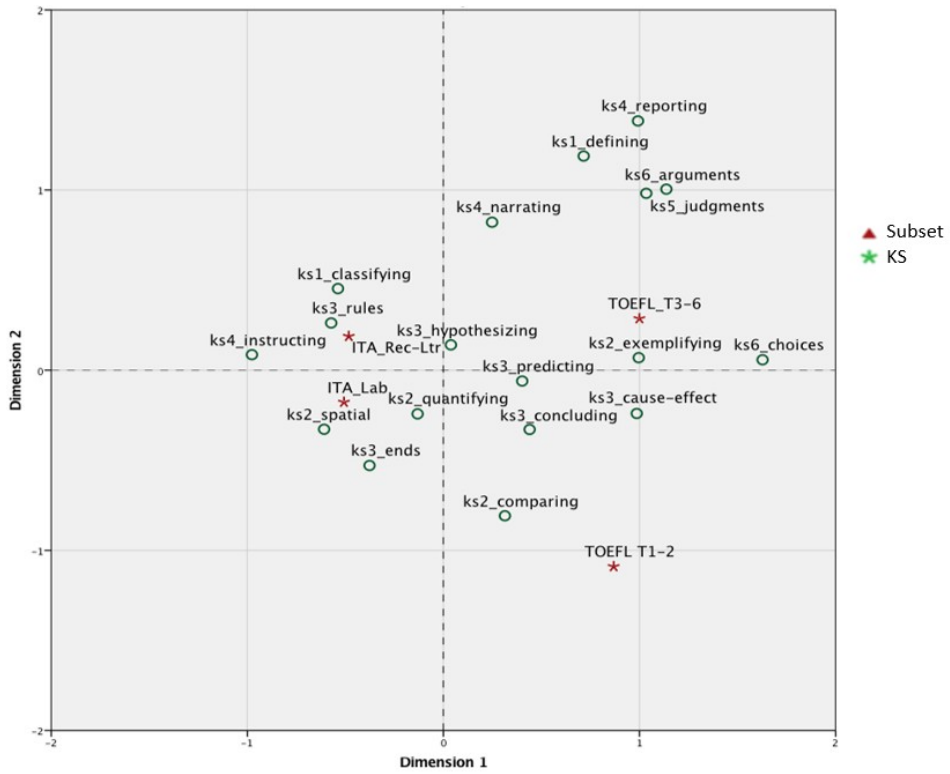

Figure 7 Perceptual map of the relations among functions and subsets.

Because drawing lines and angles between the origin and the profile points of functions and subsets (as in Figure 6) to show associations would make the perceptual map unreadable, we summarize the results in Table 8. Proportionally speaking, KS2 *comparing* is related to TOEFL speaking Tasks 1–2, whereas integrated TOEFL speaking tasks (Tasks 3–6) elicited from test taker a variety of functions, including KS1 *defining*, KS2 *exemplifying*, KS4 *reporting*, KS4 *narrating*, KS5 *making judgments*, KS6 *making choices*, and KS6 *making arguments*. The proportion of KS3 *demonstrating cause–effect* appears to be commonly higher in TOEFL speaking tasks than in ITA discourse. The functions prevalent in the ITA subsets are less disperse on the perceptual map than those prevalent in TOEFL tasks. It means that many of those functions were frequently used in both lab and recitation/lecture genres. On the basis of the dimensions in which functions and subsets are located on the perceptual map, lab appears to have close associations with KS2 *spatial positioning* and KS3 *specifying ends*. On the other hand, recitation/lecture is closely related to KS3 *setting rules* and KS1 *classifying*. As it is close to the origin, KS4 *instructing* appears to be used in both ITA genre subsets, while having a slightly closer relationship with recitation/lecture. The map also suggests that KS2 *quantifying* and KS3 *hypothesizing* are commonly used in both TOEFL speaking responses and ITA discourse, compared to the other functions, as they are located very close to the origin.

Table 8 Associational Relations Between Functions and Genre Subsets

Genre subsets	Function
ITA lab	• KS4 *instructing* • KS2 *spatial positioning* • KS3 *specifying ends*

	• KS3 *setting rules*
ITA recitation/lecture	• KS4 *instructing* • KS1 *classifying* • KS3 *setting rules* • KS2 *spatial positioning*
TOEFL Tasks 1–2	• KS2 *comparing* • KS3 *demonstrating cause–effect*
TOEFL Tasks 3–6	• KS4 *reporting* • KS1 *defining* • KS6 *presenting arguments* • KS5 *making judgments* • KS6 *making choices* • KS4 *narrating* • KS2 *exemplifying* • KS3 *demonstrating cause–effect*

Note. ITA = international teaching assistant.

It has not escaped our notice that our CA results clearly discriminated between the ITA discourse and the TOEFL responses, as mapped onto Dimension 1. Importantly, mapping onto Dimension 2 supported our initial assumption that ITA lab discourse should be considered comparable with TOEFL independent Tasks 1–2 and recitation and lecture discourse with integrated Tasks 3–6.

DISCUSSION

Summing up the results, it can be concluded that functional discourse in the curriculum genres of the target domain of language use (lecture, recitation, lab) and in responses elicited by integrated and independent TOEFL iBT speaking tasks can be realized with 6 KSs and 29 language functions. All 6 KSs were used to construct knowledge in both the TOEFL iBT and ITA speech corpora. Of the 29 functions identified in the subsets of both corpora, 8 functions (KS5 *evaluating*, KS2 *describing*, KS6 *expressing opinions*, KS3 *predicting*, KS4 *indicating order*, KS3 *explaining*, KS4 *indicating process*, and KS4 *instructing*) amounted to more than 70%. Such distribution suggests that some functions were more characteristic than others in both our corpora. The presence of the functions that occurred with lower frequencies indicates that there is considerable variation in functional language use across the four subsets. The implications of these findings depend on the purpose of the description of the target language domain. Highlighting key features focusing on fewer functions, perhaps the eight functions in the top group, would be a more informative outcome given the motivation of this study. A more comprehensive description, on the other hand, might recount all 29 functions.

KS use was not drastically different in the target instructional genres, and both subsets of the TOEFL iBT speaking tasks seemed to elicit functions relatively analogously across the KSs. The only discrepancy we detected was in the use of two KSs, KS3 PRINCIPLES being more frequent in test taker responses and KS4 SEQUENCE in ITA discourse. In terms of associational relations, KS3 PRINCIPLES appeared to be closely related with Tasks 3–6, KS4 SEQUENCE with ITA lab discourse. At the level of functions, the TOEFL tasks seem to have elicited more frequent KS3 *explaining*, while the CA results showed relative associations with nine functions. Of these, eight functions were related to Tasks 3–6, and only KS3 *demonstrating cause–effect* was related to both task types. Interestingly, KS4 *reporting* stood out on the

perceptual map indicating a strong association with Tasks 3–6. Although confident interpretive claims cannot be made because the frequencies of this function were low, such an association is not unexpected—in the process of corpus annotation, we observed that the test takers often reiterated the information from the reading or listening passage.

On the other hand, KS4 *instructing* predominated in recitation/lecture and lab; in fact, it was the only function with the highest frequencies and strongest associations with both types of genre subsets. This is not unexpected, though, because KS4 *instructing* naturally reflects the general purpose of teaching. Instructional discourse often engages directives and various question types, including questions for clarification or confirmation purposes. In labs, in particular, ITAs normally have to guide individual students' or small groups' work on-site by soliciting actions imperatively. The more pronounced use of KS4 *instructing* by ITAs can also be explained by the interactive nature of ITA discourse. Unlike the testing context, face-to-face interaction in the classroom entailed the use of many short, fragmented, and incomplete utterances (e.g., *How? Sorry?*). Consider a few examples:

a. *Just follow the general me– general method. What do we do first?* (#001, recitation)
b. *"Compound" is the correct answer here. What is com– com– uh compound, everyone? In this sentence, what does it mean? To figure out?* (#003, lecture)
c. *So what this means is that in the ploidal organisms you have two copies in each cell whereas in haploid organisms you have just one copy of each chromosome per cell. So uh, can you tell me what human beings are? Are they haploid or diploid?* (#004, lab)

Even though previous research on ITA discourse is very limited and not specifically focused on functional language use, our findings are comparable to some earlier studies. The variation of KSs and language functions in our subsets resonates with Axelson and Madden's (1994) conclusion that the demands on language behaviors vary depending on the type of instructional context. Rounds (1987) studied successful classroom discourse of both domestic and international teaching assistants in mathematics classes and reported that effective teaching of course content requires specification of processes, which is a direct parallel to KS4 SEQUENCE *indicating process*, which prominently surfaced in our ITA speech corpus. Similarly, the description by Levis et al. (2012) of how ITAs organize and connect content ideas is similar to the functions of SEQUENCE. That being said, it is important to remark that in Levis et al., the language patterns of ITAs were not the same as those found in the discourse of native-English-speaking teaching assistants. Therefore a replication of our study with a native-speaker corpus is warranted.

Furthermore, although this study did not scrutinize the language choices that instantiated the KS functions, it is worth mentioning that ITAs' language choices largely reflected the nature of interpersonal communication, plus disciplinary content. The functions and language choice patterns observed in the TOEFL iBT speech corpus appeared to be determined to a great extent by the task prompts. Therefore a possible implication may rest with the design of TOEFL iBT speaking tasks. Perhaps new tasks could be designed to elicit functional language that would better represent the action knowledge constructed in the academic contexts of higher education. However, this may not be justifiable for a secondary use of the test, such as ITA certification, because the benefit of maintaining a close correspondence between the test tasks and the tasks in the target language use domain might come at a greater expense in terms of fairness.

CONCLUSION

This study investigated functional language use in ITA-facilitated instructional activities to inform decisions in higher education regarding the secondary use of TOEFL iBT speaking scores for the purpose of ITA placement and certification. Through an in-depth analysis of the discourse in the target domain of language use, we sought evidence for the domain description inference in the TOEFL interpretive argument. Our assumption was that the spoken performance of TOEFL iBT test takers may contain language functions similar to those used in the target domain. Unlike previous studies, we analyzed the language production of the target population with a particular focus on functional discourse features at the micro level of interactive language exchanges vis-à-vis test takers' responses to TOEFL iBT speaking tasks. Overall, the results produced mostly positive evidence supporting our assumption regarding functional language use, as the language functions were fairly similar at higher levels in both test responses and in the curriculum genres of the target domain. Yet our data also pointed out some differences, the most notable one being in the use of KS4 *instructing*. Consequently, a follow-up study should further explore the extent to which this difference could have a measurable impact on score interpretation and decisions related to secondary test use and more generally on the chain of inferences that follow, particularly to utilization.

Ideally, a future study would also account for the challenges we encountered. For example, because recruiting a large number of ITAs on duty proved to be difficult, we had to compromise and record more than one speech sample from most participants. Additionally, transcribing audio files was at times challenging because in some cases, intelligibility was affected by the interfering sound effects in lab settings. In such cases, the researchers had to spend additional time following the time stamp in the transcribed files to listen to the original audio and decipher the utterances that were not clear to the transcribers. More importantly, the degree to which our study results can be generalized more broadly to ITA discourse is hindered due to the characteristics of the ITA corpus (e.g., compiled at one university, predominantly science and engineering). Moreover, a considerable limitation we have to acknowledge is not being able to annotate the entire ITA data set. Timewise, it was not feasible to annotate texts in full (lengths ranging between 3,000 and 9,000 words). With a corpus fully annotated, one would be able to analyze the macro level of pedagogic discourse and identify patterns of organization and rhetorical composition in terms of communicative goals and functional strategies (Swales, 1981), as has been extensively done for many genres of academic writing in various disciplines (see Biber, Connor, & Upton, 2007). Such a study will enable mapping the language functions coded in this study onto rhetorical units of discourse as their specific linguistic realizations, which would be an innovative approach to domain description.

From a theoretical standpoint, the comparison of functions in TOEFL iBT speaking responses and authentic ITA speech can be more broadly interpreted through the lens of Mohan's (1998) distinction between background knowledge and action knowledge (theory and practice), which relies on the linguistic choices of KS pairs. Our results indicate that TOEFL iBT speaking tasks successfully target the CLASSIFICATION–DESCRIPTION and EVALUATION–CHOICE pairs but do not entirely account for the full complexity of the PRINCIPLES–SEQUENCE theory–practice dyad. Overall, the study supports the suitability of the knowledge framework approach for identifying the characteristics of functional discourse in different text types.

Finally, the results of this study suggest implications for automated text analysis of learner language, both spoken and written. If test tasks are to be designed to elicit functional

language, then appropriate linguistic models are needed to automatically assess functional language. SFL is not new in natural language processing work (O'Donnell & Bateman, 2005). According to Kappagoda (2009), who argued for using SFL in automated text mining, high accuracy can be achieved in "predicting word functions in unseen text in co-training with other grammatical information, providing the basis for further grammatical and semantic text processing" (p. 1). The knowledge framework has not yet been employed in automated text analysis, largely because of the lack of fine-grained descriptors and of corpora annotated with SFL formalisms, which is needed to train and test machine learning models. Now two such corpora have been created in our study, and they could serve as the prime means for developing KS-based computational models for automated detection of functional language.

APPENDIX A: THE KNOWLEDGE FRAMEWORK

Background knowledge (theory)		Action knowledge (practice)	
Knowledge structure	Functional categories	Knowledge structure	Functional categories
CLASSIFICATION	concepts	DESCRIPTION	particulars
PRINCIPLES	rules, cause–effect	SEQUENCE	procedures, schedules
EVALUATION	evaluations	CHOICE	reason–action

Note: Adapted from "Knowledge Structures in Oral Proficiency Interviews for International Teaching Assistants" by B. Mohan, 1998, in R. Young & A.W. He (Eds.), Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency, Philadelphia, PA: John Benjamins.

APPENDIX B: FREQUENCIES OF FUNCTIONS IN THE INTERNATIONAL TEACHING ASSISTANT SPEECH CORPUS

Function	Lab	Recitation/lecture
KS1 *classifying*	237 (0.95%)	459 (1.80%)
KS1 *defining*	121 (0.48%)	125 (0.49%)
KS2 *comparing*	244 (0.98%)	260 (1.02%)
KS2 *describing*	**3,198 (12.78%)**	**3,714 (14.57%)**
KS2 *exemplifying*	169 (0.68%)	252 (0.99%)
KS2 *quantifying*	711 (2.84%)	806 (3.16%)
KS2 *spatial positioning*	548 (2.19%)	363 (1.42%)
KS3 *concluding*	600 (2.40%)	665 (2.61%)
KS3 *demonstrating cause–effect*	346 (1.38%)	314 (1.23%)
KS3 *explaining*	791 (3.16%)	930 (3.65%)
KS3 *hypothesizing*	737 (2.95%)	898 (3.52%)
KS3 *predicting*	**2,051 (8.20%)**	**1,901 (7.46%)**
KS3 *setting rules*	21 (0.08%)	25 (0.10%)
KS3 *specifying ends*	222 (0.89%)	161 (0.63%)
KS3 *specifying means*	142 (0.57%)	156 (0.61%)
KS4 *indicating order*	1,991 (7.96%)	1,544 (6.06%)
KS4 *indicating process*	1,425 (5.70%)	1,396 (5.48%)
KS4 *instructing*	**3,742 (14.96%)**	**4,014 (15.75%)**
KS4 *narrating*	42 (0.17%)	218 (0.86%)
KS4 *reporting*	71 (0.28%)	174 (0.68%)
KS5 *conceiving ideas*	294 (1.18%)	313 (1.23%)
KS5 *evaluating*	**3,462 (13.84%)**	**2,609 (10.24%)**

KS5 *making judgments*	2 (0.01%)	0 (0%)
KS6 *advising*	953 (3.81%)	952 (3.73%)
KS6 *expressing desire*	327 (1.31%)	328 (1.29%)
KS6 *expressing opinions*	**2,341 (9.36%)**	**2,596 (10.18%)**
KS6 *making choices*	12 (0.05%)	19 (0.07%)
KS6 *presenting arguments*	4 (0.02%)	13 (0.05%)
KS6 *presenting options*	217 (0.87%)	285 (1.12%)
Total	**25,021 (100%)**	**25,490 (100%)**

Note. Bolded information indicates the most frequent language functions.

APPENDIX C: FREQUENCIES OF FUNCTIONS IN THE TOEFL IBT SPEECH CORPUS

Function	Tasks 1–2	Tasks 3–6
KS1 *classifying*	23 (0.23%)	108 (0.43%)
KS1 *defining*	2 (0.02%)	313 (1.25%)
KS2 *comparing*	161 (1.61%)	247 (0.99%)
KS2 *describing*	**1,461 (14.61%)**	**3,961 (15.80%)**
KS2 *exemplifying*	177 (1.77%)	653 (2.61%)
KS2 *quantifying*	203 (2.03%)	437 (1.74%)
KS2 *spatial positioning*	44 (0.44%)	105 (0.42%)
KS3 *concluding*	340 (3.40%)	849 (3.39%)
KS3 *demonstrating cause–effect*	314 (3.14%)	982 (3.92%)
KS3 *explaining*	**914 (9.14%)**	**1,841 (7.34%)**
KS3 *hypothesizing*	176 (1.76%)	701 (2.80%)
KS3 *predicting*	**768 (7.68%)**	**2,731 (10.89%)**
KS3 *setting rules*	1 (0.01%)	7 (0.03%)
KS3 *specifying ends*	41 (0.41%)	66 (0.26%)
KS3 *specifying means*	84 (0.84%)	182 (0.73%)
KS4 *indicating order*	**1,091 (10.91%)**	**2,306 (9.20%)**
KS4 *indicating process*	**664 (6.64)**	**1,904 (7.60%)**
KS4 *instructing*	29 (0.29%)	106 (0.42%)
KS4 *narrating*	27 (0.27%)	156 (0.62%)
KS4 *reporting*	15 (0.15%)	442 (1.76%)
KS5 *conceiving ideas*	124 (1.24%)	319 (1.27%)
KS5 *evaluating*	**1,616 (16.16%)**	**3,091 (12.33%)**
KS5 *making judgments*	0 (0%)	4 (0.02%)
KS6 *advising*	269 (2.69%)	841 (3.36%)

KS6 *expressing desire*	63 (0.63%)	275 (1.10%)
KS6 *expressing opinions*	**1,251 (12.51%)**	**1,734 (6.92%)**
KS6 *making choices*	43 (0.43%)	165 (0.66%)
KS6 *presenting arguments*	4 (0.04%)	36 (0.14%)
KS6 *presenting options*	98 (0.98%)	505 (2.01%)
Total	**10,003 (100%)**	**25,067 (100%)**

APPENDIX D: FREQUENCIES OF FUNCTIONS IN THE TOEFL IBT AND INTERNATIONAL TEACHING ASSISTANT SPOKEN CORPORA

Function	TOEFL iBT corpus		ITA spoken corpus	
	Frequency	Percentage	Frequency	Percentage
KS1 *classifying*	131	0.37%	696	1.38%
KS1 *defining*	315	0.90%	246	0.49%
KS2 *comparing*	408	1.16%	504	1.00%
KS2 *describing*	**5,422**	**15.46%**	**6,912**	**13.68%**
KS2 *exemplifying*	830	2.37%	421	0.83%
KS2 *quantifying*	640	1.82%	1,517	3.00%
KS2 *spatial positioning*	149	0.42%	911	1.80%
KS3 *concluding*	1,189	3.39%	1,265	2.50%
KS3 *demonstrating cause–effect*	1,296	3.70%	660	1.31%
KS3 *explaining*	**2,755**	**7.86%**	1,721	3.41%
KS3 *hypothesizing*	877	2.50%	1,635	3.24%
KS3 *predicting*	**3,499**	**9.98%**	**3,952**	**7.82%**
KS3 *setting rules*	8	0.02%	46	0.09%
KS3 *specifying ends*	107	0.31%	383	0.76%
KS3 *specifying means*	266	0.76%	298	0.59%
KS4 *indicating order*	**3,397**	**9.69%**	**3,535**	**7.00%**
KS4 *indicating process*	**2,568**	**7.32%**	2,821	5.58%
KS4 *instructing*	135	0.38%	**7,756**	**15.36%**
KS4 *narrating*	183	0.52%	260	0.51%
KS4 *reporting*	457	1.30%	245	0.49%
KS5 *conceiving ideas*	443	1.26%	607	1.20%
KS5 *evaluating*	**4,707**	**13.42%**	**6,071**	**12.02%**
KS5 *making judgments*	4	0.01%	2	0.00%
KS6 *advising*	1,110	3.17%	1,905	3.77%
KS6 *expressing desire*	338	0.96%	655	1.30%

KS6 *expressing opinions*	**2,985**	**8.51%**	**4,937**	**9.77%**
KS6 *making choices*	208	0.59%	31	0.06%
KS6 *presenting arguments*	40	0.11%	17	0.03%
KS6 *presenting options*	603	1.72%	502	0.99%
Total	**35,070**	**100%**	**50,511**	**100%**

Note. ITA = international teaching assistant.

REFERENCES

Axelson, E. R., & Madden, C. G. (1994). Discourse strategies for ITAs across instructional contexts. In C. Madden & C. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 153–186). Alexandria, VA: TESOL.

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

Bailey, C. M. (1983). Foreign teaching assistants at U.S. universities: Problems in interaction and communication. *TESOL Quarterly*, *17*, 308–310. https://doi.org/10.2307/3586658

Bailey, C. M. (1984). The "foreign TA problem." In C. M. Bailey, F. Pialorsi, & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 3–16). Washington, DC: National Association for Foreign Student Affairs.

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/scl.28

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

Brown, K., Fishman, P., & Jones, N. (1990). *Legal and policy issues in the language proficiency assessment of international teaching assistants*. Houston, TX: University of Houston Law Center.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 Speaking framework: A working paper* (Research Memorandum No. RM-00-06). Princeton, NJ: Educational Testing Service.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Clausen, S. E. (1998). *Applied correspondence analysis*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412983426

Christie, F. (2002). *Classroom discourse analysis*. London: Continuum.

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* (Research Memorandum No. RM-04-05). Princeton, NJ: Educational Testing Service.

Dick, R. C., & Robinson, B. M. (1994). Oral English proficiency requirements for ITAs in U.S. colleges and universities: An issue in speech communication. *JACA, 2*(1), 77–86.

Farnsworth, T. (2012, April). *TOEFL iBT speaking for ITA certification: State of practice and ongoing validation questions*. Paper presented at the annual meeting of the Language Testing Research Colloquium, Princeton, NJ.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*, 274–291. https://doi.org/10.1080/15434303.2013.769548

Farnsworth, T. F. (2014). Assessing the oral English abilities of international teaching assistants in the USA. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 471–483). New York, NY: John Wiley.

Field, A. (2013). *Discovering statistics using SPSS* (4th ed.). London, England: Sage.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*, 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Gaithersburg, MD: Advanced Analytics.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. Baltimore, MD: University Park Press.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (Research Memorandum No. RM-00-03). Princeton, NJ: Educational Testing Service.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*, 135–170. https://doi.org/10.1207/s15366359mea0203_1

Kappagoda, A. (2009). *The use of systemic–functional linguistics in automated text mining*. Edinburgh, Australia: Defense Science and Technology Organization.

Lazaraton, A., & Wagner, S. (1996). *The revised Test of Spoken English (TSE): Discourse analysis of native speaker and nonnative speaker data* (Research Memorandum No. RM-96-10). Princeton, NJ: Educational Testing Service.

Levis, J., Levis, G. M., & Slater, T. (2012). Written English into spoken: A functional discourse analysis of American, Indian, and Chinese TA presentations. In G. Gorsuch (Ed.), *Working theories for teaching assistant development: Time-tested and robust theories, frameworks, and models for TA and ITA learning* (pp. 529–573). Stillwater, OK: New Forums.

Liao, S. (2009). Variation in the use of discourse markers by Chinese teaching assistants in the US. *Journal of Pragmatics, 41*, 1313–1328. https://doi.org/10.1016/j.pragma.2008.09.026

Lijffijt, J., & Gries, S. T. (2012). Correction to "Dispersions and adjusted frequencies in corpora." *International Journal of Corpus Linguistics, 17*, 147–149. https://doi.org/10.1075/ijcl.17.1.08lij

Lim, H., Kim, H., Behney, J., Reed, D., Ohlrogge, A., & Lee, J. E. (2012, March). *Validating the use of iBT Speaking scores for ITA screening.* Paper presented at the TESOL Annual Convention and Exhibit, Philadelphia, PA.

Madden, C. G., & Myers, C. L. (1994). *Discourse and performance of international teaching assistants.* Alexandria, VA: Teachers of English to Speakers of Other Languages.

Mohan, B. (1998). Knowledge structures in oral proficiency interviews for international teaching assistants. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 173–204). Philadelphia, PA: John Benjamins. https://doi.org/10.1075/sibil.14.12moh

Mohan, B. A. (1986). *Language and content.* Reading, MA: Addison-Wesley.

Mohan, B. A. (1989). Knowledge structures and academic discourse. *Word, 40*, 99–115. https://doi.org/10.1080/00437956.1989.11435799

Mohan, B.A. (2007). Knowledge structures in social practices. In J. Cummins & C. Davison (Eds.). *International handbook of English language teaching* (pp. 303-315). London: Kluwer Academic Publishers.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2014). 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*, 39–62. https://doi.org/10.1177/0265532214538014

O'Donnell, M., & Bateman, J. (2005). SFL in computational contexts: A contemporary history. In R. Hasan, C. M. I. M. Matthiessen, & J. Webster (Eds.), *Continuing discourse on language: A functional perspective* (Vol. 1, pp. 343–382). London, England: Equinox.

Plakans, B. S., & Abraham, R. G. (1990). The testing and evaluation of international teaching assistants. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (pp. 68–81). Washington, DC: NAFSA.

Rose, D. (2017). Pedagogic register analysis: Mapping choices in teaching and learning. *Functional Linguistics, 5*, 1–33.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *Identifying the reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (Research Memorandum No. RM-01-03). Princeton, NJ: Educational Testing Service.

Rounds, P. L. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly, 21*, 643–671. https://doi.org/10.2307/3586987

Ruderman, A. (2000, December 27). Colleges are moving to ensure English fluency in teaching assistants. *New York Times*. Retrieved from https://www.nytimes.com/2000/12/27/nyregion/colleges-are-moving-to-ensure-english-fluency-in-teaching-assistants.html

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL iBT test. *Language Testing, 26*, 5–30. https://doi.org/10.1177/0265532208097335

Slater, T., & Gleason, J. (2011). *Integrating language and content: The knowledge framework*. Retrieved from https://lib.dr.iastate.edu/engl_conf/7

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunananthan, . . . Bergman, H. (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of Clinical Epidemiology, 63*, 638–646. https://doi.org/10.1016/j.jclinepi.2009.08.008

Swales, J. M. (1981). *Aspects of article introductions* (Aston ESP Research Report No. 1). Birmingham, England: University of Aston, Language Studies Unit.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*, 713–729. https://doi.org/10.2307/3586870

Wagner, E. (2016). *A Study of the Use of the TOEFL iBT® Test Speaking and Listening Scores for International Teaching Assistant Screening* (Research Memorandum No. RR-16-18). Princeton, NJ: Educational Testing Service.

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly, 26*, 693–711. https://doi.org/10.2307/3586869

Wylie, E. C., & Tannenbaum, R. J. (2006). *TOEFL Academic Speaking test: Setting a cut score for international teaching assistants* (Research Memorandum No. RM-06-01). Princeton, NJ: Educational Testing Service.

Xi, X. (2007). Validating TOEFL iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly, 4*, 318–351. https://doi.org/10.1080/15434300701462796

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL Speaking scores for ITA screening and setting standards for ITAs* (TOEFL iBT Research Report No. TOEFLiBT-

03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02088.x

Xi, X. (2015). *Language constructs revisited for practical test design, development and validation*. Paper presented at the Language Testing Research Colloquium, Toronto, Canada.

Yelland, P. M. (2010). An introduction to correspondence analysis. *Mathematica Journal, 12*, 1–21. Retrieved from http://www.mathematica-journal.com/2010/09/an-introduction-to-correspondence-analysis

NOTES

[1] Many universities have developed their own tests, others have adopted or adapted the American Council on the Teaching of Foreign Languages Oral Proficiency Interview, and some have been experimenting with computer-based tests, such as the Pearson Test of English Academic (PTE Academic).

[2] ITA certification is higher stakes because the proficiency level may determine the stipend amount, progress in the academic program, eligibility for certain courses (e.g., Preparing Future Faculty at Iowa State University), etc.

[3] Five other sets were used in the training of annotators.

[4] CA was chosen because it permits occurrences of individual responses in multiple cells of cross-tabulated data (Hair, Black, Babin, & Anderson, 2009). Additionally, CA is free from the assumptions of normal distribution, homoscedasticity, and independence of residuals that are mandatory for methods like MANOVA and loglinear regression. Using such methods requires that each person, item, or entity contributes to only one cell of the contingency table (Field, 2013). In our study, each individual participant contributed to multiple cells in our contingency tables. Specifically, individual TOEFL test takers contributed multiple responses to Tasks 1–2 and Tasks 3–6. Some lab and recitation sessions were taught by the same ITA participants. Both TOEFL and ITA performances contained a variety of KS and functions in the same text.

[5] Scaling of the perception map is also important, because the normalization of row profiles can distort the normalization of column profiles, and vice versa. In our study, symmetrical normalization was used for proper exploration of the data.

[6] This technique is used to identify the lexical items that show distributional patterns distinctive from their expected proportions in the corpus. We adopted it to identify the functions that demonstrate distributional patterns from what is expected given the total use of functions in each corpus subset.

Suggested citation:

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* (TOEFL Research ReportNo. RR-85). Princeton,NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12233