ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–19

Discourse Characteristics of Writing and Speaking Task Types on the *TOEFL iBT*® Test: A Lexico-Grammatical Analysis

Douglas Biber

Bethany Gray

March 2013

Discourse Characteristics of Writing and Speaking Task Types on the *TOEFL iBT*® Test: A Lexico-Grammatical Analysis

Douglas Biber
Northern Arizona University, Flagstaff

Bethany Gray
Iowa State University, Ames

RR-13-04

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Abstract

One of the major innovations of the *TOEFL iBT*® test is the incorporation of integrated tasks complementing the independent tasks to which examinees respond. In addition, examinees must produce discourse in both modes (speech and writing). The validity argument for the TOEFL iBT includes the claim that examinees vary their discourse in accordance with these considerations as they become more proficient in their academic language skills (the explanation inference). To provide evidence in support of this warrant, we undertake a comprehensive lexico-grammatical description of the discourse produced in response to integrated versus independent tasks, across the spoken and written modes, by test takers from different score levels.

Discourse descriptions at several linguistic levels are provided, including vocabulary profiles, collocational patterns, the use of extended lexical bundles, distinctive lexico-grammatical features, and a multidimensional (MD) analysis that describes the overall patterns of linguistic variation. In sum, we undertake a comprehensive linguistic analysis of the discourse of TOEFL iBT responses, interpreting observed linguistic patterns of variation relative to three parameters that are relevant in the TOEFL iBT context: mode, task type, and score level of test takers.

Key words: task variation, spoken/written differences, proficiency levels, vocabulary, grammatical variation, multi-dimensional analysis

TOEFL® was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT®. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2012-2013) members of the TOEFL Committee of Examiners are:

John M. Norris - Chair	Georgetown University
Maureen Burke	The University of Iowa
Yuko Goto Butler	University of Pennsylvania
Barbara Hoekje	Drexel University
Ari Huhta	University of Jyväskylä, Finland
Eunice Eunhee Jang	University of Toronto, Canada
James Purpura	Teachers College, Columbia University
John Read	The University of Auckland, New Zealand
Carsten Roever	The University of Melbourne, Australia
Steve Ross	University of Maryland
Norbert Schmitt	University of Nottingham, UK
Ling Shi	University of British Columbia, Canada

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Table of Contents

List of Tables

Page

List of Figures

Page

1. Background

Numerous studies have described linguistic characteristics of the discourse produced by different learner groups in different contexts. One important research objective of these studies has been to investigate the linguistic characteristics of discourse associated with different developmental stages or different proficiency levels, while many of the studies have additionally considered differences across task types. Such research provides the foundation for practice in language assessment and teaching.

Within the context of the *TOEFL iBT*® test, both objectives are important. Thus, the validity argument for the TOEFL iBT begins with the *domain description* to document the range of spoken and written tasks that students encounter in university settings (see Chapelle, Enright, & Jamieson, 2008, pp. 19–21; Enright & Tyson, 2008, p. 3). Building upon that research, the second stage in the validity argument is the development of appropriate tasks for the exam itself (including independent and integrated tasks in both the spoken and written modes) and the development of appropriate scoring rubrics for the discourse produced in those tasks (Enright & Tyson, 2008, Table 1). The validity argument is then further supported by the *explanation inference* that "expected scores are attributed to a construct of academic language proficiency" (Chapelle et al., 2008, p. 20). Evidence to support this proposition—the focus of the present project—comes from linguistic analyses of the discourse produced by examinees across task types and across score levels. That is:

> For writing and speaking tasks, the characteristics of the discourse that test takers produce is expected to vary with score level as described in the holistic rubrics that raters use to score responses. Furthermore, the rationale for including both independent and integrated tasks in the TOEFL iBT speaking and writing sections was that these types of tasks would differ in the nature of discourse produced, thereby broadening representation of the domain of academic language on the test. (Enright & Tyson, 2008, p. 5)

Two previous studies carried out pilot investigations of this type. Cumming et al. (2005, 2006) analyzed the written independent and integrated responses from 36 examinees on a prototype version of the TOEFL iBT. That study found significant differences across both score levels and task types for a range of discourse characteristics including length of response, lexical diversity, T-unit (clause) length, grammatical accuracy, use of source materials, and

paraphrasing. Brown, Iwashita, and McNamara (2005) focused on spoken responses but similarly considered differences across score levels and independent versus integrated tasks. That study found weaker patterns of linguistic variation associated with fluency, vocabulary, grammatical accuracy, and complexity.

The present project complements these previous studies by focusing on the lexico-grammatical characteristics of examinee responses on the TOEFL iBT, considering a much larger inventory of linguistic features than in previous research, and analyzing a larger corpus of exam responses. Similar to the two studies cited above, though, this study focuses on the primary considerations relevant to the explanation proposition of the *TOEFL*® validity argument: analysis of the discourse characteristics of responses produced across task types, by examinees from different score levels. Thus, the study investigates three major research questions:

1. Do test takers systematically vary the linguistic characteristics of discourse produced in the spoken versus written modes across different task types? If so, how?
2. In what ways do exam scores correspond to systematic linguistic differences in the discourse produced by test takers?
3. How does the relationship between linguistic discourse characteristics and score level vary across the spoken/written modes and/or task types?

The first question adopts a register perspective, disregarding proficiency level. The issue here is the extent to which the texts produced by test takers reflect awareness of the linguistic differences across the spoken and written modes and between integrated versus independent task types; that is, have test takers developed proficiency in the appropriate use of linguistic features (e.g., vocabulary and grammar) associated with spoken versus written language, and with integrated versus independent tasks?

The second question concerns the ways in which TOEFL iBT score levels correspond to systematic linguistic differences in the language produced by test takers. As noted above, the analytical focus of this study is on the lexical and grammatical characteristics of the discourse produced by the test taker groups.

Finally, the third question brings the first two perspectives together, considering the interactions of score levels, mode, and task differences as predictors of the patterns of lexico-grammatical variation.

To address these research questions, this study presents an empirical linguistic analysis of a corpus of TOEFL iBT exam responses, providing a comprehensive lexico-grammatical description of the discourse of exam responses. As set out in the TOEFL validity argument, the linguistic characteristics of examinees' discourse are predicted to vary in systematic ways with task type, mode, and score level. The investigations reported below are a first step toward describing those relationships.

In Section 2, we briefly summarize previous research that has described the use of a variety of lexico-grammatical features in the spoken and/or written production of English language learners. In Section 3, we introduce the TOEFL iBT context and corpus, followed by a description of our research design and methods in Section 4. We then present and discuss the results of our investigations into the lexico-grammatical characteristics of spoken and written TOEFL iBT discourse in Section 5, and conclude with a brief summary and discussion of implications for the TOEFL iBT in Section 6.

2. A Brief Survey of Previous Research

Several previous studies have described linguistic characteristics of the discourse produced by different learner groups in attempts to document the linguistic changes associated with language development and different levels of proficiency. Table 1 surveys many of the most important of these studies. Rather than undertaking an exhaustive survey of previous research, the purposes here are to illustrate the wide range of discourse characteristics that have been investigated in these studies.

Table 1

Features Investigated in Spoken and Written Language Production, as Related to Proficiency and/or L1 (Language 1)

Category	Study	Linguistic features	Findings
Lexical features	Grant & Ginther (2000)	Lexical specificity (i.e., type/token ratios, word length), conjuncts, hedges, amplifiers, emphatics, demonstratives, downtoners	As proficiency increased, lexical specificity increased (i.e., longer and more varied words were used). Uses of conjuncts, amplifiers, emphatics, demonstratives, and downtoners increased.
	Ferris (1994)	Word length, special lexical classes	Higher proficiency writers used more specific lexical classes (e.g., emphatics, hedges). Word length was one of the most significant predictors of holistic scores assigned to essays.
	Engber (1995)	Lexical variation (i.e., type/token variation), error-free variation, percentage of lexical error, lexical density	Lexical variation and holistic scores assigned to compositions were highly correlated. Error-free variation and holistic scores were also highly correlated.
	Jarvis, Grant, Bikowski, & Ferris (2003)	Mean word length, type/token ratio, conjuncts, hedges, amplifiers, emphatics, downtoners	Cluster analysis revealed that clusters of highly-rated texts varied little in terms of lexical diversity and use of conjuncts.
	Jarvis (2002)	Lexical diversity (type/token ratios)	Results indicated that lexical diversity did contribute to writing quality, but this relationship was dependent on the writer's L1.
	Laufer & Nation (1995)	Lexical frequency profiles based on proportions of UWL, GSL 1K, GSL 2K, and offlist words	Lexical frequency profiles discriminate between proficiency levels and correlate well with other measures of vocabulary size with lower proficiency learners using higher proportion of high frequency words and higher proficiency learners using more words from the less frequent or offlist words.

Category	Study	Linguistic features	Findings
	Cumming et al. (2005)	Lexical sophistication (word length, type/token ratios)	All proficiency levels tended to use longer words in integrated tasks. Higher proficiency learners had higher type/token ratios.
Grammatical and syntactic features	Grant & Ginther (2000)	Nouns, nominalizations, personal pronouns, verbs, modals, adjectives, adverbs, prepositions, articles, subordination, complementation, relative clauses, adverbial subordination, passives	The frequency of several features increased with proficiency: nominalizations, modals, first and third-person pronouns, more varied verb tense uses, passives, subordination, and complementation.
	Ferris (1994)	Verb tenses, pronouns, adverbials, modals, negation, coordination, prepositional phrases, definite article reference, passives, relative clauses, stative forms, coordination, participials, coherence features	Higher proficiency writers produced more of the more difficult syntactic constructions such as stative forms, participial constructions, relative clauses, and adverbial clauses. Higher proficiency writers used more passives, existential *there*, preposed adverbials, clefts, topicalizations to show "pragmatic sensitivity" and "promote textual coherence" (p. 418).
	Jarvis et al. (2003)	Nouns and nominalizations, pronouns, adverbials, prepositions, definite articles, present tense verbs, stative verb *be*, passives, adverbial subordination, relative clauses, complementation	Using cluster analysis, Jarvis et al. found that judgments of essay quality depended on how linguistic features were used together rather than on the use of individual features. Clusters of highly rated texts could differ in terms of mean word length, nouns and nominalizations, prepositions, and present tense verbs. Highly rated texts varied less in terms of text length and lexical diversity.

Category	Study	Linguistic features	Findings
	Cumming et al. (2005)	Syntactic complexity (clauses per T-unit, words per T-unit)	More proficient learners produced more words per T-unit. The mean number of clauses per T-unit differed across task types, but no difference was found across proficiency level.
	Wolfe-Quintero, Inagaki, & Kim (1998)	Linguistic complexity (clauses per T-unit, dependent clause ratio)	Surveyed previous empirical research on complexity and language development, identifying the most promising lexico-grammatical complexity features.
	Ortega (2003)	Syntactic complexity (especially T-unit measures)	Surveyed 25 previous studies of syntactic complexity in L2 writing.
Rhetorical structure	Hirose (2003)	Deductive vs. inductive organizational patterns	L2 organization scores did not significantly correlate with L1 organization scores. Choice of organizational pattern (deductive or inductive) did not contribute alone to the evaluation of organization; rather, factors such as coherence between/within paragraphs also influenced how organization was evaluated.
	Kubota (1998)	Location of main idea, rhetorical pattern/ organization	About half of the participants used similar rhetorical patterns in L1 and L2 essays. A positive correlation was found between L1 and L2 organization, indicating that writing proficiency in the L2 may be related to writing proficiency in the L1. Little evidence for transfer of rhetorical patterns from L1 to L2.
	Coffin (2004)	Argument structure	Lower-level learners tend to use arguments composed using exposition structures rather than a discussion-based argument.
	Cumming et al. (2005)	Quality of argument structure, orientations to source evidence	In integrated tasks, highly proficient learners often summarized and synthesized information from source materials, while learners in the midproficiency ranges used more phrases directly from the prompts.

Category	Study	Linguistic features	Findings
Formulaic language	Cortes (2004)	Lexical bundles	Student writers rarely used lexical bundles used by professional writers.
			When student writers did use the target bundles, they did not use them in the same way as professional writers.
	Hyland (2008)	Lexical bundles	Student writers employed a higher proportion of lexical bundles that outline research procedures as compared to published writers, which may be related to the nature of the student genres as a way of displaying knowledge.
			Student writers tended to avoid participant-oriented bundles, perhaps due to influences from the L1 culture and educational experience.
	Howarth (1998)	Collocational density	Advanced learners are able to internalize restricted collocation or semi-idioms, but there are too many less restricted combinations to learn as unitary items.
	Altenberg & Granger (2001)	Grammatical patterns, meanings, collocations of *make*	When compared to native English-speaking student writers, advanced level learners underused delexical *make* and used inappropriate collocations.

Note. L2 = Language 2; UWL = University Word List, 808 common word families in academic writing; GSL 1K = 1,000 most frequent words in the General Service List; GSL 2K = second 1,000 most frequent words in the General Service List.

As Table 1 shows, previous research has investigated the use of linguistic features at all grammatical levels associated with English language development. Thus, the features considered in previous studies include the following:

- Lexical features (e.g., type/token ratio, average word length, use of academic and general service words)
- Word classes and general grammatical features (e.g., nouns, nominalizations, adjectives)
- Grammatical features that specifically relate to linguistic complexity (e.g., relative clauses, adverbial clauses, average T-unit length, depth of embedding)
- Rhetorical organization (e.g., move structure of written essays)
- Formulaic language (e.g., collocational patterns, lexical bundles)

It is worth noting that (almost) all lexico-grammatical characteristics of English are useful indicators of register and communicative task differences (see Biber & Conrad, 2009, especially Chapter 3). By extension, it is likely that these same linguistic features are associated with language development and differences in language proficiency. These relationships exist because lexico-grammatical features are functional and are used to differing extents in association with the communicative purposes and production circumstances of different registers. For example, writing development entails the productive use of lexico-grammatical features that are not naturally acquired in speech, including an increased range of vocabulary, increased range of grammatical structures (e.g., nonfinite relative clauses), and increased complexity in noun phrase constructions (especially with phrasal modifiers). Language development in speech follows a different progression and is focused more on clausal (rather than phrasal) modification and vocabulary diversification. As a result, the linguistic features listed in Table 1 represent a relatively comprehensive subset of the possible lexico-grammatical characteristics of English discourse.

Beginning in the 1970s, numerous researchers have focused on L2 (Language 2) writing development with an overt focus on the linguistic structures used in student texts (see, e.g., Cooper, 1976; Ferris & Politzer, 1981; Flahive & Snow, 1980; Gipps & Ewen, 1974). This trend has continued to the present time, so that it is common now to find second language researchers who focus on "measures of fluency, accuracy, and complexity" in second language writing (as in the title of the 1998 book by Wolfe-Quintero, Inagaki, & Kim). More recent studies include

Brown et al. (2005), Ellis and Yuan (2004), Larsen-Freeman (2006), and Nelson and Van Meter (2007).

Across these decades, when writing development research has focused on the linguistic description of student texts, one of the key concerns has been the analysis of grammatical complexity. Most of these studies have adopted a deductive approach, beginning with an a priori definition of grammatical complexity as elaborated structures added on to simple phrases and clauses (see, e.g., Purpura, 2004, p. 91; Willis, 2003, p. 192). Specifically, most studies of L2 writing development have relied on T-unit-based measures, based on the average length of structural units and/or the extent of clausal subordination, assuming that longer units and more subordination reflect greater complexity. The early reliance on clausal subordination (and T-unit-based measures) is documented by Wolfe-Quintero et al. (1998), and subsequent studies have continued this practice (e.g., Ellis & Yuan, 2004; Larsen-Freeman, 2006; Li, 2000; Nelson & Van Meter, 2007; Norrby & Håkansson, 2007). The two previous studies of TOEFL iBT spoken and written responses (Brown et al., 2005; Cumming et al., 2006) have similarly relied heavily on T-unit based measures for their analyses of syntactic complexity. Ortega (2003) provided strong confirmation that current research continues to employ these same two measures, based on a meta-analysis of empirical research on grammatical complexity in college level ESL/EFL writing. Of the 27 studies included in her survey, 25 studies relied on the mean length of T-unit (MLTU) to measure grammatical complexity, while 11 studies used the related measure of dependent clauses per T-unit (C/TU). No other measure was used widely across these studies.

Biber and Gray (2010) and Biber, Gray, and Poonpon (2011) challenged this pervasive practice, arguing instead that phrasal embedding is a much more important indicator of advanced writing development than clausal embedding; these structures function mostly as noun phrase modifiers, such as attributive adjectives, premodifying nouns, prepositional phrase postmodifiers, and appositive noun phrase postmodifiers. Based on corpus analysis, these two studies show that there is no empirical basis for treating all dependent clauses as a single construct reflecting complexity. Rather, different types of dependent clauses are distributed in quite different ways across spoken and written registers, indicating that they represent quite different types of structural complexity. Thus, for the purposes of the present research project, the full range of linguistic features associated with both clausal embedding and phrasal embedding is considered (see Research Design and Methods below).

Corpus-based research on English grammar has provided the foundation for much of the previous research on discourse produced by learners at different proficiency levels. In fact, the linguistic features investigated in many developmental studies have been adopted directly from earlier grammatical studies that analyze differences across spoken and written registers (e.g., Biber, 1988). The *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad, & Finegan, 1999) documented systematic patterns of variation for the grammatical features listed in Table 1, showing how their frequency of use corresponds to the discourse requirements of different spoken and written registers (see also Biber & Conrad, 2009). Based on previous research of this type, we are able to interpret observed frequency differences in the use of linguistic features across exam responses in relation to the underlying communicative functions served by these features in discourse.

The logic underlying our general approach (which has also been widely adopted in previous research) can be summarized as follows:

1. Previous corpus-based research has shown in great detail how the grammatical characteristics of spoken discourse are dramatically different from the grammatical characteristics of written discourse (especially academic writing).
2. These differences are mostly due to the fact that linguistic variation is functional: speakers and writers rely on different lexico-grammatical characteristics because they produce discourse under different circumstances for different communicative purposes and tasks.
3. English-language learners must learn to control the discourse characteristics of academic writing to succeed at the university level; as a result, language development and increased language proficiency are strongly associated with increased control over the lexico-grammatical resources associated with academic writing, including appropriate use of these features across different communicative tasks.
4. Numerous empirical studies have directly documented the association of these core lexico-grammatical features with language development and proficiency.
5. Taken together, these studies indicate that any lexico-grammatical feature that distinguishes among spoken and written registers will probably also be an important indicators of language development and proficiency. At the same time, these studies indicate that no single developmental parameter exists. Rather, different sets of

discourse characteristics have different functional associations and, as a result, are associated with different types of development.

Building on the same general approach employed in these previous studies, the current project investigates the full set of lexico-grammatical characteristics in the discourse produced by TOEFL iBT test takers at different score levels, also considering differences in spoken versus written language production and differences in independent versus integrated task types. The resulting descriptions provide a comprehensive linguistic description of the discourse produced in the TOEFL iBT context.

3. Overview of the TOEFL iBT Context and Corpus

The project employs a series of corpus-based analyses to describe the discourse patterns of linguistic variation and use among TOEFL iBT responses across multiple external parameters of variation (score level, task type, and mode). This section details the context of the TOEFL iBT and the corpus utilized in the study.

Each TOEFL iBT exam consists of six spoken responses and two written responses, representing independent and integrated task types in each mode. Independent tasks require test takers to give their opinion about a topic with no supporting materials, while integrated tasks require test takers to describe or explain information based on reading and listening passages that they first comprehend. The four major categories differ with respect to several parameters, summarized in Table 2. The full prompts and questions for these exams are given in Appendix A.

Table 2

Summary of Some Major Situational Characteristics of the TOEFL iBT Text Categories

Text category	Mode of production	Planning/editing time	Support from external text	Communicative purposes
Spoken independent	Speech	Minimal: 15-second planning time; 45-second response	None	Give personal opinions based on individual personal experiences.
Spoken integrated	Speech	Little: 20-second planning time; 60-second response Preplanning is possible while reading and listening to the external texts	Yes—both written and spoken texts	Describe/summarize the content of the external texts; sometimes also take a position.
Written independent	Writing	Considerable: 30 minutes to plan and write	None	Give personal opinions about life choices or general issues.

Text category	Mode of production	Planning/editing time	Support from external text	Communicative purposes
Written integrated	Writing	Considerable: Pre-planning is possible while reading and listening to the external texts.	Yes—both written and spoken texts	Describe/summarize the content of the external texts.

Each individual response had been previously assigned a holistic score. TOEFL iBT raters receive extensive training in the use of evaluation rubrics specific to each mode and task type, incorporating a range of discourse and content characteristics (see Appendix B). Raters consider a wide range of factors, including the overall content, relevance of the response to the assigned task, fluency (in speech), coherence and clear progression of ideas, word choice, and control of grammatical structures (see, e.g., Lumley, 2002). Similar to instructors in university courses, raters consider this range of factors to determine a single overall quality score for each response. As a result, responses at a given score level can differ considerably in their use of particular linguistic features (see Jarvis et al., 2003).

Spoken and written responses are scored using different scales: The spoken TOEFL iBT is graded on a 4-point scale (1, 2, 3, 4), while the reported scores for the written TOEFL iBT use a 9-point scale (1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5). Thus, both the overall magnitude of the two scales as well as the degree of possible variability differ between the two modes, making it impossible to directly compare the two in a statistical test. To address this problem, the written scores were transformed to a 4-point scale as shown in Table 3.

Table 3

Transformation of Scores for Written Responses on the TOEFL iBT Test

Original score	Transformed score
1.0 , 1.5, 2.0	1
2.5, 3.0	2
3.5, 4.0	3
4.5, 5.0	4

We considered two criteria for this transformation: the need for identical scales for the spoken and written responses, and achieving roughly comparable representation of score levels. Table 4 shows the overall composition of the corpus following score transformation. (Note that the sample is missing one spoken integrated response.)

The spoken responses were provided in individual sound (.spx) files, while the written responses were provided in individual text (.txt) files. We undertook a comprehensive process to prepare the responses for analysis, including transcribing all spoken responses. The corpus preparation process and our analysis procedures are described in Section 4.

4. Research Design and Methods

This study was broken down into four major procedural steps. The first step in the analysis was to prepare the corpus for analysis, including the transcription of the speaking responses and the automatic and interactive grammatical annotation of the complete corpus. The second major step was to conduct the linguistic analyses, which included investigations of lexis, grammar, and lexico-grammar. These analyses utilized existing computer programs as well as the development of new software analysis tools, resulting in quantitative rates of occurrence for each linguistic characteristic in each text. The third step involved statistical analyses of the quantitative data from the linguistic analyses, while the final step required qualitative interpretations of the patterns of variation. An overview of the major steps in the analysis is provided in Table 5. Each of these procedural steps is described in more detail in the following sections.

Table 4

Total Corpus Composition

Task	Score level	Number of texts	Number of words	Mean text length	Min. text length	Max. text length
Spoken independent tasks	1	36	1,778	49.4	13	88
(2 responses per exam)	2	368	27,968	76.1	29	140
	3	440	41,228	93.7	49	172
	4	116	12,447	107.3	71	164
Subtotal		960	83,421			
Spoken integrated tasks	1	105	6,121	58.3	12	116
(4 responses per exam)	2	764	73,115	95.7	17	195
	3	826	99,120	119.9	48	213
	4	224	31,248	139.5	85	212
Subtotal		1,919	20,9604			
Written independent	1	46	9,890	215.0	61	351
tasks	2	177	51,118	288.8	160	507
(1 response per exam)	3	155	52,452	338.4	206	549
	4	102	39,300	385.3	261	586
Subtotal		480	152,760			

Task	Score level	Number of texts	Number of words	Mean text length	Min. text length	Max. text length
Written integrated tasks	1	128	21,286	166.3	45	293
(1 response per exam)	2	118	23,683	200.7	102	303
	3	122	25,962	212.8	108	367
	4	112	26,264	234.5	145	388
Subtotal		480	97,195			
Total		3,839	542,980			

Note. The sample is missing one spoken integrated response.

4.1. Corpus Preparation: Phase 1

The first step was to transcribe all spoken responses. To begin this task, we established procedures for transcriber training and transcription conventions. Initially, 19 transcribers were trained, although only 14 individuals chose to actually transcribe texts after training. The training process involved an initial training meeting, detailed checking of trial transcriptions, meetings to discuss questions/problems, and repetition of the training cycle. After an individual transcriber had demonstrated his or her ability to consistently achieve accurate transcriptions, spot checks were carried out on the reliability of the transcriptions (one file per batch of 18 sound files). Approximately 6% of the total transcribed texts were evaluated for quality and reliability.

Table 5

Major Procedural Steps in the Analysis

	Procedural steps
1. Corpus preparation	Transcription of the spoken corpus
	Recoding of score level in written responses
	Automatic grammatical annotation (tagging) of the spoken and written corpora
	Evaluation of automatic tag accuracy
	Development of additional computer programs for more accurate automatic tagging
	Interactive hand-editing of problematic grammatical features
	Re-evaluation of automatic tag accuracy
	Verification of required minimum text length for quantitative lexico-grammatical analyses
2. Linguistic analyses	Vocabulary and collocational patterns
	Extended lexical phrases

Procedural steps	
	Lexico-grammatical features
3. Quantitative/statistical analyses	Chi-squared and log-likelihood methods to compare word distributions for the vocabulary, collocational, and lexical-phrase analyses
	Exploratory correlations for preliminary investigation of the association between lexico-grammatical features and TOEFL iBT score
	General linear models for more detailed investigations of particular linguistic features associated with score level, task type, and individual test taker (in both the spoken and written modes)
	Overall textual patterns (multidimensional analysis): Factor analysis, with posthoc comparisons across modes, score levels, task types, and individual test takers
4. Interpretation	Qualitative interpretation of functions of linguistic features

4.2. Corpus Preparation: Phase 2 – Annotation & Evaluation

The second step in the corpus preparation was the linguistic analysis of both spoken and written texts. This step began with the application of a computational tool—the Biber Tagger—that automatically annotates texts for a wide range of lexico-grammatical characteristics. The Biber Tagger has both probabilistic and rule-based components, uses multiple large-scale dictionaries, and runs under Windows; this tagger has been used for many previous large-scale corpus investigations, including MD studies of register variation (e.g., Biber, 1988, 1995), the *Longman Grammar of Spoken and Written English* (Biber et al., 1999), and a major study of university spoken and written registers for ETS (Biber, 2006; Biber, Conrad, Reppen, Byrd, & Helt, 2002; Biber, Conrad, Reppen, et al., 2004).

Like all grammatical taggers, the Biber Tagger annotates a text by automatically identifying the part of speech (e.g., noun, verb, preposition) of each word in the text. However, while this tagger achieves accuracy levels comparable to other existing taggers, it is especially robust, having different processing options for oral and literate texts. The Biber Tagger also has more extensive coverage than most other taggers, identifying not only basic parts of speech, but also many other grammatical and syntactic features, such as the tense and aspect of verbs, passive voice, relative clauses, and other postnominal modifier types, complement clause types, and so on.

To ensure the accuracy of the grammatical annotation in the present application, we employed a cyclical process of automatic analysis, evaluation of tagging accuracy, revision, and

development of additional computer programs, and hand-editing of the annotation codes. This process included the following major tasks, described in more detail below:

1. Automatically tagged all responses in the spoken and written subcorpora
2. Evaluated and edited the corpus for textual and formatting issues affecting tagger accuracy and automatically retagged all responses in the spoken and written subcorpora
3. Detailed tagchecking Phase 1: Identified tagging problems, followed by programming revisions to the tagger and automatic retagging
4. Detailed tagchecking Phase 2: Line-by-line evaluation of tags and calculation of initial reliability rates for the tagging process
5. Edited the corpus to remove further textual/formatting issues affecting tagger accuracy and automatically retagged the complete corpus
6. Wrote, tested, and ran Perl scripts to correct systematic lexically governed errors in the automatic annotation (both item-specific errors and corpus-wide errors)
7. Undertook a fix-tagging process, including the development of a fix-tagger computer tool to aid interactive tag checking and correcting, the development of training materials for fix-tagging, and the recruitment and training of fix-taggers, ending with fix-tagging selected features in the entire corpus
8. Analyzed and calculated reliability rates for the final annotated corpus

After all the responses in the spoken and written subcorpora were automatically tagged using the Biber Tagger, the tags were evaluated and several textual and formatting issues were discovered that affected the accuracy of the tagger. These issues (e.g., no spaces after punctuation in many written responses) were corrected in the entire corpus, and the corpus was retagged.

Subsequently, a detailed tagchecking process (Tagchecking Phase 1) was undertaken to identify systematic tagging problems. This initial analysis led to programming revisions of the Biber Tagger and retagging of the corpus, with cyclical evaluation to determine whether changes were effective.

A second detailed tagchecking process (Tagchecking Phase 2) was then undertaken to systematically evaluate the reliability of the automatic tags. For this step, a 5% sample of texts from both the written and spoken subcorpora was randomly selected across test forms, items, and

score levels. Training materials and error-marking conventions were developed, and two independent coders were recruited and trained to complete a line-by-line evaluation of the automatically assigned tags. The project research assistant (RA) served as second coder for the complete 5% sample. Where disagreement with the first tagchecker occurred, a third coder was consulted to resolve the issue. After the 5% sample had been coded for tagging errors, we developed additional computer programs to analyze corpus files coded for errors, measuring accuracy in terms of both precision and recall. The precision scores give the proportion of the automatic tags that are accurate, while the recall scores give the proportion of all actual occurrences of a target linguistic feature that are identified by the automatic software.

We analyzed accuracy rates separately in the written subcorpus and spoken subcorpus, checking for the possibility that the tagging software would encounter different problems in the two registers. Most linguistic features were automatically tagged with a high degree of accuracy, with both precision and recall rates over 90%. However, based on this analysis, we determined that some features required further tag-editing (*fix-tagging*). Both automatic and manual tag-editing were employed in this study.

To address the systematic tagging errors uncovered during Tagchecking Phase 2, a series of computer programs (scripts) were developed to automatically correct tags that were systematically incorrect in certain lexical or grammatical contexts. The scripts corrected errors that were specific to a particular test item (e.g., *fungus* should be tagged as a noun in responses to Speaking Item 6 on Form 2 of the exam), as well as corpus-wide errors (e.g., the verb in the sequence *be able to VERB* should be tagged as an infinitive). These scripts were evaluated for accuracy and then run on the complete corpus.

Other features were better addressed through manual tag-editing. In the written corpus, those features included the following: (a) all occurrences of *that*, to correctly determine their grammatical function as adjective complement clause, noun complement clause, verb complement clause, relative clause and so forth; and (b) all occurrences of past participles (except those tagged as finite passive voice verbs or perfect aspect verbs) to determine their grammatical function as finite past-tense verb versus nonfinite relative clause versus other functions (e.g., attributive adjective). In the spoken corpus, all occurrences of past participles (except those tagged as finite passive voice verbs or perfect aspect verbs) were fix-tagged, because this was the only feature to show major problems with the automatic tagging. In order to

carry out the manual tag-editing process, we developed a fix-tagger computer tool to aid in interactive tag checking and correction, along with training materials that included information on tag descriptions, how to distinguish the various functions of the features that required hand checking, and how to use the fix-tagger tool. We recruited and trained 10 fix-taggers for the hand corrections of problematic tags. All fix-taggers went through a training of 15 text files; the edited texts were then checked by the research assistant, and feedback was given to each fix-tagger. Then, the entire spoken corpus was fix-tagged for past participle forms, and the written corpus was fix-tagged for all instances of *that* and most past participle forms. Separate training materials were developed for each of these linguistic features (available on request).

Accuracy of the grammatical coding was evaluated throughout the fix-tagging process by checking a random subsample of files from each fix-tagger (checked 6.5% of writing files, 7% of speaking files). In addition, a second set of Perl scripts was developed, tested, and run to correct additional tagging errors.

After all grammatical tag-correction was completed (both manual and automatic fix-tagging), the accuracy rates for the grammatical codes in the corpus were reevaluated. Specifically, the research assistant for the project compared the 5% sample of texts originally coded for tagger errors with the same texts after they had been corrected in the interactive fix-tagging and automatic error correction processes. This recoding was performed in order to calculate final reliability measures for the grammatical coding, calculated again in terms of precision and recall. The detailed results are given in Appendix C and D: Nearly all linguistic features in the final versions of the annotated corpora are identified with a high degree of accuracy (precision and recall rates over 90%) and many features have extremely high rates of accuracy (approaching 100%).

4.3. Quantitative Linguistic Analyses

After the corpus preparation was completed, we developed additional computer programs to analyze the quantitative distribution of linguistic features. We analyzed the discourse of TOEFL iBT exam responses at several linguistic levels: vocabulary distributions, collocational differences, phraseological patterns, grammatical features, and lexico-grammatical patterns.

4.3.1. Vocabulary distributions. To investigate vocabulary distributions, a computer program was developed to calculate lexical frequency profiles (see Laufer & Nation, 1995). Laufer and Nation have shown that lexical frequency profiles are useful in distinguishing

learners across proficiency levels, and it is possible that variation also occurs across mode (spoken versus written discourse) and task type (independent versus integrated tasks). Our computer program calculates the percentages of words in a corpus that come from specific vocabulary lists. In this study, the proportion of words (based on word tokens) came from the 1,000 most frequent words in the General Service List (GSL 1K; see Nation, 1990; West, 1953), the second 1,000 most frequent words in the GSL (GSL 2K), and the Academic Word List (see Coxhead, 2000). The program looks up each word in a test taker response, determines the list that the word belongs to (including offlist words), creates a word count for each of those lists, and calculates the proportion of total words in the text that come from each of the lists.

4.3.2. Collocational differences. In order to investigate differences in collocational patterns between score levels, task types, and mode, this study focused on the patterns of use for five high-frequency verbs: *get*, *give*, *have*, *make*, and *take.* For this analysis, we developed a computer program to identify the frequently co-occurring words (the collocates) for each verb. The program identified collocations of all forms of these verbs, automatically identifying content words that co-occurred within three words after the target verb. Collocation was defined using simple distributional criteria: any content word that co-occurred with the target verb in more than 10 texts at a rate of more than five times per 100,000 words. Separate analyses were carried out for the spoken subcorpus and the written subcorpus.

4.3.3. Phraseological patterns. The third approach used to investigate lexical patterns in the TOEFL iBT corpus was to identify extended fixed sequences of words, or *lexical bundles* (see Biber, Conrad, & Cortes 2004; Biber et al., 1999, Chapter 13). For the analysis of the spoken corpus, lexical bundles were defined as any four-word sequence that occurred in at least 15 texts with an overall rate of at least 5 occurrences per 100,000 words. (The Independent Level 1 category was dropped from this analysis because there were too few texts for reliable quantitative results.) There are many fewer texts in the written corpus than in the spoken corpus, but individual texts tend to be much longer. Thus, in writing, the range requirement was reduced to 10 texts but retained the same requirement for rate of occurrence (five times per 100,000 words). The program to investigate lexical bundles identifies each four-word sequence in a corpus, tracking the rate of occurrence for each potential bundle. Bundles meeting the frequency and distribution requirements described above were then analyzed functionally.

4.3.4. Grammatical and lexico-grammatical patterns. The programs that counted the grammatical and lexico-grammatical features were simpler because they were based directly on the previously edited grammatical tags in the corpus. Thus, these programs simply counted the occurrences of each tag in each text of the corpus, including word classes (e.g., pronouns, nouns), grammatical distinctions (e.g., past tense verbs, passive voice verbs, prepositional phrases), syntactic features (e.g., nouns and adjectives as premodifiers of nouns, relative clauses, adverbial clauses), and lexico-grammatical features (e.g., mental verbs controlling *that* complement clauses). For lexico-grammatical features, these programs identified occurrences of specific target words occurring together with the target grammatical construction. All counts were normalized to a rate of occurrence (per 1,000 words of text) so that quantitative measures would be comparable across texts regardless of text length. Appendix E lists the major grammatical and lexico-grammatical features analyzed for the project.

4.4. Quantitative Analyses

Two major research designs were employed for the linguistic analyses: treating each subcorpus as an observation and treating each individual text as an observation (see Biber & Jones, 2009). For the vocabulary and collocational analyses, each subcorpus was treated as an observation. This design was employed because lexical investigations require large corpus samples:Individual words occur much less frequently than grammatical constructions. Thus, all texts in each subcategory were combined into a single sample for the purposes of the lexical analyses. Consequently, the results for these analyses are based on overall rates of occurrence for each subcategory (e.g., an overall rate of occurrence for the collocation *have time* in the spoken corpus), but no parametric statistics are possible.

In contrast, the lexico-grammatical analyses employed a research design where each individual text was treated as an observation. In this case, rates of occurrence (per 1,000 words) were computed for each grammatical feature in each text. Then, it was possible to apply correlational techniques to investigate the relations among variables and to compute means and standard deviations for each linguistic feature in each text category. In addition, this design allows for the application of factor analysis, which is used in MD analysis (see Section 5.4 below).

The major drawback of this second research design is that the results are unreliable when applied to extremely short texts. That is, it is possible to obtain reliable measures for the rates of

occurrence of most grammatical features in texts that are longer than 100 words (see Biber, 1990, 1993). However, two problems arise with quantitative analyses of shorter texts: (a) many features simply do not occur in such texts and (b) the normalized rates of occurrence can be greatly inflated for rare features when they do happen to occur in a short text.

In the present study, an additional confounding factor was present with these short texts: They usually received low scores by TOEFL iBT raters. Thus, all of the spoken-independent texts with a score of 1 in our corpus are shorter than 100 words, and 97% of the spoken-integrated texts with a score of 1 in our corpus are shorter than 100 words. The pattern for written responses is somewhat different, because some longer responses also received a score of 1. However, all written responses shorter than 100 words received a score of 1.

Thus, two methodological problems are caused by the inclusion of short texts in our corpus: (a) the unreliability of rates of occurrence for linguistic features and (b) the confounding influence of text length and TOEFL iBT score. For both reasons, we decided to omit all texts shorter than 100 words for the purposes of the quantitative grammatical analyses (including the MD analysis). The resulting corpus composition is shown in Table 6. (The three spoken-integrated texts with a score of 1 are also omitted from these analyses, because a sample of three observations does not provide an adequate representation of that cell.)

The following sections present the quantitative findings for the discourse characteristics of these TOEFL iBT texts, organized by linguistic level. Results of the lexical analyses (vocabulary distributions, collocational analyses, and lexical bundles) are presented first, followed by the results of the lexico-grammatical analyses. Finally, results of an MD analysis describe the overall patterns of linguistic variation in this discourse domain.

Table 6

Corpus for the Statistical Analyses (i.e., Excluding Texts Shorter Than 100 Words)

Task	Score level	Number of texts	Text length	Mean text length	Min. text length	Max. text length
Spoken independent tasks	1	--	--	--	--	--
(2 responses per test taker)	2	39	4,376	112.2	101	140
	3	142	16,245	114.4	101	172
	4	67	7,953	118.7	101	164
Subtotal		248	28,574			
Spoken integrated tasks	1	3	323	109.3	104	116
(4 responses per test taker)	2	313	37,153	118.7	101	195

Task	Score level	Number of texts	Text length	Mean text length	Min. text length	Max. text length
	3	654	84,104	128.6	101	213
	4	216	30,521	141.3	101	212
Subtotal		1,186	152,106			
Written independent tasks	1	42	9,597	228.5	123	351
(1 response per test taker)	2	177	51,118	288.8	160	507
	3	155	52,452	338.4	206	549
	4	102	39,300	385.3	261	586
Subtotal		476	152,467			
Written integrated tasks	1	119	20,587	173.0	101	293
(1 response per test taker)	2	118	23,683	200.7	102	303
	3	122	25,962	212.8	108	367
	4	112	26,264	234.5	145	388
Subtotal		471	429,643			
Total		2,381	429,643			

5. The Quantitative-Linguistic Descriptions of TOEFL iBT Exam Responses

This study analyzed the discourse of TOEFL iBT exam responses at several linguistic levels: vocabulary distributions, phraseological patterns, grammatical patterns, and an overall MD analysis of the patterns of variation.

5.1. Vocabulary Distributions

Although vocabulary use is not a major focus of the present study, it is useful as background to compare the inventory of words used across modes, tasks, and score levels. Tables 7 and 8 show that most of the words (tokens) used in these TOEFL iBT responses belong to the most common vocabulary items: the top 1,000 words from the GSL (see Nation, 1990; West, 1953). Surprisingly, the pattern is very similar for spoken and written responses, with 80–85% of all words in both modes coming from the top 1,000 GSL.

At the same time, Tables 7 and 8 show small but consistent differences between speech and writing, and across tasks/levels: Written responses—especially integrated responses and higher level independent responses—use more words from the Academic Word List (see Coxhead, 2000) than spoken responses. In contrast, spoken responses use more function words (including pronouns; see Section 5.4) than written responses. Although the differences are relatively small, there is also a trend toward higher level responses using fewer of the most frequent words (GSL 1K words) than lower levels, and more of the less-common words (GSL 2K words) and Academic Word List words.

Table 7

Distribution of Words Across Vocabulary Classes: Spoken Responses

Task type & score	Number of texts	Total words	Number of GSL 1K words	% GSL 1K words	Number of GSL 2K words	% GSL 2K words	Number of AWL words	% AWL words	Number of function words	% function words
Independent score 1	36	1,592	1,348	85%	54	3%	49	3%	141	9%
Independent score 2	368	27,456	22,783	83%	1,101	4%	724	3%	2848	10%
Independent score 3	440	40,563	33,658	83%	1,581	4%	1,051	3%	4273	11%
Independent score 4	116	11,959	9,851	82%	488	4%	305	3%	1315	11%
Integrated score 1	105	6,107	4,903	80%	314	5%	165	3%	725	12%
Integrated score 2	764	72,720	59,784	82%	3,839	5%	1,804	2%	7293	10%
Integrated score 3	826	98,288	80,998	82%	4,827	5%	2,618	3%	9845	10%
Integrated score 4	224	30,901	25,242	82%	1,588	5%	984	3%	3087	10%

Note. GSL 1K = General Service List 1,000 most frequent words; GSL 2K = General Service List second 1,000 most frequent words; AWL = Academic Word List.

Table 8

Distribution of Words Across Vocabulary Classes: Written Responses

Task type & score	Number of texts	Total words	Number of GSL 1K words	% GSL 1K words	Number of GSL 2K words	% GSL 2K words	Number of AWL words	% AWL words	Number of function words	% function words
Independent score 1	46	9,880	8,475	86%	243	2%	438	4%	724	7%
Independent score 2	177	51,056	43,191	85%	1,465	3%	2,421	5%	3,979	8%
Independent score 3	155	52,371	43,750	84%	1,520	3%	3,033	6%	4,068	8%
Independent score 4	102	39,214	32,186	82%	1,209	3%	2,579	7%	3,240	8%
Integrated score 1	128	21,243	17,265	81%	950	4%	1,332	6%	1,696	8%
Integrated score 2	118	23,630	19,267	82%	1,002	4%	1,510	6%	1,851	8%
Integrated score 3	122	25,860	20,875	81%	1,091	4%	1,722	7%	2,172	8%
Integrated score 4	112	26,159	20,863	80%	1,187	5%	1,906	7%	2,203	8%

Note. GSL 1K = General Service List 1,000 most frequent words; GSL 2K = General Service List second 1,000 most frequent words; AWL = Academic Word List.

5.2. Phraseological Patterns

A second perspective on the lexical level of discourse is the investigation of phraseological patterns. In the present study, we approach this issue in two different ways: through consideration of the collocational associations of light verbs (5.2.1), and through consideration of the most common lexical bundles in TOEFL iBT responses (5.2.2).

5.2.1. Collocational associations of light verbs. We focused on the collocational patterns for five semantically light verbs: *get, give, have, make, take* (see Altenberg & Granger, 2001 for a similar approach, focusing on *make* in L2 student writing compared to L1 student writing). Collocations of a given word are lexical items that commonly co-occur with the target word (see, e.g., Partington, 1998). Collocations in our analysis were identified using simple distributional criteria: any content word that co-occurred with one of the five target verbs in more than 10 texts at a rate of more than five times per 100,000 words. Only collocates following the target verbs, including a span of three words, were considered. We carried out separate analyses for the spoken subcorpus and the written subcorpus.

Appendices F and G provide the full lists of collocates for each of the five target verbs in speech and writing. A comparison of the lists shows that spoken responses have many more collocates for these verbs than the written responses. Interestingly, this is especially the case for collocations that have their source in the prompts: Twenty-three prompt-specific collocations were frequently used in the spoken responses (e.g., *get money, give a gift, give an example*), but only two such collocations appeared in the written responses (*HAVE crystals* and *MAKE a/the point*).

The more interesting collocates are those that did not have their source in the prompts, and overall the responses demonstrate awareness of numerous such collocations. Many of these collocations are relatively specialized combinations of words with idiomatic meanings. At least some test takers used these forms in written tasks, where they had time for careful planning, as well as in spoken tasks, under more constrained production circumstances. Examples of relatively idiomatic collocations include the following:

Examples from spoken responses:

GET + rid

GIVE + an assignment

HAVE + a/the chance, a class, a good day, an exam, fun, the opportunity, a problem, a question, a reaction

TAKE + an exam/test/midterm, care, part

Examples from written responses

GET + along, better, a job, good grades

GIVE + an example

HAVE + the ability to, an advantage, the chance, a choice, a career, an effect, no interest, a job, limitations, an opinion, the opportunity to, a problem, time

MAKE + a decision, money, sense

TAKE + care, classes/course/subjects, the example of

In addition, many of the spoken collocations that are labeled *prompt-specific* in Appendices F and G are found in both independent and integrated responses; in the case of their use in independent tasks, these collocations did not occur in the question and so their use must be attributed to the test takers themselves. Examples include *GIVE a gift, HAVE money, HAVE time, MAKE sure, TAKE time*.

A few of these combinations occur with especially high frequencies. In speech, these include prompt-related collocations (e.g., *GIVE an/the example, TAKE a class/classes*) as well as collocations that are attributed directly to the test takers (e.g., *HAVE a problem, HAVE time, MAKE up a/the exam(ination)/test (later)*). In writing, only test taker-initiated collocations occur with especially high frequencies, such as *GET a job, HAVE the ability to*, and *TAKE a class/a course/subjects*.

Tables 9 and 10 summarize the overall distributional patterns for these collocations, suggesting some general differences across the modes, tasks, and score levels. These tables list the number of collocations found in each task/level, defined as the number of word combinations that occur with a frequency of at least five times per 100,000 words in that text category.

In speech (Table 9), these collocations are much more likely to be used in integrated tasks than independent tasks, suggesting that the extra planning time associated with the integrated tasks permits the recall and use of such collocational combinations. This greater use is further facilitated by the occurrence of some collocations in the prompts themselves, although many of those same combinations are used in independent responses where they must be attributed directly to the test takers. Level 1 in the independent responses has the fewest collocations, suggesting that low-level students have not yet acquired many of these lexical combinations.

In contrast, we see a different trend across levels for the use of collocations in the integrated tasks: Levels 2–4 all use a large number of prompt-specific collocations, but other collocations (i.e., directly attributable to the test takers) are more prevalent in responses at Levels 2 and 3 than in Level 4 responses. This trend suggests that intermediate-level students rely on these prepackaged/formulaic expressions to a greater extent than the most proficient students do.

A similar trend can be observed in Table 10, which summarizes the number of collocations found across tasks/levels in the written responses. In contrast to the overall patterns for speech, in writing we see more collocational combinations in the independent tasks than in the integrated tasks. Overall, these patterns suggest that collocational sequences are acquired at intermediate levels and that their use requires some planning and processing time but, at the highest levels and in the tasks with the most opportunity for planning and production, they are less commonly used, possibly because they are stigmatized as being clichés and less creative.

Table 9

Number of Co-Occurring Collocates (Frequency > 5 per 100,000 Words) With Each Verb: Spoken Responses

Target verb	Task and level/Source of collocates = Test taker versus prompt															
	Independent 1		Independent 2		Independent 3		Independent 4		Integrated 1		Integrated 2		Integrated 3		Integrated 4	
	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt
GET	1	0	3	2	3	2	2	1	2	3	3	4	2	4	1	4
GIVE	0	0	0	1	0	1	0	0	1	2	2	3	2	3	1	3
HAVE	1	1	4	3	7	3	6	3	6	4	12	5	11	5	9	5
MAKE	0	0	2	0	2	0	2	1	1	3	6	4	6	4	3	5
TAKE	0	0	1	1	4	1	1	1	4	3	4	5	4	6	3	4
Total	2	0	17	0	23	0	17	0	14	15	27	21	25	22	17	21

Table 10

Number of Co-Occurring Collocates (Frequency > 5 per 100,000 Words) With Each Verb: Written Responses

Target verb	Task and level/Source of collocates = Test taker versus prompt															
	Independent 1		Independent 2		Independent 3		Independent 4		Integrated 1		Integrated 2		Integrated 3		Integrated 4	
	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt	Test taker	Prompt
GET	1	0	2	0	4	0	3	0	2	0	0	0	1	0	0	0
GIVE	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0
HAVE	10	0	15	0	13	0	14	0	8	1	5	1	5	1	5	1
MAKE	1	0	4	0	5	0	5	0	2	1	0	1	1	1	0	1
TAKE	3	0	5	0	4	0	5	0	1	0	0	0	0	0	0	0
Total	15	0	26	0	26	0	28	0	14	2	6	2	8	2	6	2

5.2.2. Lexical bundles. Phraseological patterns can also be investigated through consideration of extended fixed sequences of words, referred to as *lexical bundles* (see Biber et al., 1999, Chapter 13). For the analysis of the spoken corpus, we defined lexical bundles as any four-word sequence that occurred in at least 15 texts and had an overall rate of at least five occurrences per 100,000 words. (The Independent Level 1 category was excluded from this analysis because it contained too few texts for reliable quantitative results; see Section 4.2.) There are many fewer texts in the written corpus than in the spoken corpus, but individual texts tend to be much longer. Thus, in writing, we defined lexical bundles as any four-word sequence found in at least 10 texts with the same rate of occurrence (five per 100,000 words).

We classified lexical bundles into five major functional categories (extending the framework developed in Biber, Conrad, & Cortes, 2004: personal/epistemic bundles, attitudinal/evaluative bundles, information source bundles, information organizers, and discourse organizers. Appendices H and I provide complete lists of the lexical bundles used in the spoken and written responses and grouped into these major functional categories, while Tables 11 and 12 summarize the breakdown of bundle types across TOEFL iBT text categories.

Previous research on lexical bundles has shown that these fixed lexical sequences are generally more prevalent in spoken registers than they are in written registers. However, a comparison of Tables 11 and 12 shows that lexical bundles in the TOEFL iBT context are prevalent in both speech and writing, although each text category relies on different functional types.

One major reason for this distribution is that many lexical bundles directly reflect the exam questions in both speech and writing. Thus, for example, spoken independent tasks note that "Others think it is better to go…" and specifically ask questions beginning with "What do you think is the best way for…" and "Do you think your life is…." Responding to these questions results in an extremely frequent use of epistemic and attitudinal lexical bundles that are essentially copied from the prompts, such as: *I think my life*, *I think the best*, *the best way for*, and *it is better to go*. We also find even longer recurrent sequences of words that are taken from the prompts, such as *I think it's better to*, *it is better to go*, and even *I think it's better to go*.

Table 11

Lexical Bundle Types in Spoken Responses

Bundle type	Number of bundles	Rates per 100,000 words						
		Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
Personal/epistemic bundles	18	630.1	554.7	401.4	16.4	74.3	73.3	38.8
Attitudinal/evaluative bundles	23	775.8	574.4	501.7	49.1	130.6	131.2	80.9
Information source	5	58.3	37.0	33.4	16.4	42.7	25.4	29.1
Information organizers	8	105.7	39.4	33.4	65.5	48.1	53.9	51.8
Discourse organizers	6	14.6	66.6	92	32.7	48.1	91.6	84.1

Note. Ind = independent task; Int = integrated task.

Table 12

Lexical Bundle Types in Written Responses

Bundle type	Number of bundles	Rates per 100,000 words							
		Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
Personal/epistemic bundles	7	50.6	58.8	45.8	58.7	42.4	21.2	61.9	65
Attitudinal/evaluative bundles	41	1194.3	1333.8	943.3	673.2	37.7	25.4	50.3	38.2
Information source	16	20.2	19.6	28.6	5.1	287.2	474.0	263	328.8
Information organizers	16	50.6	33.3	55.4	17.9	211.8	241.2	278.4	263.8
Discourse organizers	4	30.4	60.7	64.9	51.0	70.6	84.6	46.4	38.2

Note. Ind = independent task; Int = integrated task.

Similarly in the written independent tasks are questions like the following: "Do you agree or disagree with the following statement? It is more important to choose to study…." Not surprisingly, this prompt results in frequent attitudinal/evaluative lexical bundles like *is more important to choose* and *I agree with the/this statement*. Many information-organizing bundles in writing are also prompt-dependent. For example, one prompt has a reading passage and lecture that discusses three theories about bird navigation. Correspondingly, we find frequent bundles like *there are three theories*, *the first theory is*, *the second theory is*, and so forth.

At the same time, there are many other bundles in the TOEFL iBT texts that should be attributed to the test takers rather than the prompts. In speech, these include the hedging tag *or something like that* as well as attitudinal expressions like *if you want to*, *you don't want to*, *to be able to*, and *not be able to.* Epistemic bundles that incorporate *I think* are among the most frequent bundles found in natural conversation as well as university classroom teaching (see Biber, Conrad & Cortes, et al., 2004), so the extremely high frequency of these bundles in spoken independent tasks is probably also due in part to the test takers themselves. This interpretation is supported by the fact that three of these same bundles are found in written independent responses (*think that it is*, *I think it is*, *I think that it*), where there is no corresponding bundle used in the prompt or question. In addition, we find test taker–sourced epistemic bundles of factuality being used in the written responses: *it is true that*, *to the fact that*, *the fact that the*, and *a matter of fact*.

In the general functional domain of discourse organizing bundles, we find mostly lexical sequences that originated with the test takers rather than language from the prompts. Three subcategories comprise this general functional domain: Bundles that identify the source of information (i.e., the test taker, a lecturer, a reading passage, and so forth), bundles that organize the specific information in a response (*the second theory is*, *the first reason is that*), and general purpose discourse organizers (*on the other hand*, *at the same time*).

Source-of-information bundles are especially prevalent in written responses. These are mostly used to identify sources other than the test taker: either a lecturer (e.g., *according to the professor*) or a reading passage (e.g., *in the reading passage*). As a result, these bundles are used almost entirely in integrated written responses. Overall, there are fewer source-of-information bundles in speech, but they include both those identifying external sources, used in integrated

responses (e.g., *in the listening passage*) as well as those used to overtly signal the test taker's own opinion, used in independent responses (e.g., *in my opinion I*, *I agree with the*).

The second subcategory of discourse organizing bundles is used to organize the specific information in texts. As noted above, most of these bundles are closely tied to either the prompts or the particular questions that students respond to. This is especially the case in written responses to integrated tasks. For example, in one question test takers are asked to summarize the points made in a lecture that discusses three theories about bird navigation. Correspondingly, we find frequent bundles like *the points made in the*, *there are three theories*, *the first theory is*, *the second theory is*, and so forth.

Many of the tasks required for spoken integrated responses involve itemized responses (e.g., a task that requires the test taker to "explain two ways that fungus indirectly benefits trees"), including discussion of the reasons underlying an opinion: "Explain the reasons she gives for holding that opinion," "Explain the reasons for your recommendation," "Describe his opinion and his reasons for holding that opinion." As a result, we find frequent bundles identifying different possibilities (*the first one is*, *the second one is*) and specifically identifying different reasons (*the first reason is*, *the second reason is*). Interestingly, spoken independent responses also frequently use these same bundles, even though there is no mention of the need to give reasons in the specification of the task itself. Instead we find only the requirement to "explain why." However, many test takers decide that the best way to structure the explanation of their opinion is by using these lexical bundles that identify reasons for their opinion.

The third subcategory of discourse organizing bundles—general purpose discourse organizers—are found in both spoken and written responses. It is somewhat surprising, though, that there is a larger inventory of these bundles used in speech than in writing. Two of these are especially frequent in both speech and writing, used in both independent and integrated tasks: *at the same time* and *on the other hand*.

Finally, Tables 11 and 12 reveal an interesting trend in the distribution of lexical bundles across score levels: In general, the two intermediate score levels (2 and 3) use these bundles to a greater extent than either the lowest level (1) or the highest level (4). In speech, the only real exception to this pattern is for general discourse organizers, which are most common in Level 4 independent responses and Level 3/4 integrated responses. In writing, the only real exception is for epistemic bundles, which are used more frequently in Level 4 responses, in both independent

and integrated tasks. Overall, this pattern suggests a general developmental progression in which low level test takers are just beginning to acquire the use of these fixed expressions (and thus use them less frequently), intermediate level test takers have acquired the expressions but tend to overuse them (resulting in the highest frequencies), and the highest scoring test takers control these fixed expressions but often choose to use alternative (more creative) expressions in their discourse.

5.3. Lexico-Grammatical Patterns

The main goal of the present project was to provide a comprehensive quantitative description of lexico-grammatical characteristics in the discourse of TOEFL iBT responses. Descriptive statistics for the 171 grammatical and lexico-grammatical features investigated are available by request, while Appendix J provides descriptive statistics for the 36 most important grammatical features (see discussion below. As explained in Section 4.3, no spoken texts with Score Level 1 were included in the corpus for grammatical analysis, and all texts shorter than 100 words were excluded from this analysis.)

In general, the largest quantitative differences found in the investigation were between spoken versus written responses. For example, Appendix J shows that prepositional phrases occur circa 80 times per 1,000 words in speech, with a range of 75.6–85.2 among the spoken tasks/levels. In contrast, prepositional phrases occur circa 103 times per 1,000 words in writing, with a range of 99.3–106.1 among the written tasks/levels. These ranges of variation are nonoverlapping for the two modes, with all written tasks/levels having rates of occurrence that are circa 25% higher than any spoken task/level. The difference between the two modes is even more dramatic for features like nominalizations, where all written tasks/levels have rates of occurrence circa 10 times higher than any spoken task/level.

One major research question considered in the project was the extent to which the rater score was influenced by the use of particular grammatical characteristics. To begin the investigation of this question, we carried out exploratory correlations of each linguistic feature (rate of occurrence) with score level; we carried out separate analyses for speech and writing, reflecting the likelihood that particular grammatical features would be used to different extents in the two modes.

Surprisingly, few of the lexico-grammatical features considered in our study correlate with score level. Thus, in speech only 14 of the 170-plus grammatical features investigated here

have even a minimal correlation (> .1) with score level. Seven features have positive correlations with score: word length (.17), adverbs (.16), finite passive verbs (.13), stance adverbials (.11), attributive adjectives (.10), split auxiliaries (.10), *as* (.10); another seven features have inverse correlations with score: non-past tense (-.12), third-person pronouns (-.12), place nouns (-.12), human nouns (-.12), all modals (-.11), possibility modals (-.11), and desire verb + *to*-clause (-.10).

Scores for the written responses have stronger correlations with grammatical features. Fifteen features have positive correlations greater than .1: split auxiliaries (.27), finite passive verbs (.23), non-finite passive postnominal clauses (.19), perfect aspect verbs (.18), word length (.17), attributive adjectives (.16), adverbs (.16), progressive aspect verbs (.14), ability adjective + *to*-clause (.14), *as (*.14), *-ing* complement clauses (.12), certainty verb + *that*-clause (.12), adjective + *to*-clause (.12), relational adjectives (.11), certainty stance adverbials (.10); another six features have negative correlations: non-past tense (-.19), first-person pronouns (-.10), place nouns (-.15), possibility modals (-.16), clausal *and* (-.10), main verb *have (*-.12). However, in both speech and writing, the large majority of grammatical features are essentially uncorrelated with score level.

In addition, even the features listed above have only a weak relationship to score. Thus, the strongest correlation in speech—for word length—represents only a 3% relationship with score (i.e., $r^2 = .029$). The correlations are slightly stronger in writing, with the strongest correlation—for split auxiliaries—representing a 7% relationship with score (i.e., $r^2 = .073$). In sum, variation in the use of independent linguistic features is largely uncorrelated with TOEFL iBT score, even when spoken texts are analyzed separately from writing.

It is possible, however, that these lexico-grammatical features might be important for distinguishing among task types, which in turn interact with the score ratings. To investigate this possibility, together with other possible interactions among the external factors, we undertook full factorial analyses of mode, task type, and score level as predictors of the variation in the use of 36 major grammatical features. These 36 features were chosen based on three primary considerations: (a) they had been identified as theoretically important in previous studies of L2 language development; (b) they had been shown to have some relationship to score in the exploratory correlational analysis; or (c) they occurred frequently enough in the TOEFL iBT corpus to warrant further statistical analysis. To adjust for these repeated tests of statistical

significance, we set an experiment-wise required probability level of $p < .001$ (that is, .05 / 36 = .0014). Descriptive statistics for these 36 features, broken down by mode, task type, and score level, are presented in Appendix J.

We used general linear models in SAS for the statistical analysis of these grammatical features. Four categorical variables were used as independent variables: mode (spoken or written), task (independent or integrated), score level (1, 2, 3, 4), and test taker. The last variable was required because most of the test takers included in our sample produced multiple texts included in the corpus. Thus, for both statistical and theoretical reasons, it was necessary to consider the possible influence of individual students as a predictor of linguistic variation. Statistically, this is a type of repeated measure design; thus, it was necessary to control for the possible influence of individual student. However, in this case, this variation is also of theoretical interest, because such variation might reflect patterns of individual language use or development: cases where an individual examinee relies on a grammatical feature to a greater extent than expected, across modes, task types, and score levels. (Note, however, that in most cases our corpus did not include complete exams from individual test takers, so we were only able to carry out a restricted analyses of the influence of individual variation; see the discussion in Section 3.)

As noted above, we ran statistical tests for 36 grammatical features with a required probability level of $p < .001$ for the overall model in each case. Then, for those models that were significant, we considered the effects of each independent variable and all interactions. For this purpose, we used Type III sums of squares, which included variation that was unique to an effect after adjusting for all other effects that were included in the model. (This approach was especially important in the present study because the subcategory samples were not balanced, and thus any simple comparison of high-level categories would otherwise have been confounded.)

Table 13 summarizes the results of the factorial comparisons. In addition to the information about individual grammatical features, there are a few interesting general patterns that can be observed from Table 13. First of all, most of these features are associated with significant and important differences in the TOEFL iBT Corpus, with overall model r^2 values ranging from circa 40% to circa 75%. These significant models are mostly associated with strong differences between the spoken and written modes and with independent versus integrated tasks. In addition, 23 of these features have significant interaction effects between mode and task type

(Table 13). These findings highlight the importance of mode and task-type differences in the TOEFL iBT, providing strong confirmation to the validity argument for the inclusion of both independent and integrated tasks in speech and writing. In contrast, score level is not a significant predictor of variation in the use of most grammatical features, either as an independent factor or in interaction with mode/task.

For those grammatical features and predictive factors that showed significant differences, it is possible to interpret the patterns of use by examining the mean scores for each category. Appendix J presents the mean scores and standard deviations for each feature, while Table 14 summarizes the major patterns of use. The symbols used in Table 14 represent significant effects: + and ++ mark significant main effects; * and ** mark significant interactions. In addition, based on consideration of the actual mean scores, Table 14 identifies the particular mode/task-type/score-level that used the feature most and describes the major patterns of interaction.

The features listed in Table 14 are grouped to highlight those that behave in similar ways. Two major categories of grammatical features emerge from this analysis:

1. Those that are more frequent in speech and in independent tasks; some of these features are also more common in low-scoring responses; and
2. Those that are more frequent in writing and in integrated tasks; some of these features are also more common in high-scoring responses.

The features associated with speech and independent tasks include verbs (present tense, past tense, perfect aspect, possibility modals), pronouns (especially 1st person, but also second-person and third-person), adverbial structures (total adverbs, stance adverbials, adverbial clauses), clauses connected by *and*, and desire verbs (especially *want*) controlling a *to*-clause. Some of these features are also associated with low-scoring responses: frequent use of present tense verbs, first-person pronouns, possibility modals, and desire verbs controlling a *to*-clause (e.g., *I want to…*).

Table 13

Summary of the Full Factorial Models for 36 Grammatical Features

Linguistic feature	Model	R^2	Mode (sp/wr)	Task	Score level	Mode*task	Mode*score	Task*score	Mode*task*score	Test taker
Word length	< . 0001	0.652	< . 0001	< . 0001	< . 01	< . 0001	ns	< . 05	ns	< . 0001
Non-past tense verbs	< . 0001	0.464	< . 001	ns	ns	ns	ns	< . 001	< . 05	< . 0001
Past tense verbs	ns									
Perfect aspect verbs	< . 0001	0.449	ns	< . 0001	ns	< . 05	ns	ns	ns	< . 0001
Progressive aspect verbs	< . 0001	0.413	< . 01	ns	ns	< . 05	ns	ns	ns	< . 0001
Passive voice verbs	< . 0001	0.539	< . 0001	< . 0001	ns	< . 0001	ns	< . 001	ns	< . 001
Copula *BE* as main verb	ns									
Phrasal verbs	ns									
Possibility modals	< . 0001	0.416	ns	ns	< . 01	< . 001	ns	ns	ns	< . 0001
Prediction modals	< . 0001	0.402	ns	ns	ns	< . 0001	ns	ns	ns	ns
Clausal *and*	< . 0001	0.558	ns	< . 05	ns	< . 05	ns	ns	ns	< . 0001
Adverbs	< . 0001	0.483	< . 0001	< . 0001	ns	< . 0001	ns	ns	ns	< . 01
Split auxiliaries	< . 001	0.397	ns	ns	ns	< . 0001	ns	ns	ns	ns
Stance adverbials	< . 0001	0.49	ns	ns	ns	ns	< . 05	ns	ns	< . 0001
First-person pronouns	< . 0001	0.643	< . 0001	< . 0001	ns	< . 0001	ns	< . 01	< . 01	ns
Second-person pronouns	ns									
Third-person pronouns	< . 0001	0.417	< . 0001	< . 0001	ns	< . 0001	ns	ns	ns	ns
Linking adverbials	< . 0001	0.442	ns	< . 05	ns	< . 05	ns	ns	ns	< . 0001
Nouns	< . 0001	0.702	< . 0001	< . 0001	ns	< . 0001	ns	< . 001	ns	< . 0001
Nominalizations	< . 0001	0.774	< . 0001	< . 001	ns	< . 001	< . 01	< . 01	< . 01	< . 0001
Prepositions	< . 0001	0.557	< . 0001	< . 05	ns	ns	ns	ns	ns	< . 0001
Of genitives	< . 0001	0.509	< . 0001	< . 0001	ns	< . 0001	ns	ns	ns	< . 0001
Attributive adjectives	< . 0001	0.474	< . 0001	< . 0001	< . 05	ns	< . 05	ns	ns	< . 001
Premodifying nouns	< . 0001	0.564	< . 0001	< . 0001	ns	< . 0001	ns	< . 01	ns	< . 0001
Finite adverbial clauses	< . 0001	0.421	< . 0001	< . 0001	ns	< . 05	ns	ns	ns	< . 001
WH complement clauses	ns									
Verb + *that*-clause	< . 0001	0.533	ns	< . 0001	0.05	< . 0001	ns	ns	ns	< . 0001
Adjective + *that*-clause	ns									
Noun + *that*-clause	< . 0001	0.472	< . 05	< . 0001	ns	< . 0001	ns	ns	ns	< . 0001

Linguistic feature	Model	R^2	Mode (sp/wr)	Task	Score level	Mode*task	Mode*score	Task*score	Mode*task*score	Test taker
Verb + *to*-clause	ns									
Desire verb + *to*-clause	< . 0001	0.413	ns	< . 0001	ns	< . 0001	ns	< . 05	ns	ns
Adjective + *to*-clause	< . 0001	0.429	ns	< . 0001	ns	ns	ns	ns	ns	< . 01
Noun + *to*-clause	< . 0001	0.484	< . 0001	< . 0001	ns	< . 0001	ns	< . 05	ns	< . 0001
Verb + *ing*-clause	ns									
Finite relative clauses	< . 0001	0.408	ns	< . 01	ns	< . 01	ns	ns	ns	< . 0001
Passive *-ed* relative clause	< . 0001	0.508	< . 0001	< . 0001	ns	< . 0001	ns	< . 0001	ns	< . 0001

Note. See Appendix J for detailed descriptive statistics. Sp/wr = spoken mode/written mode; ns = not significant.

Table 14

Summary of the Major Patterns for Linguistic Features Across Mode (Speech Versus Writing), Task Type (Independent Versus Integrated), and Score Level

Linguistic features that are generally more common in speech, independent tasks, and lower score levels	Mode		Task		Score level		Interactions
	SP	WR	IND	INT	1	4	
Nonpast tense verbs	++*		**		**		Most in low-scoring independent tasks
Perfect aspect verbs	*		++				More in spoken independent texts
1st person pronouns	++**		++**		*		Most in low-scoring independent texts; especially spoken
3rd-person pronouns	++**			++**			Most common in spoken integrated
Linking adverbials	*	*	*	*			Most common in spoken independent and written integrated
Possibility modals	**		**		+		More in spoken/independent/low-scoring texts
Stance adverbials	*					*	Most in high-scoring spoken (independent) texts
Adverbs	++**		++**				Most in spoken independent texts
Finite adverbial clauses	++*		++*				Most in spoken/independent texts; rare in written integrated texts
Clausal *and*	*			+			More in spoken, integrated (low-scoring) texts
Desire verb + *to*-clause	**		++		*		Most spoken/independent/low-scoring texts
Adjective + *to*-clause			+				More common in independent texts
Word length		++**		++**		+*	Spoken independent has the shortest words; written/integrated / high-scoring has longer words

Linguistic features that are generally more common in speech, independent tasks, and lower score levels	Mode		Task		Score level		Interactions
	SP	WR	IND	INT	1	4	
Passive voice verbs		++**		++**		**	More in written/integrated/high-scoring texts
Nouns		++**		++**		**	Most in written integrated texts; interaction with score is hard to interpret
Nominalizations		++**		++**			Most in written/integrated texts
Prepositions		++		+			Most in written (integrated) texts
Noun + *of*-phrase		++**		++**			Most in written integrated texts
Attributive adjectives		++*		++		+*	Most in written / integrated / high-scoring texts
Premodifying nouns		++**		++**			Most in written integrated texts
Verb + *that*-clause		**		++		+	Most in high-scoring written integrated texts
Split auxiliaries		**		**			Most in (high-scoring) written integrated texts
Noun + *that*-clause		+**		++**			Most in written integrated texts
Passive *–ed* relative clauses		++**		++**		**	Most in high-scoring written integrated texts
Noun + *to*-clause		++**	++**				Most common in written independent texts
Progressive aspect		+		*			More in writing; least in spoken independent texts
Finite relative clauses		*		+			More common in written integrated texts

Note. This table is based on significance for the main effects and interaction effects, considered together with the descriptive statistics for each group. SP = spoken mode; WR = written mode; IND = independent task; INT = integrated task.

+ marks main effects at < .05; ++ marks main effects at < .001; * marks interaction effects at < .05; ** marks interaction effects at < .001.

The features associated with writing and integrated tasks are mostly noun phrase features: noun classes (nouns, nominalizations) and noun phrase modifiers (prepositions, noun+*of*-phrase, attributive adjectives, premodifying nouns, noun+*that*-clause, noun+*to*-clause, and passive *–ed* relative clauses). Longer words, which are often morphologically derived forms, also have the same distribution. In addition, a few verbal/clausal features are associated with writing and/or integrated tasks: passive voice verbs, verb+*that*-clause, split auxiliaries, and progressive aspect verbs.

When we compare these patterns to those documented in previous research on oral/literate differences, it is clear that the tasks included in the TOEFL iBT effectively represent a range of the register variation found in English university discourse and that many of these test takers control these register differences. For example, recent research on grammatical complexity in spoken and written registers (e.g., Biber, 2009; Biber & Gray, 2010; Biber, Gray, & Poonpon, 2011) has shown that *oral* registers are characterized by frequent use of verbs, adverbs, pronouns, and finite dependent clauses. In contrast, informational written registers are to a large extent nonclausal, being instead characterized by a very dense use of nouns and phrasal constructions used as noun modifiers. The TOEFL iBT responses conform to these same general characteristics: Spoken responses tend to use verbs, pronouns, clauses and clausal modifiers (adverbials); written responses tend to use nouns and phrasal noun modifiers. Independent tasks, where test takers give an opinion on a topic, are relatively similar to some of the typical communicative purposes of conversation, and thus they tend to use oral linguistic features. In contrast, integrated tasks are embedded in a literate context, with test takers reading a written text or listening to a scripted passage as background before producing their response; thus, test takers tend to use *literate* linguistic features in integrated tasks.

Register awareness is a major component of language development. Thus, higher-proficiency students will use these oral versus literate groups of features appropriately in spoken versus written registers. In contrast, lower-proficiency students will still be developing this register awareness, and specifically they will probably continue to rely on oral features even in written tasks (see Biber, Gray, & Poonpon, 2011, pp. 29–32).

To some extent, linguistic differences associated with TOEFL iBT scores also conform to these expectations, especially regarding the association of some literate features with higher-scoring written responses: long words, passive voice constructions (finite passive verbs and

nonfinite passive relative clauses), and attributive adjectives. (Interestingly, verb + *that*-clause constructions are also associated with high-scoring written integrated responses, even though this feature is much more common in conversation than in academic writing generally.)

Passive verbs show the strongest association with TOEFL iBT scores, but as an interaction effect with task type rather than as a significant main effect. Two related grammatical features were included in our analysis: finite passive verbs and nonfinite passive relative clauses. For example:

> Finite passive verb:
> This theory *was criticized* by some scientists.
>
> Nonfinite passive relative clause:
> The lecture emphasizes the difference between the aspects *shown in the reading* and what really happens.

Figures 1 and 2 plot the mean rates of occurrence for these grammatical features across TOEFL iBT score levels and task types. Both features show the same general patterns:

1. These passive features are much more common in written-integrated tasks than in the other three task types, and
2. Within written-integrated responses, there are consistent and relatively strong differences across iBT score levels, with higher level scores using passives to a greater extent than lower scores (r = .34 for finite passive verbs; r = .26 for non-finite passive relative clauses).

Passive voice verbs are a perceptually salient grammatical feature that has strong associations with academic writing. The patterns displayed in Figures 1 and 2 show that examinees and exam raters are aware of the associations and positively reward the use of these features in written-integrated tasks.

However, most other grammatical features are weak predictors of score level, with many features having no significant relationship to TOEFL iBT score at all. The overall generalization here is that variation in the use of most grammatical features has little relationship to TOEFL iBT score.

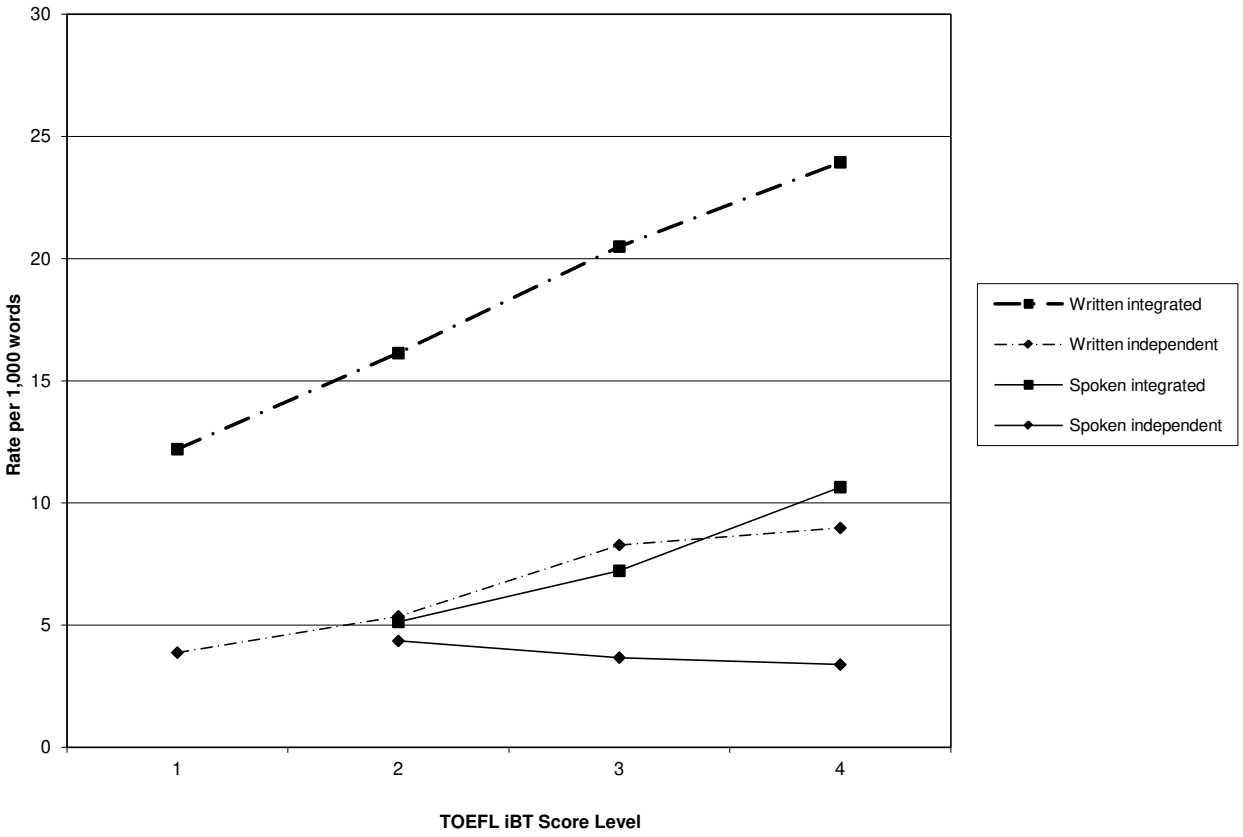

Figure 1. **Finite passive-voice verbs across score levels and task types.**

In contrast, the test taker control variable is a significant and strong predictor of variation for nearly all of these linguistic features (see Table 13). That is, across the multiple responses produced by each test taker, there are significant differences in the extent to which individual test takers use these linguistic features. This is the pattern of use for most grammatical complexity features, including nouns, nominalizations, prepositional phrases, premodifying nouns, noun complement clauses (both *that*-clauses and *to*-clauses), and finite relative clauses. These features are strongly associated with mode and task differences, being generally used more in written integrated tasks. In addition, there is extensive individual variation in the use of these features, with some test takers using these features across responses and other examinees rarely using these features. But in contrast, none of these features is a significant predictor of score level differences.

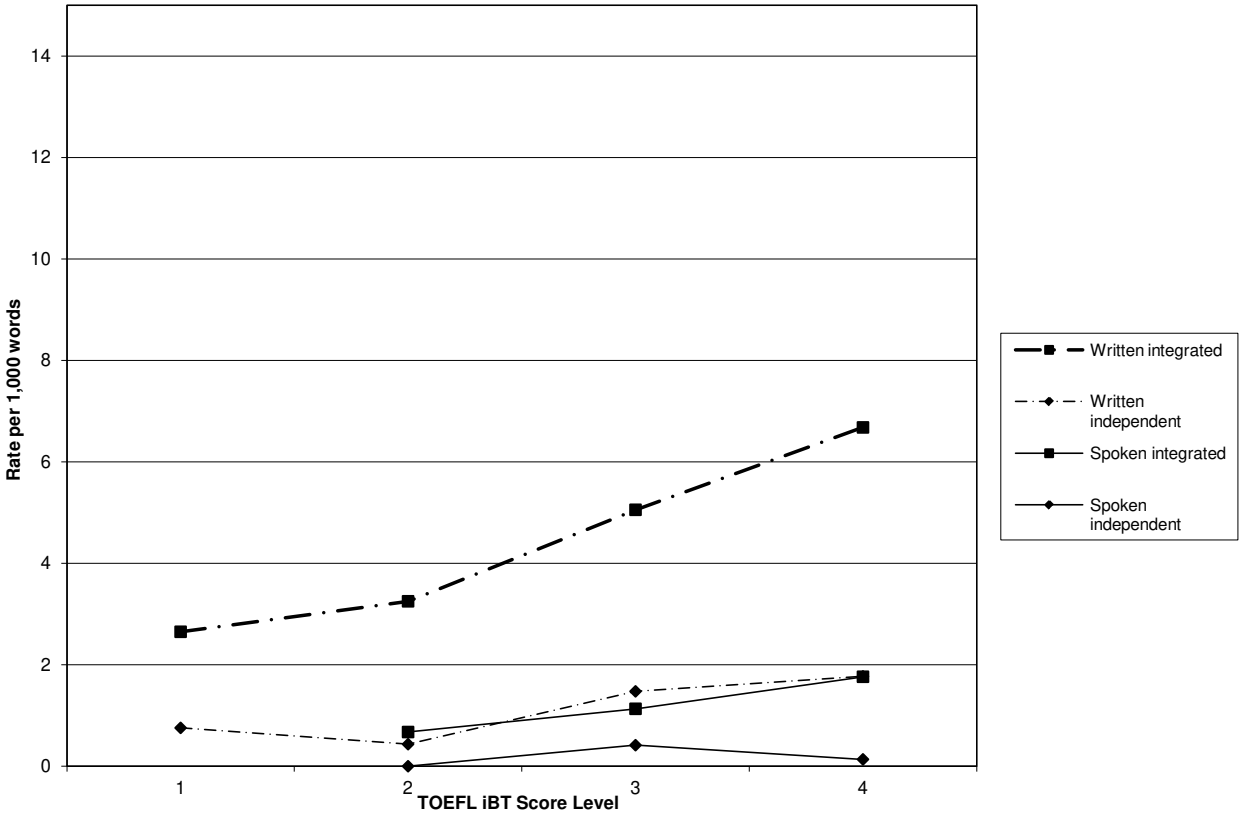

Figure 2. **Nonfinite passive relative clauses across score levels and task types.**

A simple inspection of the descriptive statistics for these lexico-grammatical features (see Appendix J) illustrates the extent of this variation. For example, Figure 3 below shows the range of values (rate of occurrence per 1,000 words) for the use of nominalizations in written integrated responses across the four score levels. Within each score level, some of these texts use almost no nominalizations, and some of these texts have a very dense use of nominalizations. There is clearly extensive linguistic variation here: some of these test takers employ frequent nominalizations, and others do not. This same general pattern exists for many of the other grammatical complexity features considered in this section. The data clearly shows that test takers vary considerably in their use of these linguistic features, but that this variation has little or no relation to TOEFL iBT score level.

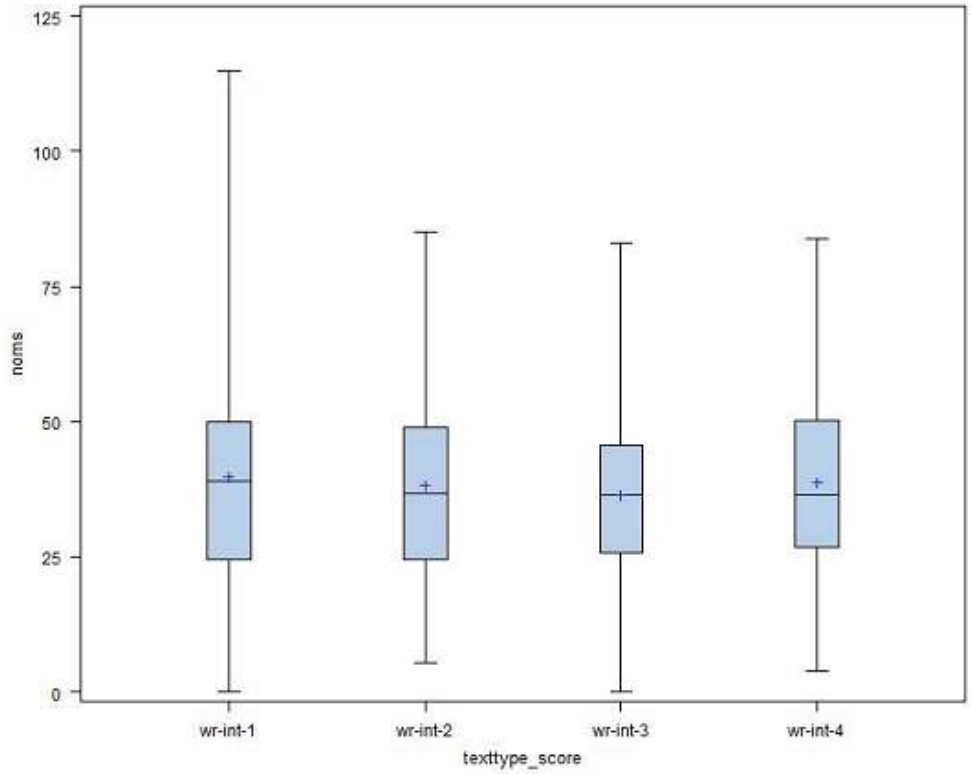

Figure 3. **Box plot of the use of nominalizations across score level in written integrated responses.**

This finding is consistent with previous research that has shown a weak and inconsistent relationship between holistic scores and the use of individual grammatical features. For example, Cumming et al. (2006, pp. 29–30) found a significant difference across score levels for the length of T-units (number of words), but no difference for the complexity of T-units (number of clauses per T-unit). Similarly, Brown et al. (2005, pp. 69–70) found a significant difference across score levels for utterance length but no consistent score-level differences for the T-unit complexity measure (clauses per T-unit) or the dependent clause ratio.

Jarvis et al. (2003) directly tackled this surprising general pattern, documenting the ways in which highly rated written essays can vary in their lexico-grammatical characteristics. In particular, that study shows that there are multiple linguistic profiles that students employ to achieve high-quality writing. The underlying claim of this research is that instructors/raters and students/examinees are much more tuned in to constellations of linguistic features used

effectively than they are to the use of any individual linguistic feature: "The quality of a written text may depend less on the use of individual linguistic features than on how these features are used in tandem" (Jarvis et al., 2003, p. 399).

Section 5.3 is a research methodology developed for research problems of this type (see, e.g., Biber 1988, 1995). Thus, in the following section, we consider the underlying dimensions of linguistic variation among TOEFL iBT responses, investigating the differences among task types with respect to those dimensions, as well as the extent to which score level differences are captured by those parameters of co-occurring linguistic features.

5.4. Multidimensional (MD) Analysis

The quantitative approach of MD analysis allows the researcher to compare many different registers and text categories with respect to several linguistic parameters—the dimensions. Each dimension represents a set of co-occurring linguistic features, that is, linguistic features that tend to be used together in texts. Thus, MD analysis offers a complementary perspective to analyses based on consideration of individual linguistic features (as in the previous section).

Registers can be more or less different with respect to each dimension. By considering all linguistic dimensions, it is possible to describe both the ways and the extent to which registers differ from one another, and ultimately, the overall patterns of register variation in a discourse domain.

In the previous section, we undertook a detailed investigation of linguistic variation based on the distribution of individual linguistic features. That approach identified strong linguistic differences across modes and task types. However, most individual linguistic features were not associated with significant differences across score levels.

In the present section, we consider a complementary perspective, investigating the ways in which linguistic features co-occur in texts and thus work together as underlying *dimensions* of variation. As the following discussion shows, some of these linguistic dimensions are associated with systematic differences across the modes, task-types, and score levels, indicating that these constellations of co-occurring linguistic features are much more important discourse characteristics than linguistic features considered individually.

The notion of linguistic co-occurrence is central to the MD approach, in that different co-occurrence patterns are analyzed as underlying dimensions of variation. The first step in an MD

analysis is to analyze the co-occurrence patterns among linguistic features, using a factor analysis of the rates of occurrence for each linguistic feature. Then factor scores for each text with respect to each factor are computed, and the mean factor scores for each register are compared to analyze the linguistic similarities and differences among registers. Finally, factors are interpreted functionally as underlying dimensions of variation. (See Biber, 1988, Chapters 4–5; Biber, 1995, Chapter 5, for detailed discussion of the methods for MD analysis.) Within the context of the TOEFL iBT, MD analysis has been applied to the description of spoken and written registers in American universities; see Biber, 2006, Chapter 7.)

For the present MD analysis, we began with the major linguistic features analyzed in the MD Analysis section above plus a few additional features that have theoretical importance. Some features were dropped from the analysis because they shared little variance with the overall factorial structure (as shown by the communality estimates). For the final factor analysis, 28 linguistic features were retained.

Appendix K gives the full factorial structure for this analysis. The solution for four factors was selected as optimal. Solutions with fewer factors resulted in a collapsing of linguistic features onto single factors, making the interpretation of those factors more difficult. Solutions with additional factors accounted for little additional variance, and those factors were represented by only a few features. The choice of a four-factor solution was further supported by visual inspection of a scree plot of the eigenvalues. Taken together, these four factors account for 44% of the shared variance (see Appendix K, Table K2).

Following initial extraction, the factor solution was rotated using a Promax rotation, which allows for correlated factors. Only small correlations (less than .3) exist in the present solution (see Appendix K, Table K3).

Table 15 presents the important linguistic features loading on to each factor (i.e., features with factor loadings over + or - .3). This table also includes interpretive labels for each factor; these are explained in the discussion below.

Table 15

Summary of the Important Linguistic Features Loading on Each Factor

Dimensions	Features with positive loadings	Features with negative loadings
Dimension 1 Literate versus oral responses	Nouns: common nouns (.64), concrete nouns (.64), premodifying nouns (.39)	Verbs: present tense (-.33), mental verbs (-.62), modal verbs (-.36)
	Prepositional phrases (.52), noun + *of*-phrase (.47)	Pronouns: third person (-.55)
	Adjectives: attributive (.61), topical (.40)	*That*-clauses: controlled by likelihood verbs (-.45), *that*-omission (-.48)
	Word length (.40)	Finite adverbial clauses (-.31)
	Passives: finite (.41), postnominal (.32)	
Dimension 2 Information source: Text versus personal experience	Nouns (.37), place nouns (.45), premodifying nouns (.39)	Pronouns: first person (-.33), second person (-.39)
	Third-person pronouns (.41)	Abstract nouns (-.37)
	That-clauses controlled by communication verbs (.68)	
	Communication verbs (.80)	
Dimension 3 Abstract opinion versus concrete description/summary	Word length (.49)	Concrete nouns (-.38) Activity verbs (-.47)
	Nouns: nominalizations (.62), mental nouns (.51), abstract nouns (.38)	
	Noun + *to*-complement clause (.33)	
	Mental verbs (.31)	
Dimension 4 Personal narration	First-person pronouns (.35)	Present-tense verbs (-.70)
	Past-tense verbs (.74)	

The second major step in an MD analysis is to compute factor scores for each text by summing the rates of occurrence of the features having salient loadings on that factor. (The rates of occurrence are standardized before computing factor scores, so that all linguistic features have the same scale, with an overall corpus mean score = 0.0, and units of ±1 representing one standard deviation; see Biber, 1988, 1995.)

As Table 16 shows, all four dimensions are significant and strong predictors of differences among the TOEFL iBT text categories; the GLM models for three of the four dimensions have r^2 values of circa 65%, while the fourth dimension has an r^2 value of almost 50%. Mode (speech versus writing) and task (independent versus integrated) are significant factors for all four dimensions. Score level has a much weaker relationship to these linguistic dimensions: It is a significant predictor only for Dimension 1, and significant in interaction with mode or task for Dimensions 2 and 3 (see discussion below).

Figures 4–7 plot the mean scores for each text category with respect to each dimension. The descriptive statistics for dimension scores, broken down by each text category, are given in Appendix L.

The underlying assumption of MD analysis is that linguistic co-occurrence patterns are functional: Linguistic features occur together in texts because they serve related communicative functions. Dimensions are therefore interpreted in functional terms, based on (a) analysis of the communicative function(s) most widely shared by the set of co-occurring features, and (b) analysis of the similarities and differences among registers with respect to the dimension. In the present case, the following functional labels are proposed:

Dimension 1: Literate versus oral responses

Dimension 2: Information source: text versus personal experience

Dimension 3: Abstract opinion versus concrete description/summary

Dimension 4: Personal narration

Dimension 1 is the easiest to interpret, because it is so similar to Dimension 1 in previous MD studies of other discourse domains (e.g., Biber, 1988, 1995, 2006). Dimension 1 is composed of both positive and negative features: The positive features occur together frequently in texts, and the negative features occur together frequently in texts. The two groupings constitute a single dimension because they occur in complementary distribution: When the positive features occur with a high frequency in a text, that same text will have a low frequency of negative features, and vice versa. Considering both the co-occurring linguistic features that define this dimension, together with the distribution of text categories shown in Figure 4, it is straightforward to propose a functional interpretation for Dimension 1: Literate versus oral tasks.

Table 16

Summary of the Full Factorial Models for Dimensions 1–4

	Model	R^2	Mode (sp/wr)	Task	Score level	Mode* task	Mode* score	Task* score	Mode* task* score	Test taker
Dimension 1: Literate versus oral responses	< . 0001	0.685	< . 0001	< . 0001	< . 01	< . 0001	ns	< . 05	ns	< . 0001
Dimension 2: Information source	< . 0001	0.678	< . 0001	< . 0001	ns	< . 0001	ns	< . 01	ns	ns
Dimension 3: Abstract vs. concrete	< . 0001	0.654	< . 0001	< . 0001	ns	< . 0001	ns	ns	ns	< . 01
Dimension 4: Personal narration	< . 0001	0.485	< . 05	< . 0001	ns	< . 0001	ns	ns	< . 01	ns

Note. Sp/wr = spoken mode/written mode; ns = not significant.

.

The positive features on Dimension 1 are mostly nouns and other features used to modify noun phrases (i.e., nouns premodifying a head noun, attributive adjectives, *of*-phrases, and other prepositional phrases). These features co-occur with long words and passive constructions. A similar grouping of features has been found in previous MD studies, associated with written (as opposed to spoken) registers, and especially associated with informational written registers for specialist readers. (Biber & Gray, 2010, and Biber, Gray & Poonpon, 2011, focused on a similar set of grammatical features to document the surprising fact that the complexity of written academic discourse is phrasal, arguing that the emphasis on dependent clauses in studies of writing development and assessment is misdirected.)

In contrast to the nouns and phrasal structures with positive loadings on Dimension 1, the negative features on this dimension are verbs, pronouns, and clausal structures. In previous MD studies, such features have been associated with speech and with registers having involved communicative purposes.

Dimension 1 in the present study is a very strong predictor of the text category differences found in the TOEFL iBT (cf. Table 16). Figure 4 plots the mean scores for each text category. Written text categories are shown in lower case; spoken text categories are shown in capital letters. Independent task types are shown in bold; integrated task types are shown in italics. Score level is marked by the hyphenated numbers.

Figure 4 shows that all three situational parameters of variation are systematically distinguished along this dimension:

- Written responses are mostly more literate than spoken responses
- Within each mode, integrated responses are mostly more literate than independent responses
- Within each mode/task category, higher scoring responses are mostly more literate than lower scoring responses

The written mode offers the most opportunity for careful production (including revision and editing), permitting the use of a nominal/phrasal discourse style. Integrated tasks have literate textual support (i.e., the reading and listening passages that students comprehend before text production), and those supporting texts apparently also enable more literate grammatical characteristics. Raters are also responsive to these discourse characteristics, so they tend to rate texts with literate Dimension 1 characteristics higher within all four text categories.

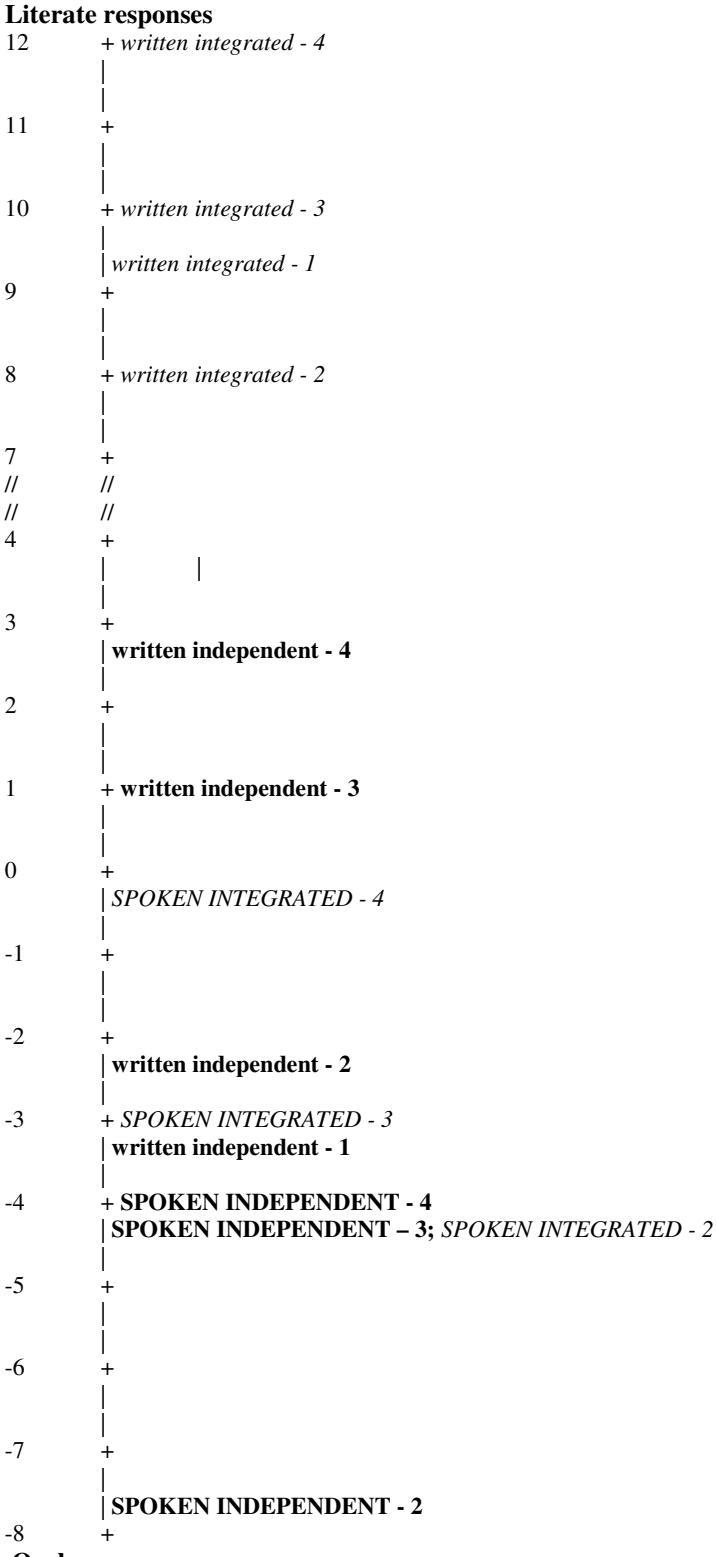

Figure 4. **Mean scores of the TOEFL iBT text categories along Dimension 1: Oral versus literate tasks.**

Interestingly, Dimension 1 corresponds to highly systematic differences across TOEFL iBT score levels. Thus, Level 4 responses are generally the most literate in their Dimension 1 scores, while the lowest scoring responses are generally the most oral. Figure 4 shows that this pattern consistently holds for all four task types. Correlations between Dimension 1 and TOEFL iBT score level, carried out separately for each task type, are all significant at $p < .01$:

Spoken-independent: $r = .19$

Spoken-integrated: $r = .16$

Written-independent: $r = .39$

Written-integrated: $r = .13$

These correlations are only moderately strong, showing that there is considerable variation in the use of Dimension 1 features that is not associated with TOEFL iBT score-level. At the same time, the results here show that the grouping of co-occurring linguistic features on Dimension 1 is a much better predictor of TOEFL score differences than any linguistic feature considered individually.

As noted above, the most important distinction made by Dimension 1 is between speech and writing, and then between integrated versus independent tasks within each mode. In contrast, Figure 5 shows that the top-level distinction made by Dimension 2 is between integrated versus independent tasks, with a secondary distinction between speech and writing within each of the task types. Thus, integrated tasks have positive scores along Dimension 2, with the written integrated tasks having larger positive scores than the spoken integrated tasks. At the other extreme, independent tasks have negative scores along Dimension 2, with the spoken independent tasks having larger negative scores than the written independent tasks.

The positive features defining Dimension 2, associated with the integrated tasks, include nouns, third-person pronouns, and communication verbs (often controlling a *that*-clause). This collection of features is important for describing and summarizing the content of another text: the main communicative goal of integrative tasks. In contrast, the negative features on Dimension 2 include first and second-person pronouns, and abstract nouns. Normally, first and second-person pronouns are associated with highly interactive discourse. In this case, though, these features serve the purpose of talking about typical events and consequences based on the speaker/writer's

own personal experience. First-person pronouns are used in the obvious way to refer directly to the speaker/writer; for example:

I agree with…

I think it is better to…

I like this car…

The most important gift I ever received…

However, second-person pronouns do not have the literal meaning of referring to the addressee. Rather, these pronouns are almost always used with an impersonal third-person meaning, as in the following:

If you sleep in the morning and wake up at night it's not enough sleeping

When you wake up you'll find yourself starved

Nowadays you can find different types of transportation

Nowadays you have many different decisions that you have to make

Thus, considering both the defining linguistic features and the distribution of text categories, Dimension 2 seems to relate primarily to the source of information used for a response, captured in the interpretive label: Information source: Text versus personal experience.

TOEFL iBT score level shows an interesting interaction effect with respect to Dimension 2, with an inverse relationship in independent versus integrated tasks. In both task types, higher scoring responses use these linguistic features to a lesser extent than lower scoring responses. Thus, higher-scoring responses tend toward the unmarked middle of this dimension, while lower-scoring responses tend toward the two extremes of the dimension. This tendency results in opposite characteristics for low-scoring responses in integrated versus independent tasks: Low-scoring integrated responses have the most frequent use of textual information source features, while low-scoring independent responses have the most frequent use of personal experience information source features. In both cases, this pattern might be interpreted as overuse of these features, since high-scoring responses tend to use the features in question to a lesser extent (resulting in Dimension 2 scores closer to 0.0).

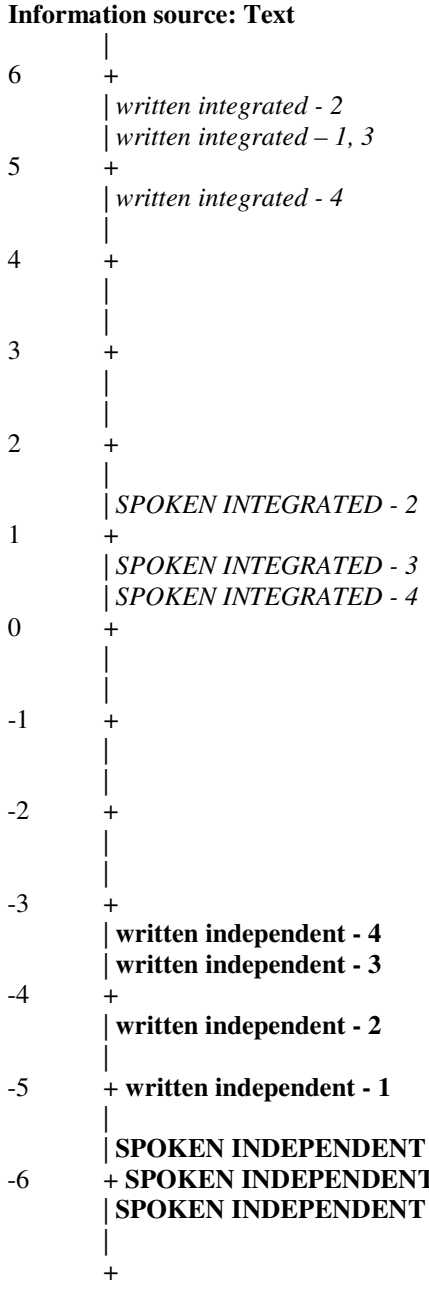

***Figure 5*. Mean scores of the TOEFL iBT text categories along Dimension 2: Information source: Text versus personal experience.**

Dimension 3 is more difficult to interpret. The defining positive linguistic features include long words, nominalizations, special noun classes (mental nouns and abstract nouns), and mental verbs. There are only two negative features: concrete nouns and activity verbs. Thus, the functional opposition here seems to be between abstract content and concrete activities.

Surprisingly, though, written independent responses are by far the most marked for large positive scores along this dimension, as shown in Figure 6. It might be expected that written integrated responses would be more abstract in content than the independent responses, but Figure 6 shows that this is not the case, regardless of score level. At the other extreme, all spoken responses—whether for independent or integrated tasks—have negative scores along Dimension 3, reflecting their focus on concrete activities rather than abstract content. TOEFL iBT score levels have no systematic relationship with Dimension 3 features.

A more detailed consideration of the specific tasks required for each of these categories helps to explain the surprising fact that written independent tasks (rather than integrated tasks) are the most marked for abstract content. Spoken and written independent tasks share the characteristic that the test taker produces a response with no supporting texts. They differ, though, in the specific communicative tasks that are required. In spoken independent tasks, the test taker is asked to give his or her opinion about life choices and normal everyday practices based on personal experiences, such as the best way to relax, or whether it is better to go to bed early or stay up late. In contrast, test takers are asked to give their opinions about larger personal/societal issues in written independent tasks, such as:

> *Do you agree or disagree with the following statement? It is more important to choose to study subjects you are interested in than to choose subjects to prepare for a job or career.*
>
> *Do you agree or disagree with the following statement? In today's world, the ability to cooperate well with others is far more important than it was in the past.*

As a result, spoken versus written independent tasks are polar opposites in their Dimension 3 characteristics. Integrated tasks are less marked along this dimension, because they are so closely tied to the supporting text. Written integrated tasks are more abstract than spoken integrated tasks, presumably because the test takers have more opportunity for planning and careful production, permitting use of longer words (especially nouns). However, because the specific task involved is to summarize the content of an external text, integrated written responses are less abstract with respect to these linguistic features than independent written responses. Thus, considering both the linguistic features as well as the distribution of text

categories, the interpretive label “abstract opinion versus concrete description/summary” can be proposed for Dimension 3.

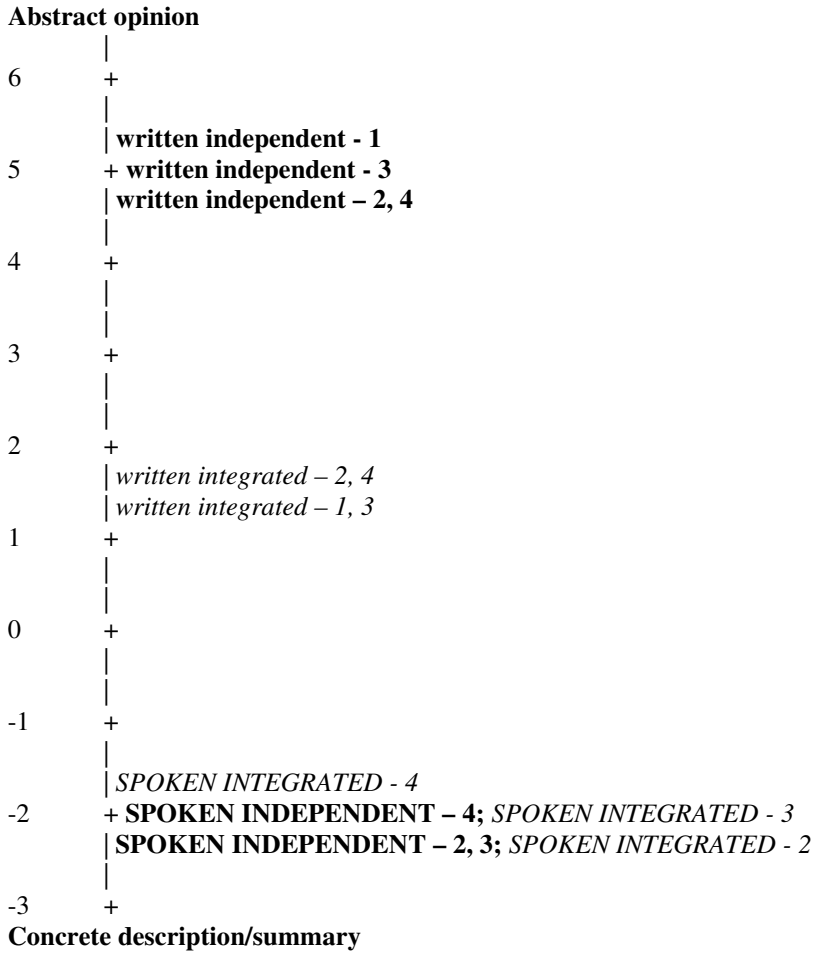

***Figure 6*. Mean scores of the TOEFL iBT text categories along Dimension 3: Abstract opinion versus concrete description/summary.**

Finally, Dimension 4 is defined by only three linguistic features with factor loadings over |.3|: past tense verbs and 1st person pronouns with positive loadings, versus present tense verbs with a negative loading. Normally, a factor should be represented by at least five or six variables with meaningful loadings to enable interpretation. In this case, though, the functional associations are so obvious that the dimension can be interpreted as a personal narration dimension based on only these few features.

Table 16 and Figure 7 show that Dimension 4 is less important for distinguishing among the TOEFL iBT text categories. In general, independent tasks employ these personal narrative features to a greater extent than integrated tasks, with spoken independent tasks being the most

marked. Further, Table 16 shows a significant interaction effect between mode, task type, and score level. This effect is due mostly to differences within the spoken independent tasks (see Figure 7), with higher-scoring responses using these features to a greater extent than lower-scoring responses. Overall, though, this dimension of variation is less important than Dimensions 1–3 in the TOEFL iBT domain.

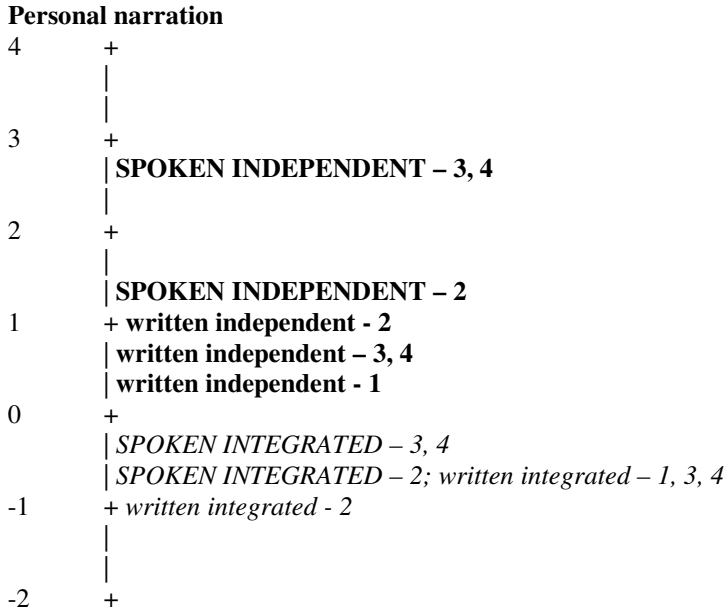

Figure 7. **Mean scores of the TOEFL iBT text categories along Dimension 4: Personal narration.**

6. Discussion and Implications for the TOEFL iBT

As noted in the introduction, the present study set out to investigate three important research questions relating to the discourse characteristics of test taker language production on TOEFL iBT exam responses:

1. Do test takers systematically vary the linguistic characteristics of discourse produced in the spoken versus written modes across different task types? If so, how?
2. In what ways do exam scores correspond to systematic linguistic differences in the discourse produced by test takers?
3. How does the relationship between linguistic discourse characteristics and score level vary across the spoken/written modes and/or task types?

To achieve this goal, we undertook comprehensive linguistic investigations of the discourse produced by TOEFL iBT test takers, categorized according to the mode of production (speech or writing), task type (independent or integrated), and score level. Our linguistic analyses included features at multiple levels, including vocabulary distributions, collocational associations of individual verbs, extended lexical bundles, word class features, simple grammatical devices, and more complex phrasal and clausal structures. In addition, we carried out a MD analysis to identify the underlying parameters of linguistic variation in this discourse domain (the dimensions) and to describe the similarities and differences among TOEFL iBT text categories with respect to each dimension. Based on these analyses, we can now return to the general research questions identified in the introduction.

Do test takers systematically vary the linguistic characteristics of discourse produced in the spoken versus written modes across different task types? If so, how? The answer to this question is clearly yes. In fact, this is by far the strongest general finding from our investigation: TOEFL iBT test takers—at all proficiency levels—demonstrate the ability to vary their linguistic styles across the spoken and written modes and across independent/integrated task types. We found evidence of this ability in all linguistic analyses, including lexical patterns, grammatical variation, and the overall multidimensional patterns of variation. By comparing these specific patterns of linguistic variation to more general patterns identified in previous research, we can conclude that TOEFL iBT test takers vary their linguistic expression in appropriate ways. For example, test takers are more likely to use colloquial features (e.g., pronouns, modal verbs, stance features) in speech than in writing, and they are more likely to use literate grammatical devices (e.g., long words, passive voice verbs, nominalizations) in written responses. Further, test takers employ many of these same linguistic devices to distinguish between independent and integrated tasks: Independent tasks are more personal and involved, and therefore test takers generally use more colloquial features; integrated tasks are more informational, and therefore test takers use more literate features in those tasks.

In terms of the TOEFL validity argument, the findings here provide strong evidence in support of the first two propositions listed in Enright and Tyson (2008, Table 1): "The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings," and "tasks . . . are appropriate for obtaining evidence of test takers' academic language abilities." Previous research provides

detailed descriptions of the patterns of linguistic variation across a very wide range of spoken and written registers (see, e.g., Biber, 1988, 2006; Biber & Conrad, 2009; Biber, Conrad & Cortes, 2004; Biber et al., 1999). In almost every regard, the linguistic patterns of linguistic variation found in TOEFL iBT responses parallel these more general patterns of variation found across spoken and written registers in English.

Interestingly, variation among the TOEFL iBT responses also conforms to the less salient—but equally important—distinction between clausal versus phrasal grammatical complexity. That is, several studies have shown that one of the most important discourse characteristics of academic writing in English is its preference for phrasal modification as opposed to the frequent use of verbs and clausal embedding (see, e.g., Biber & Gray, 2010; Biber, Gray, & Poonpon, 2011). In the present study, both the detailed grammatical analyses as well as the MD analysis show that TOEFL iBT test takers have developed some awareness of these differences and are able to apply them to their own discourse production.

In what ways do TOEFL iBT scores correspond to systematic linguistic differences in the discourse produced by test takers? In contrast to the first research question, our answer to this second question is much less definitive. Overall, we found few general linguistic differences in the discourse produced by test takers from different score levels. The lexical bundle analysis uncovered a general trend showing that the lowest-level responses (and the highest level responses) use lexical bundles to a lesser extent than intermediate-level responses. In the grammatical analysis, there were only four significant main effects for score level: Possibility modals are used more in low-level responses; longer words, attributive adjectives, and verb+*that*-clause constructions are used more in high-level responses. Otherwise, there are no general significant effects for score level. There are, however, some features where score level differences are significant interacting with mode and/or task; these are discussed in the following subsection.

How does the relationship between linguistic discourse characteristics and TOEFL iBT score level vary across the spoken/written modes and/or task types? To the extent that we uncovered systematic lexico-grammatical differences across score levels, they were mostly in interaction with mode and task. Given the fundamental importance of register variation, this is not a surprising finding. That is, the most important linguistic differences found in any discourse domain are associated with register variation (which is in turn associated with mode,

communicative purpose, interactivity, and so forth.). It is thus not surprising that we would find significant differences across score levels within a given mode/task, rather than overall linguistic differences across score levels that apply generally to both spoken and written responses, for both independent and integrated tasks.

Examples of this type emerged from the grammatical analysis, including significant interactions for nonpast tense verbs, first-person pronouns, stance adverbials, desire verb + *to*-clause, word length, and attributive adjectives. The strongest interactions of this type were for finite passive verbs and nonfinite passive relative clauses. The analysis shows that at least some grammatical features vary in a systematic way across score levels in interaction with task type differences. For example, passive verbs are considerably more common in written integrated tasks, and they have a relatively strong association with higher scores within that task type.

The strongest example of this interaction effect came from Dimension 1 in the MD analysis, which was defined by literate linguistic features (e.g., premodifying nouns, attributive adjectives, *of*-phrases, and other prepositional phrases) versus oral linguistic features (e.g., verbs, pronouns, and clausal structures). This linguistic parameter distinguished among all text categories in the TOEFL iBT Corpus: Written responses are more literate than spoken responses; within each mode, integrated responses are more literate than independent responses; and within each mode/task category, higher scoring responses are more literate than lower scoring responses. This pattern existed for the distribution of score levels within all four mode/task type categories.

However, for other linguistic features considered in our analysis (and most dimensions in the MD analysis), there are only small and insignificant differences across score levels, whether considered as a main effect or in interaction with mode and task type. This is the pattern of use for most grammatical complexity features, including nouns, nominalizations, premodifying nouns, prepositional phrases, *of*-genitive phrases, noun complement clauses (both *that*-clauses and *to*-clauses), and finite relative clauses. These are all grammatical features associated with advanced writing development. Further, in the analyses above, these features are all strongly associated with mode and task differences, being generally used more in written integrated tasks. But these features are not associated with TOEFL iBT score level differences.

On first consideration, this finding is surprising: Our prior expectation was that grammatical variation (especially for complexity features) would correlate in systematic ways

with score level. This expectation underlies much of the grammatical research on language development and writing development (e.g., see the summaries of research in Wolfe-Quintero et al., 1998; Ortega, 2003).

However, more careful consideration of the TOEFL validity argument makes it clear that there is little reason to expect that TOEFL iBT scores would correlate with the use of individual grammatical features. Rather, the intended interpretation of those scores is that they "reflect targeted language abilities" (Chapelle et al., 2008, p. 19) which can be "attributed to a construct of academic language proficiency" (p. 20). Scores can then be extrapolated to predict "the quality of linguistic performance" in American universities (p. 21). To achieve these goals, scoring rubrics have been developed and evaluated for each of the four mode/task-type categories of the TOEFL iBT (see Appendix B). TOEFL iBT raters consider a wide range of factors, including the overall content, relevance of the response to the assigned task, fluency (in speech), coherence and clear progression of ideas, word choice, and control of grammatical structures. Similar to the evaluations provided by instructors in actual university courses, ratings in the TOEFL iBT context are carried out holistically, assigning a single score to each response.

Raters generally have high agreement in their assignment of holistic scores (see, e.g., Chapelle et al., 2008, Chapters 5 and 6), indicating that a general construct of academic language proficiency is reliably assessed by these scores. However, it is less clear what the specific considerations are that influence raters. Thus, Lumley (2002, p. 246) noted that "the process of rating written language performance is still not well understood" and that "the relationship between scale contents and text quality remains obscure." Studies of rater cognition generally employ think-aloud methods to identify the considerations that are most influential for raters (Lumley, 2002; Cumming, Kantor, & Powers, 2002). However, there has been less direct empirical research to manipulate specific linguistic characteristics of texts and determine the effects of such linguistic variation on holistic ratings. Thus, while we know a great deal about the stability of holistic ratings and the reported cognitive processes of raters, we know less about the specific linguistic characteristics of texts that are most influential to raters.

Within the context of the validity argument for the TOEFL iBT, the important point to note here is that score levels are not intended to directly measure linguistic development in the use of particular lexico-grammatical features. For example, the rubric for evaluating written independent responses includes several different characteristics, including the extent to which the

response addresses the topic/task, is well organized and developed, uses clear explanations and examples, is coherent, and so forth. The use of appropriate lexico-grammatical features is given comparatively little weight in the rubric, with a mention of the preference for syntactic variety and the avoidance of lexical and grammatical errors (see Appendix B). Raters must consider this full array of considerations to assign a single holistic rating to each written response. It is thus not surprising that these ratings do not correlate with development in the use of individual complexity features.

A similar set of considerations is specified in the rubric for the assessment of written integrated responses, with a focus on the well-organized presentation of information that is complete, accurate, and coherent. Grammatical errors are associated with lower score levels, but otherwise the use of lexico-grammatical features is not mentioned in this rubric.

The range of criteria considered in the rating of spoken responses is broader than those considered for written responses, with three major subcategories explicitly noted: delivery (speech is clear and fluid, intelligibility high); language use (control of basic and complex grammatical structures, effective word choice); and topic development (clear progression of ideas, appropriate detail). In this case, raters are trained to evaluate the use of lexico-grammatical features. However, each spoken response is given a single omnibus rating, reflecting the combined assessment of all three major criteria.

As noted above, the scoring of TOEFL iBT responses is very similar to the evaluations that instructors provide on academic tasks in actual university courses. That is, evaluations of discourse in university courses focus primarily on the content: whether the presentation of information is clear, coherent, well organized, well illustrated, and so forth. To the extent that grammar is overtly considered, the focus is on errors or occasionally prescriptive grammar rules. Instructors might be influenced by the use of more complex grammatical constructions, but they are unlikely to have conscious awareness of those patterns. Thus, similar to TOEFL iBT scores, instructor evaluations in the wider university context generally focus much more on content and organization than on the use of any individual grammatical feature.

In contrast, the focus of the present investigation has been on the lexico-grammatical characteristics that are associated with register variation. The findings here show that these grammatical features are important in the TOEFL iBT context for their ability to discriminate among independent versus integrated task types in speech versus writing. Advanced language

learners clearly develop proficiency in the use of these grammatical features. But this linguistic development does not have a direct relationship to the general construct of academic language proficiency as measured by holistic scores on the TOEFL iBT. As a result, most grammatical features (including complexity features) show little relationship to the holistic ratings of quality represented by TOEFL iBT scores.

It is possible that development in the use of complex grammatical features becomes a more important consideration at more advanced levels of academic performance. That is, more complex grammatical constructions are required to present advanced academic content in clear, efficient ways. Successful advanced writers of academic research writing make the transition to phrasal styles of discourse, rather than employing the clausal styles typical of speech and written narration (see, e.g., Biber & Gray, 2010; Biber et al., 2011). As a result, content considerations merge with grammatical considerations at higher levels of academic performance, making it likely that instructors pay greater consideration to the use of complex grammatical features in the evaluation of such tasks. At present, we have no direct evidence in support of this possibility, but the findings here suggest that this should be an important area for future research.

In sum, the findings here have shown that there is significant and extensive linguistic variation among TOEFL iBT texts corresponding to differences between independent and integrated tasks in the spoken and written modes. These findings strongly support the TOEFL validity argument that this range of tasks is required to capture the breadth of academic expectations in American universities. Future research is recommended to further investigate the evaluation criteria applied at different academic levels and, in particular, whether the use of complex grammatical features becomes a more relevant consideration at higher levels of performance.

References

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–195.

Biber, D. (1988). *Variation across speech and writing.* Cambridge, UK: Cambridge University Press.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, *5*, 257–269.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 243–257.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam, the Netherlands: John Benjamins.

Biber, D. (2009). Are there linguistic consequences of literacy? Comparing the potentials of language use in speech and writing. In D. R. Olson & N. Torrance (Eds.), *Cambridge handbook of literacy* (pp. 75–91). Cambridge, UK: Cambridge University Press.

Biber, D., & S. Conrad. (2009). *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*, 9–48.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., …, & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* (TOEFL Monograph Series No.MS-25). Princeton, NJ: Educational Testing Service.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*, 2–20.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* London, UK: Longman.

Biber, D., & Jones, J. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1286–1304). Berlin, Germany: Walter de Gruyter.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series No. MS-29). Princeton, NJ: Educational Testing Service.

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

Coffin, C. (2004). Arguing how the world is or how the world should be: The role of argument in IELTS tests. *Journal of English for Academic Purposes*, *3*, 229–246.

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *Journal of Educational Research*, *69*, 176–183.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*, 397–423.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*(2), 213–238.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*, 5–43.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph Series No. MS-30). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67–96.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, *26*, 59–84.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155.

Enright, M., & Tyson, E. (2008). Validity evidence supporting the interpretation and use of TOEFL iBT scores. *TOEFL iBT Research Insight*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf

Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*(2), 414–420.

Ferris, D., & Politzer, R. (1981). Effects of early and delayed second language acquisition: English composition skills of Spanish-speaking junior high school students. *TESOL Quarterly*, *15*, 263–274.

Flahive, D., & Snow, B. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 171–176). Rowley, MA: Newbury House.

Gipps, C., & Ewen, E. (1974). Scoring written work in English as a second language. *Educational Research*, *16*, 121–125.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123–145.

Hirose, K. (2003). Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students. *Journal of Second Language Writing*, *12*, 181–209.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*, 4–21.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner composition. *Journal of Second Language Writing*, *12*, 377–403.

Kubota, R. (1998). An investigation of L1-L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of Second Language Writing, 7*(1), 69–100.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*, 590–619.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

Li, Y. (2000). Linguistic characteristics of ESL writing in task-based e-mail activities. *System*, *28*, 229–245.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276.

Nation, P. (1990). *Teaching and learning vocabulary.* Boston, MA: Heinle ELT.

Nelson, N. W., & Van Meter, A. M. (2007). Measuring written language ability in narrative samples. *Reading and Writing Quarterly*, *23*, 287–309.

Norrby, C., & Håkansson, G. (2007). The interaction of complexity and grammatical processability: The case of Swedish as a foreign language. *International Review of Applied Linguistics*, *45*, 45–68.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518.

Partington, A. (1998). *Patterns and meanings.* Amsterdam, the Netherlands: John Benjamins.

Purpura, J. (2004). *Assessing grammar.* Cambridge, UK: Cambridge University Press.

West, M. (1953). *General service list of English words*. London, UK: Longman.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching*. Cambridge, UK: Cambridge University Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity* (Technical Report No. 17). Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii.

List of Appendices

Appendix A

Prompts and Questions for the Responses Used in the TOEFL iBT Corpus

TOEFL iBT Dataset—Speaking Form 1

Independent Tasks

Form 1, Question 1

Question:

Students work hard but they also need to relax. What do you think is the best way for a student to relax after working hard? Explain why.

Preparation Time: 15 seconds

Record Time: 45 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

Form 1, Question 2

Question:

Some people think it is alright to stay up late at night and sleep late in the morning. Others think it is better to go to bed early at night and wake up early. Which view do you agree with? Explain why.

Preparation Time: 15 seconds

Record Time: 45 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

Integrated Tasks

Form 1, Question 3

Reading:

The university is planning some changes in the appearance of the student library. Read the article in the student newspaper about the change. You will have 45 seconds to read the article. Begin reading now.

Library Lobby to be Renovated

Macpherson Library may become a more pleasant place to study, thanks to renovations currently being planned by library staff. “First, we plan to remove that dirty, dingy carpet in the lobby so we can restore and polish the natural wooden floors underneath it,” said Jeff Rosenthal, head librarian. “We’re also commissioning a local artist to paint a mural on the wall facing the entrance,” he added. A recent survey of students revealed that one major concern they have is that library facilities are outdated. Library officials believe that concerns revealed in the survey will be addressed by the trendy renovations being planned for the lobby.

Listening:

Now listen to two students discussing the article.

Audio:

Female student: Oh, did you read that article about the library? That should look nice.

Male student: Yeah, it may look nice, but—

Female student: But what? You sound skeptical.

Male student: Well—first of all—wooden floors are noisy. Can you imagine people walking around on wooden floors when you’re trying to study? That’s gonna echo through the whole building!

Female student: Yeah, you have a point there.

Male student: And the painted mural? I mean, who really cares about that? I can’t believe they really think that’s important.

Female student: Well, they think students wanted stuff like that.

Male student: Listen. When they asked us those questions, we listed all kinds of concerns, like—we talked about how we need a lot of new materials in the reference section. In fact, most of us listed a lot of other concerns. I can’t believe they picked this one thing to address instead of more important concerns.

Female student: I see what you mean.

Question:

The man discussed his opinion of the library's plan. Describe his opinion and his reasons for holding that opinion.

Preparation Time: 30 seconds

Record Time: 60 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

Form 1, Question 4

Reading:

Read the passage from a biology textbook. You have 50 seconds to read the passage. Begin reading now.

Allergies

The human body has a defense mechanism to protect itself against invading dangerous substances. However, the immune system, as this mechanism is called, is so sensitive in some people that it can react mistakenly. The result is what we call an allergic reaction. Allergic reactions occur when the immune system tries to fight off a normally harmless substance, or allergen, that has entered the body. Rather than treating the allergen as a harmless substance, which it is for most people, the immune system considers it a threat and mounts a biological defense against it. The unpleasant symptoms that an individual with allergies experiences all result from the body's attempt to fight off a nonexistent threat.

Lecture:

Now listen to part of a lecture in a biology class on this topic.

Audio

(Professor) As an undergraduate student, I shared a dorm room with a guy named Joe. Well, there wasn't a day that went by without Joe having a runny nose, or watery eyes, and he just couldn't stop sneezing. One day Joe told me that the sneezing and all the other stuff was the result of him being oversensitive to dust. But as I found out later, it wasn't

actually the dust itself that Joe was allergic to. It's what's in dust. There are these creatures called "dust mites" that live in it.

(Professor) Now these dust mites contain and release proteins that are light enough to float in the air. These proteins enter our body when we breathe. Generally, that's not a problem, 'cause most peoples' immune systems don't recognize the proteins from the dust mites as a threat. But even though my immune system knew this, Joe's immune system didn't, and so it started making antibodies—uh, substances the body normally uses to fight invaders.

(Professor) Now, the antibodies cause certain cells in the body to release chemicals and those chemicals are what irritate the nose, eyes, and throat. And that's when people like Joe experience an allergic reaction. You know, the runny nose, watery eyes, and those uncontrollable sneezing attacks.

Question:

Using the example given by the professor, explain what causes an allergic reaction.

Preparation Time: 30 seconds

Record Time: 60 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

Form 1, Question 5

Listening:

Listen to a conversation between two professors.

Audio

(Female professor) David, do you have a minute? I'd like your advice on something.

(Male professor) Sure Catherine, what's going on?

(Female professor) Well, there's a situation with one of my students. I think you know Kerry—she's a third-year student in the department?

(Male professor) Oh sure, yeah, she's taking my seminar on 20th century art.

(Female professor) Okay, well, ya know she's a bright student and generally does well although she's a little overextended. Probably taking one too many classes. Not to mention being on the swim team and in clubs . . . you know, you know the type.

(Male professor) Yeah, uh what's the issue?
(Female professor) Well, I need to decide what to do for her. You see, she told me yesterday that she would be out of town for an important swim team competition on the day of the midterm exam.
(Male professor) Oh—
(Female professor) Right. The art history department has never allowed makeup exams. It's a long-standing rule, but of course I wanna help her.
(Male professor) Huh, I guess you could make an exception . . . change the rule just this once, given that she has a legitimate excuse.
(Female professor) Yeah, that's possible. But I worry that other students will start to ask for makeup exams too.
(Male professor) Yeah, hmm. Or you could have her do a writing assignment. Most other students would not prefer that over an exam. I doubt they'd start requesting that.
(Female professor) I also thought of that. But I wonder if it would disadvantage Kerry a bit. Like—like I said, she's involved in a lot of classes and activities. She might not have enough time to do her best work on a paper—something that's extra.
(Male professor) I hear you. It's not an easy decision.

Question:
The speakers discuss two possible solutions to the woman's problem. Briefly summarize the problem. Then state which solution you prefer and explain why.
Preparation Time: 20 seconds
Record Time: 60 seconds
Begin speaking after the **Preparation Time** has ended.
Answer the question now.

Form 1, Question 6
Lecture:
Listen to part of a talk in an anthropology class.
Audio
(Professor) In all cultures or communities, there are recognized rules or norms for how people share or exchange goods and services. "Reciprocity" is a term used by

anthropologists to describe the ways people informally exchange goods within a society. It refers to the giving and receiving of goods, or even gifts, among members of a society. Anthropologists have identified several distinct types of reciprocity.

(Professor) The first type I want to talk about is "generalized" reciprocity. This is when people give each other goods or gifts without expecting anything in return immediately. There is a sort of understanding or social contract here, though. It's understood that at some later time in the future, the act will be reciprocated—that the giver will eventually get something in return. This type of exchange usually takes place among people who are more socially close—among family members, close friends, and so on. Say your brother has just moved into a new house. You know he doesn't have a lot of money right now, and he needs furniture. So you decide to help him out, and buy him a new bed. You don't expect anything in return, but you that someday he'll do something to help you out when you need it. Generalized reciprocity only works among people who are close because it requires a high level of trust.

(Professor) Now, a second form of reciprocity is balanced reciprocity. Balanced reciprocity is a more straightforward exchange of goods. The goods being exchanged are of similar value, and in addition, there's an explicit expectation of return—either immediately or at some specified time in the future. One gives something and knows when to expect that something of similar value will be returned. Here the social distance between giver and receiver is greater than with generalized reciprocity. Let's say this time it's your neighbor that needs a new bed, and you just happen to have one that you weren't using. So you offer to give your neighbor your extra bed. And, your neighbor understands that he or she is expected to repay you for the bed in some way. Maybe in money, maybe by giving you something of equal value. And if that doesn't happen, the relationship will suffer.

Question:

Using the examples from the talk, explain what is meant by generalized reciprocity and balanced reciprocity.

Preparation Time: 20 seconds

Record Time: 60 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

TOEFL iBT Dataset—Writing Form 1

Writing Section Directions (Overview)

This section measures your ability to use writing to communicate in an academic environment. There will be two writing tasks.

For the first writing task, you will read a passage, listen to a lecture, and then answer a question based on what you have read and heard.

For the second writing task you will answer a question based on your own knowledge and experience.

Integrated Task

Writing Section Directions (Question 1)

For this task, you will read a passage about an academic topic. A clock at the top of the screen will show how much time you have to read. You may take notes on the passage while you read. The passage will then be removed and you will listen to a lecture about the same topic. While you listen you may also take notes. You will be able to see the reading passage again when it is time for you to write. You may use your notes to help you answer the question.

You will then have to write a response to a question that asks you about the relationship between the lecture you heard and the reading passage. Try to answer the question as completely as possible using information from the reading passage and the lecture. The question does not ask you to express your personal opinion. Your response will be judged on the quality of your writing, and on the completeness and accuracy of the content. Immediately after the reading time ends the lecture will begin, so keep your headset on until the lecture is over.

Form 1, Writing Question 1

Reading:

For years, the ability of migrating birds to accurately navigate extremely long journeys has puzzled naturalists. Several different theories attempt to account for the birds‘ navigational abilities. One theory suggests that birds navigate in reference to celestial

objects like the sun or the stars. For example, some evidence seems to indicate that birds that migrate by day stay on course by orienting their flight relative to the sun's east/west path across the sky. Birds that migrate at night are thought to use the stars as a map. These birds can locate themselves in relation to the North Star. To migrate directly north, for example, the birds would keep the North Star directly in front of them. Another theory claims that birds navigate by landmarks like rivers, coastlines, and mountains. Studies have linked birds' navigational ability to the hippocampal region of the brain—the region that plays an important role in memory formation. When a bird's hippocampal region is damaged, the bird cannot perform well in tasks testing spatial ability and memory. At the same time, its ability to navigate is impaired as well. Therefore, researchers conclude, migrating birds must be using memorization skills, such as remembering landmarks, to navigate. A third theory proposes that birds use a type of internal compass that responds to Earth's magnetic field. According to this theory, birds have crystals of the mineral magnetite embedded in their beaks. Magnetite, as the name suggests, is magnetic. Supposedly, the birds can sense the way Earth's magnetic field pulls on the magnetite crystals. Sensing the direction of the pull on the crystals is like looking at a compass whose magnetized needle aligns itself with Earth's magnetic field. Thus, according to this theory, magnetite crystals serve birds as an internal compass.

Listening:

Now listen to part of a lecture on the topic you just read about.

Audio

(Professor) Each of the three theories about how birds navigate has some support. None of them explains all the situations in which birds can navigate. So each is at best a partial explanation.

(Professor) The first theory is limited by one simple observation: The sun and stars are not always visible. Obviously, they're often obscured by clouds. The fact is that many birds are able to navigate their migration accurately, even when they can't see the sun or stars. This doesn't mean that observations of celestial objects are not used by birds, but it can't be the whole story.

(Professor) The memorized landmark explanation is also limited. If it were the whole story, then birds—taken to a place they've never been—would be unable to find their

way back home or to the destination of migration. The reason would be that their memories wouldn't correspond to landmarks in the new location. But in many studies, researchers have released birds in locations that were unknown to the birds and yet the birds were still unable to navigate their way back to their nests. So birds do not rely on memorized landmarks only.

(Professor) The third explanation about magnetic crystals in birds' beaks couldn't be a complete explanation, either. Birds may use earth's magnetic field, but a compass is not enough. Just knowing that you're headed south doesn't get you to any particular place. Minimally, you still need to know where you are when you begin the journey, and how far that is from your destination. A built-in compass—as amazing as that sounds—cannot account for bird migration by itself.

Question:

Summarize the points made in the lecture, being sure to explain how they present limitations of the theories discussed in the reading passage.

Independent Task

Writing Section Directions (Question 2)

In this section you will demonstrate your ability to write an essay in response to a question that asks you to express and support your opinion about a topic or issue.

The question will be presented on the next screen and will remain available to you as you write.

Your essay will be scored on the quality of your writing. This includes the development of your ideas, the organization of your essay, and the quality and accuracy of the language you use to express your ideas. Typically an effective essay will contain a minimum of 300 words.

You will have **30 minutes** to plan, write, and revise your essay. If you finish your response before time is up, you may click on **Next** to end this section.

Form 1, Writing Question 2

Question:

Do you agree or disagree with the following statement? It is more important to choose to study subjects you are interested in than to choose subjects to prepare for a job or career. Use specific reasons and examples to support your answer.

TOEFL iBT Dataset—Speaking Form 2

Independent Tasks

Form 2, Question 1

Question:

Talk about the most important gift you have ever received. Describe the gift and explain why it was significant.

Preparation Time: 15 seconds

Record Time: 45 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

Form 2, Question 2

Question:

Do you think your life is easier or more difficult than your grandparents' lives? Use examples and details to explain your answer.

Preparation Time: 15 seconds

Record Time: 45 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

Integrated Tasks

Form 2, Question 3

Reading:

Read the following letter to the Oakdale University student newspaper. You have 50 seconds to read the letter. Begin reading now.

Letter from a Former Oakdale Student

I was a student at Oakdale University 25 years ago. Since then I have had numerous jobs, and the reason I was not more successful is that I never learned how to use a computer. So to help you avoid the problems I had, I want to make a suggestion. That's why I am writing to the campus newspaper. The suggestion is for Oakdale to require all students to take an introductory computer class to learn basic computer skills. To be successful, you need to know how to use a computer. And because computer skills are so important, it should be Oakdale's responsibility to make sure that all of you are taught them before you graduate.

Listening:

Now listen to a conversation between two students discussing the letter.

Audio

(Female student) I think this gentleman's got it all wrong.

(Male student) I agree.

(Female student) I mean—this uh, Mr. Wilson—he's right about people needing computer skills. But the problem is he graduated from Oakdale 25 years ago. Many people weren't exposed to computers back then.

(Male student) It sure is different today.

(Female student) Yes, today everyone at Oakdale knows how to use a computer. We use computers in just about every class we take.

(Male student) That's true. Like even in English and history. Even in art classes.

(Female student) Right. You can't graduate from Oakdale today without having developed computer skills along the way as part of your regular coursework.

(Male student) Who needs a special class?

(Female student) Agreed! Certainly not a required class. And the other thing is—is it really realistic to expect the university to teach people everything they're ever going to need know? Your education doesn't stop on graduation day. If you find out later that there's something that you still need to know, you can always take adult education courses.

(Male student) So you're saying—

(Female student) I'm saying—with all due respect—if Mr. Wilson never learned how to use computers properly, it's not the university's fault.

Question:

The woman expresses her opinion about the suggestion from the former student. State the woman's opinion and explain the reasons she gives for holding that opinion.

Preparation Time: 30 seconds

Record Time: 60 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

Form 2, Speaking Set 4

Reading:

Now, read a passage from a text about business decision-making. You will have 50 seconds to read the passage. Begin reading now.

Sunk Costs

When individuals invest money in a project, their money can sometimes be recovered if the project is discontinued. However sometimes money that is invested cannot be recovered. In that case it is considered a "sunk cost": even if the project is abandoned, the money is lost. These sunk costs can affect people's decisions. Economists have noticed that when there are sunk costs, people often continue projects that should be discontinued. Even when a project seems unlikely to provide a benefit, people will stick with it because of the money they have already spent.

Lecture:

Now listen to part of a lecture on this topic in an Economics class.

Audio

(Professor) Say you decide to treat yourself and buy a ticket to a football game. You spend quite a bit of money because you want a really good seat. But when the night of the game rolls around, it's freezing cold and snowing. You really hate the idea of sitting out there in that outdoor stadium getting all cold and wet. And besides, the game's gonna be on TV.

(Professor) So what you really want is to stay home and watch the game from your warm, cozy living room. But if you're like most people, you'll find yourself thinking, "I spent so much money on that ticket, I've gotta go to the game." And for that reason alone, you might make yourself go out and endure the miserable weather when you could be watching the same game at home.
(Professor) So at the end of the day, you've paid for your ticket, and you've gotten cold and wet. Not a great deal, right? But what if you'd done the opposite? You'd paid for the ticket, and then had a nice warm evening in front of the TV. Look, the ticket is already paid for and you don't get your money back no matter what you do. So what's the point of having a cold, unpleasant time out there in the snow when staying home would make you much happier?

Question:
Using the example given by the professor, explain what sunk costs are and how they affect people's decisions.
Preparation Time: 30 seconds
Record Time: 60 seconds
Begin speaking after the **Preparation Time** has ended.
Answer the question now.

Form 2, Question 5
Listening:
Now listen to a conversation between two students.
Audio
(Female student) Hi, David. You're coming to the review session Saturday morning, right?
(Male student) The review for the physics exam? I'd really like to. We've covered so much material this semester. It'd be really helpful to go over everything before the test.
(Female student) Yeah, I know. I'm so glad the professor scheduled this review.
(Male student) I just wish he scheduled it for a different day. I promised my cousin Janet I'd help her move into her new apartment on Saturday. She asked me like a month ago.

(Female student) Oh—well, the review session is in the morning, and should only last a couple of hours. Couldn't you help your cousin that afternoon, after the review?
(Male student) Well, yeah—and I'm sure Janet would understand. But I know she really wanted to get an early start. Moving can take a long time, you know? And she has a lot of stuff.
(Female student) Hmm, well, if you decide not to come to the review, you're welcome to borrow my notes.
(Male student) Thanks. That's a really nice offer. I know you take great notes. The only is—if I'm not there, I won't be able to ask any questions—make sure I understand everything.
(Female student) True, it's not like being there. But you know, I'd be glad to answer your questions—I mean, if I can.
(Male student) Well, let me think about it, and I'll get back to you.
(Female student) OK—good luck.

Question:
Briefly summarize the man's problem. Then state which solution you would recommend. Explain the reasons for your recommendation.
Preparation Time: 20 seconds
Record Time: 60 seconds
Begin speaking after the **Preparation Time** has ended.
Answer the question now.

Form 2, Question 6
Lecture:
Now listen to part of a talk in a botany class.
Audio
(Professor) You've probably all seen old trees that are covered with fungus. That look like they have little mushrooms sticking to them? Now, there's also fungus inside an old tree. You might think the fungus is harming the tree, but actually, fungus indirectly helps the tree—brings benefits to it.

(Professor) See—the trunk of an old tree is full of dead wood, and dead wood's useless to the tree. Fungus feeds on that dead wood. It literally eats it up, and the trunk becomes empty inside—hollow—and being hollow helps the tree in a couple of ways.
(Professor) For example, after fungus eats away the dead wood, well, you'd think that being hollow would make the tree weak. But actually, a hollow tree is very stable. It's so much lighter with the dead wood eaten away that its roots—uh, under the ground—can anchor the tree very well. So it won't blow over as easily in a strong wind. And when a big storm arrives, old hollow trees are often left standing because their roots hold them in place. But younger trees, which are too heavy for their roots—they may fall over.
(Professor) Another benefit is that once the fungus makes the tree hollow, that big hollow tree makes a great home for animals. Birds make their nests there, for example. And the tree is a shelter for other animals.
(Professor) Now how does this help the tree? Well, animals produce waste matter. And if they live in the tree, they're gonna leave their waste there. These waste products are food for the tree. They get broken down, and the tree absorbs the nutrients from the animals' waste products through its roots. Trees get important nutrients as a result of the animals that live inside them.

Question:

Using points and examples from the talk, explain two ways that fungus indirectly benefits trees.

Preparation Time: 20 seconds

Record Time: 60 seconds

Begin speaking after the **Preparation Time** has ended.

Answer the question now.

TOEFL iBT Dataset—Writing Form 2

Writing Section Directions (Overview)

This section measures your ability to use writing to communicate in an academic environment. There will be two writing tasks.

For the first writing task, you will read a passage, listen to a lecture, and then answer a

question based on what you have read and heard.

For the second writing task you will answer a question based on your own knowledge and experience.

Integrated Task

Writing Section Directions (Question 1)

For this task, you will read a passage about an academic topic. A clock at the top of the screen will show how much time you have to read. You may take notes on the passage while you read. The passage will then be removed and you will listen to a lecture about the same topic. While you listen you may also take notes. You will be able to see the reading passage again when it is time for you to write. You may use your notes to help you answer the question.

You will then have to write a response to a question that asks you about the relationship between the lecture you heard and the reading passage. Try to answer the question as completely as possible using information from the reading passage and the lecture. The question does not ask you to express your personal opinion. Your response will be judged on the quality of your writing and on the completeness and accuracy of the content. Immediately after the reading time ends the lecture will begin, so keep your headset on until the lecture is over.

Form 2, Writing Question 1

Reading:

Since the 1960s, fish farming—the growing and harvesting of fish in enclosures near the shoreline—has become an increasingly common method of commercial fish production. In fact, almost one third of the fish consumed today are grown on these farms. Unfortunately fish farming brings with it a number of harmful consequences and should be discontinued. One problem with fish farming is that it jeopardizes the health of wild fish in the area around the farm. When large numbers of fish are confined to a relatively small area like the enclosures used in farming, they tend to develop diseases and parasitic infections. Although farmers can use medicines to help their own fish, these illnesses can easily spread to wild fish in the surrounding waters, and can endanger the local populations of those species. In addition, farm-raised fish may pose a health risk to

human consumers. In order to produce bigger fish faster, farmers often feed their fish growth-inducing chemicals. However, the effects of these substances on the humans who eat the fish have not been determined. It is quite possible that these people could be exposed to harmful or unnatural long-term effects. A third negative consequence of fish farming relates to the long-term wastefulness of the process. These fish are often fed with fish meal, a food made by processing wild fish. Fish farmers must use several pounds of fish meal in order to produce one pound of farmed fish. So producing huge numbers of farm-raised fish actually reduces the protein available from the sea.

Listening:

Now, listen to part of a lecture on the topic you just read about.

Audio

(Professor) The reading passage makes it seem that fish farming is a reckless, harmful enterprise. But each of the arguments the reading passage makes against fish farming can be rebutted.

(Professor) First, what are the wild, local fish that fish farms are supposed to harm? The fact is that in many coastal areas, local populations of wild fish were already endangered—not from farming, but from traditional commercial fishing. Fish farming is an alternative to catching wild fish. And with less commercial fishing, populations of local species can rebound. The positive effect of fish farming on local, wild fish populations is much more important than the danger of infection.

(Professor) Second, let's be realistic about the chemicals used in fish farm production. Sure, farmers use some of these substances. But the same can be said for most of the poultry, beef, and pork that consumers eat. In fact, rather than comparing wild fish with farm fish as the reading does, we should be comparing the consumption of fish with the consumption of these other foods. Fish has less fat and better nutritional value than the other farm-raised products, so consumers of farm-raised fish are actually doing themselves a favor in terms of health.

(Professor) Finally, the reading makes claims that fish farming is wasteful. It's true that some species of farm-raised fish are fed fishmeal. But the species of fish used for fishmeal are not usually eaten by humans. So fish farming is a way of turning inedible

fish into edible fish. Contrary to what the reading says, fish farming increases the number of edible fish, and that's what's important.

Question:

Summarize the points made in the lecture, being sure to explain how they challenge the specific points made in the reading passage.

Independent Task

Writing Section Directions (Question 2)

In this section you will demonstrate your ability to write an essay in response to a question that asks you to express and support your opinion about a topic or issue. The question will be presented on the next screen and will remain available to you as you write.

Your essay will be scored on the quality of your writing. This includes the development of your ideas, the organization of your essay, and the quality and accuracy of the language you use to express your ideas. Typically an effective essay will contain a minimum of 300 words.

You will have **30 minutes** to plan, write, and revise your essay. If you finish your response before time is up, you may click on **Next** to end this section.

Form 2, Writing Question 2

Question:

Do you agree or disagree with the following statement? In today's world, the ability to cooperate well with others is far more important than it was in the past. Use specific reasons and examples to support your answer.

Appendix B

Scoring Rubrics for TOEFL iBT Text Categories

TOEFL iBT Test Independent Speaking Rubrics (Scoring Standards)

Score	General description	Delivery	Language use	Topic development
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, and usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.

Score	General description	Delivery	Language use	Topic development
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress, and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; there are frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limits (or prevents) expression of ideas and connections among ideas. Some low level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

TOEFL iBT Test Integrated Speaking Rubrics (Scoring Standards)

Score	General description	Delivery	Language use	Topic development
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Speech is generally clear, fluid and sustained. It may include minor lapses or minor difficulties with pronunciation or intonation. Pace may vary at times as speaker attempts to recall information. Overall intelligibility remains high.	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relevant ideas. Contains generally effective word choice. Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning).	The response presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate detail, though it may have minor errors or minor omissions.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation or pacing and may require some listener effort at times. Overall intelligibility remains good, however.	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message.	The response is sustained and conveys relevant information required by the task. However, it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas.

Score	General description	Delivery	Language use	Topic development
2	The response is connected to the task, though it may be missing some relevant information or contain inaccuracies. It contains some intelligible speech, but at times problems with intelligibility and/or overall coherence may obscure meaning. A response at this level is characterized by at least two of the following:	Speech is clear at times, though it exhibits problems with pronunciation, intonation or pacing and so may require significant listener effort. Speech may not be sustained at a consistent level throughout. Problems with intelligibility may obscure meaning in places (but not throughout).	The response is limited in the range and control of vocabulary and grammar demonstrated (some complex structures may be used, but typically contain errors). This results in limited or vague expression of relevant ideas and imprecise or inaccurate connections. Automaticity of expression may only be evident at the phrasal level.	The response conveys some relevant information but is clearly incomplete or inaccurate. It is incomplete if it omits key ideas, makes vague reference to key ideas, or demonstrates limited development of important information. An inaccurate response demonstrates misunderstanding of key ideas from the stimulus. Typically, ideas expressed may not be well connected or cohesive so that familiarity with the stimulus is necessary in order to follow what is being discussed.
1	The response is very limited in content or coherence or is only minimally connected to the task. Speech may be largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning. Delivery is choppy, fragmented, or telegraphic. Speech contains frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limits (or prevents) expression of ideas and connections among ideas. Some very low-level responses may rely on isolated words or short utterances to communicate ideas.	The response fails to provide much relevant content. Ideas that are expressed are often inaccurate, limited to vague utterances, or repetitions (including repetition of prompt).
0	Speaker makes no attempt to respond or response is unrelated to the topic.			

TOEFL iBT Test Integrated Writing Rubrics (Scoring Standards)

Score	Task description
5	A response at this level successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.
4	A response at this level is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information in relation to the relevant information in the reading, but it may have minor omission, in accuracy, vagueness, or imprecision of some content from the lecture or in connection to points made in the reading. A response is also scored at this level if it has more frequent or noticeable minor language errors, as long as such usage and grammatical structures do not result in anything more than an occasional lapse of clarity or in the connection of ideas.
3	A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following: Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading The response may omit one major key point made in the lecture Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise
2	A response at this level contains some relevant information from the lecture, but is marked by significant language difficulties or by significant omission or inaccuracy or important ideas from the lecture or in the connections between the lecture and the reading; a response at this level is marked by one of the following: The response contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture The response contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture
1	A response at this level is marked by one of more of the following: The response provides little or no meaningful or relevant coherent content from the lecture The language level of the response is so low that it is difficult to derive meaning
0	A response at this level merely copies sentences from the reading, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

TOEFL iBT Test Independent Writing Rubrics (Scoring Standards)

Score	Task description
5	An essay at this level largely accomplishes all of the following: Effectively addresses the topic and task Is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details Displays unity, progression, and coherence Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4	An essay at this level largely accomplishes all of the following: Addresses the topic and task well, though some points may not be fully elaborated Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details Displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning
3	An essay at this level is marked by one of more of the following: Addresses the topic and task using somewhat developed explanations, exemplifications, and/or details Displays unity, progression, and coherence, though connection of ideas may be occasionally obscured May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning May display accurate but limited range of syntactic structures and vocabulary
2	An essay at this level may reveal one or more of the following weaknesses: Limited development in response to the topic and task Inadequate organization or connection of ideas Inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task A noticeable inappropriate choice of words or word forms An accumulation of errors in sentence structure and/or usage
1	An essay at this level is seriously flawed by one or more of the following weaknesses: Serious disorganization or underdevelopment Little or no detail, or irrelevant specifics, or questionable responsiveness to the task Serious and frequent errors in sentence structure or usage
0	An essay at this level merely copies words from the topic, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Appendix C

Precision and Recall Measures for Grammatical Tags in the Final Written Subcorpus

Linguistic feature	Precision	Recall
Attributive adjectives	0.98	0.97
Predicative adjectives	0.99	0.96
Nouns	0.96	.99
Gerunds	0.97	.96
Nominalizations	0.99	n/a
Adverbs	0.98	0.97
THAT—adjective complement clauses	1.00	1.00
THAT—noun complement clauses	0.89	1.00
THAT—verb complement clauses	1.00	0.99
THAT deletion	0.94	0.67
THAT relative clauses	0.98	0.94
TO–infinitive marker	0.96	0.96
Base form of BE–main verb	1.00	1.00
DO–auxiliary verb	0.97	1.00
DO–main verb	1.00	0.92
Base form of HAVE–main verb	0.96	1.00
Other verbs—present tense or nonfinite	0.97	0.95
Past form of BE–main verb	1.00	1.00
Past form of DO–auxiliary verb	1.00	1.00
Past form of DO–main verb	1.00	1.00
Past form of HAVE–auxiliary verb	1.00	1.00
Past form of HAVE–main verb	1.00	1.00
Past-tense verbs—other verbs	0.99	0.94
Third-person present form of BE–main verb	0.94	1.00
Third-person present form of DO–auxiliary verb	1.00	1.00
Third-person present form of DO–main verb	1.00	1.00
Third-person present form of HAVE–main verb	1.00	1.00

Linguistic feature	Precision	Recall
Third-person present tense verbs—other verbs	0.96	0.93
Infinitive verbs in TO–clauses	0.99	0.94
WH relative clauses	1.00	0.98
Finite passive-voice verbs	0.97	0.99
Perfect aspect verbs	1.00	0.92
Finite progressive-aspect verbs	0.97	0.91
Passive nonfinite relative clauses	0.90	0.93
WH questions	1.00	1.00
Modal verbs	1.00	0.99
Prepositions	0.99	0.99
Pronouns	1.00	0.99
Linking adverbials	0.99	0.99
Demonstrative determiners	0.96	0.99
Demonstrative pronouns	0.97	0.94

Appendix D

Precision and Recall Measures for Grammatical Tags in the Final Spoken Subcorpus

Linguistic feature	Precision	Recall
Attributive adjectives	0.98	0.96
Predicative adjectives	0.99	0.84
Nouns	0.97	.98
Nominalizations	0.99	n/a
Adverbs	0.97	0.97
THAT—adjective complement clauses	1.00	1.00
THAT—noun complement clauses	0.80	0.89
THAT—verb complement clauses	0.97	0.92
THAT deletion	0.88	0.88
THAT relative clauses	0.90	0.89
TO–infinitive marker	0.99	0.95
Base form of BE–main verb	1.00	0.98
DO–auxiliary verb	0.98	1.00
DO–main verb	1.00	0.95
Base form of HAVE–main verb	1.00	0.92
Other verbs—present tense or nonfinite	0.97	0.97
Past form of BE–main verb	1.00	0.96
Past form of DO–auxiliary verb	1.00	1.00
Past form of DO–main verb	1.00	1.00
Past form of HAVE–auxiliary verb	1.00	1.00
Past form of HAVE–main verb	0.92	1.00
Past-tense verbs—other verbs	1.00	0.99
Third-person present form of BE–main verb	1.00	0.96
Third-person present form of DO–auxiliary verb	0.95	1.00
Third-person present form of DO–main verb	1.00	0.67
Third-person present form of HAVE–main verb	0.97	0.97
Third-person present tense verbs—other verbs	0.98	0.98

Linguistic feature	Precision	Recall
Infinitive verbs in TO–clauses	1.00	0.96
WH relative clauses	0.98	0.98
Finite passive-voice verbs	0.97	0.96
Perfect-aspect verbs	0.91	1.00
Finite progressive-aspect verbs	0.97	0.90
Passive nonfinite relative clauses	0.88	1.00
WH questions	1.00	1.00
Modal verbs	1.00	0.99
Prepositions	0.98	0.99
Pronouns	1.00	1.00
Possessive nouns	0.97	0.94
Linking adverbials	0.95	0.95

Appendix E

List of Grammatical and Lexico-Grammatical Features Analyzed in the Project

Feature	Examples
1. Pronouns and pro-verbs	First-person pronouns
	Second-person pronouns
	Third-person pronouns (excluding *it*)
	Pronoun *it*
	Demonstrative pronouns (*this*, *that*, *these*, *those* as pronouns)
	Indefinite pronouns (e.g., *anybody*, *nothing*, *someone*)
	Pro-verb *do*
2. Reduced forms and dispreferred structures	Contractions
	Complementizer *that* deletion (e.g., *I think [0] he went*)
	Stranded prepositions (e.g*., the candidate that I was thinking of)*
	Split auxiliaries (e.g., *they were apparently shown to* …)
3. Prepositional phrases	For example, pain *in my leg*, went *to the store*
4. Coordination	Phrasal coordination (NOUN and NOUN; ADJ and ADJ; VERB and VERB; ADV and ADV)
	Independent clause coordination (clause initial *and*)
5. WH-questions	For example, *What's* your name?
6. Lexical specificity	Type/token ratio
	Word length
7. Nouns	Nominalizations (ending in *–tion, -ment, -ness, -ity*)
	Common nouns
7a. Semantic categories of nouns	Animate noun (e.g., *teacher*, *child*, *person*)
	Cognitive noun (e.g., *fact*, *knowledge*, *understanding*)
	Concrete noun (e.g., *rain*, *sediment*, *modem*)
	Technical/concrete noun (e.g., *cell*, *wave*, *electron*)
	Quantity noun (e.g., *date*, *energy*, *minute*)
	Place noun (e.g., *habitat*, *room*, *ocean*)
	Group/institution noun (e.g., *committee*, *bank*, *congress*)
	Abstract/process nouns (e.g., *application*, *meeting*, *balance*)
Feature	Examples
8. Verbs	
8a. Tense and aspect markers	Past tense
	Perfect aspect verbs
	Non-past tense
8b. Passives	Agentless passives
	By passives
8c. Modals	Possibility modals (*can*, *may*, *might*, *could*)
	Necessity modals (*ought*, *must*, *should*)
	Predictive modals (*will*, *would*, *shall*)

Feature	Examples
8d. Semantic categories of verbs	*Be* as main verb
	Activity verb (e.g., *smile*, *bring*, *open*)
	Communication verb (e.g., *suggest*, *declare*, *tell*)
	Mental verb (e.g., *know*, *think*, *believe*)
	Causative verb (e.g., *let*, *assist*, *permit*)
	Occurrence verb (e.g., *increase*, *grow*, *become*)
	Existence verb (e.g., *possess*, *reveal*, *include*)
	Aspectual verb (e.g., *keep*, *begin*, *continue*)
8e. Phrasal verbs	Intransitive activity phrasal verb (e.g., *come on, sit down)*
	Transitive activity phrasal verb (e.g., *carry out*, *set up*)
	Transitive mental phrasal verb (e.g., *find out*, *give up*)
	Transitive communication phrasal verb (e.g., *point out*)
	Intransitive occurrence phrasal verb (e.g., *come off*, *run out*)
	Copular phrasal verb (e.g., *turn out*)
	Aspectual phrasal verb (e.g., *go on*)
9. Adjectives	Attributive adjectives
	Predicative adjectives
9a. Semantic categories of adjectives	Size attributive adjectives (e.g., *big*, *high*, *long*)
	Time attributive adjectives (e.g., *new*, *young*, *old*)
	Color attributive adjectives (e.g., *white*, *red*, *dark*)
	Evaluative attributive adjectives (e.g., *important*, *best*, *simple*)
	Relational attributive adjectives (e.g., *general*, *total*, *various*)
	Topical attributive adjectives (e.g., *political*, *economic*, *physical*)
10. Adverbs and adverbials	Place adverbials
	Time adverbials
10a. Adverb classes	Conjuncts (e.g., *consequently*, *furthermore*, *however*)
	Downtoners (e.g., *barely*, *nearly*, *slightly*)
	Hedges (e.g., *at about*, *something like*, *almost*)
	Amplifiers (e.g., *absolutely*, *extremely*, *perfectly*)
	Emphatics (e.g., *a lot*, *for sure*, *really*)
	Discourse particles (e.g., sentence initial *well*, *now*, *anyway*)
	Other adverbs
10b. Semantic categories of stance adverbs	Nonfactive adverbs (e.g., *frankly*, *mainly*, *truthfully*)
	Attitudinal adverbs (e.g., *surprisingly*, *hopefully*, *wisely*)
	Certainty adverbs (e.g., *undoubtedly*, *obviously*, *certainly*)
	Likelihood adverbs (e.g., *evidently*, *predictably*, *roughly*)
11. Adverbial subordination	Causative adverbial subordinator (*because*)
	Conditional adverbial subordinator (*if*, *unless*)
	Other adverbial subordinator (e.g., since, *while*, *whereas*)

Feature	Examples
12. Nominal postmodifying clauses	*That* relatives (e.g., *the dog that bit me*, *the dog that I saw*) WH relatives on object position (e.g., *the man who Sally likes*) WH relatives on subject position (e.g., *the man who likes popcorn*) WH relatives with fronted preposition (e.g., *the manner in which he was told*) Past participial postnominal (reduced relative) clauses (e.g., *the solution produced by this process*)
13. *That* complement clauses	
13a. *That* clauses controlled by a verb (e.g., *we predict that the water is here*)	Communication verb (e.g., *imply*, *report*, *suggest*) Attitudinal verb (e.g., *anticipate*, *expect*, *prefer*) Certainty verb (e.g., *demonstrate*, *realize*, *show*) Likelihood verb (e.g., *appear*, *hypothesize*, *predict*)
13b. *That* clauses controlled by an adjective (e.g., *it is strange that he went there*)	Attitudinal adjectives (e.g., *good*, *advisable*, *paradoxical*) Likelihood adjectives (e.g., *possible*, *likely*, *unlikely*)
13c. *That* clauses controlled by a noun (e.g., *the view that tax reform is needed is widely accepted*)	Communication noun (e.g., *comment*, *proposal*, *remark*) Attitudinal noun (e.g., *hope*, *reason*, *view*) Certainty noun (e.g., *assertion*, *observation*, *statement*) Likelihood noun (e.g., *assumption*, *implication*, *opinion*)
14. WH-clauses	For example, I don't know *when I'll be able to go*.
15. *To*-clauses	
15a. *To*-clauses controlled by a verb (e.g., *He offered to stay*)	Speech-act verb (e.g., *urge*, *report*, *convince*) Cognition verb (e.g., *believe*, *learn*, *pretend*) Desire/intent/decision verb (e.g., *aim*, *hope*, *prefer*) Modality/cause/effort verb (e.g., *allow*, *leave*, *order*) Probability/simple fact verb (e.g., *appear*, *happen*, *seem*)
15b. *To*-clauses controlled by an adjective	Certainty adjectives (e.g., *prone*, *due*, *apt*) Ability/willingness adjectives (e.g., *competent*, *hesitant*) Personal affect adjectives (e.g., *annoyed*, *nervous*) Ease/difficulty adjectives (e.g., *easy*, *impossible*) Evaluative adjectives (e.g., *convenient*, *smart*)
15c. *To*-clauses controlled by a noun	For example, *agreement*, *authority*, *intention*

Appendix F

Collocations of Five Light Verbs in Speech

Verb	Postcollocate	Normed frequency per 1,000	Number of texts	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
GET	xx allergies/allergic	6	14	0	0	0	0	0	10	6	3
GET	xx better	5	11	0	4	2	0	0	2	4	9
GET	xx information	6	17	0	14	12	8	16	5	4	0
GET	rid	6	14	0	7	7	0	16	8	6	0
GET	up	28	58	55	92	84	136	0	3	1	0
GET[a]	xx back	28	61	0	7	2	16	16	26	47	32
GET[a]	xx cold	6	16	0	0	0	0	0	8	8	16
GET[a]	xx money	14	32	0	0	2	0	16	12	20	29
GET[a]	xx nutrient(s)/nutritions	8	24	0	8	0	0	16	13	9	6
GIVE	xx assignment	6	14	0	0	0	0	0	11	6	9
GIVE	xx money	8	22	0	0	0	0	16	16	10	0
GIVE[a]	xx bed	12	32	0	0	0	0	0	14	18	22
GIVE[a]	xx example(s)	36	105	0	0	2	0	64	66	40	41
GIVE[a]	xx gift(s)	15	42	0	21	19	0	16	18	15	6
HAVE	xx allergy/allergic	7	18	0	0	0	0	64	10	6	3
HAVE	xx chance	5	13	0	4	12	8	0	9	1	0
HAVE	xx class(es)	16	46	0	0	11	8	16	20	20	22
HAVE	xx competition	7	21	0	0	0	0	0	12	10	9
HAVE	xx day	5	13	0	11	14	8	0	1	3	0
HAVE	xx energy	5	11	0	11	14	16	0	1	0	0
HAVE	xx exam	9	25	0	4	0	0	0	15	11	12
HAVE	xx excuse	5	12	0	0	0	0	0	5	6	9
HAVE	fun	5	12	0	21	10	16	16	0	2	0
HAVE	xx opportunity(ies)	9	25	0	0	29	40	0	4	4	6

Verb	Postcollocate	Normed frequency per 1,000	Number of texts	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
HAVE	xx problem(s)	34	96	0	11	16	0	80	59	35	23
HAVE	xx question(s)	7	16	0	0	0	0	0	10	9	6
HAVE	xx reaction(s)	6	19	0	0	0	0	16	11	8	6
HAVE	xx review	8	50	0	0	0	0	0	30	26	9
HAVE	xx skills	6	15	0	0	0	0	0	11	8	3
HAVE	xx ticket	6	17	0	0	0	0	16	14	9	0
HAVE[a]	xx computer(s) (skills)	31	76	0	8	26	16	16	33	38	38
HAVE[a]	xx money	19	51	55	14	12	24	0	23	19	19
HAVE[a]	xx nose	10	29	0	0	0	0	48	14	14	6
HAVE[a]	xx runny/running	11	32	0	0	0	0	32	20	15	6
HAVE[a]	xx time	51	133	164	95	86	88	16	38	35	28
MAKE	xx better	6	17	0	11	12	8	0	5	3	3
MAKE	xx exam(s)/(ination)	47	83	0	0	0	0	0	64	64	95
MAKE	xx home	7	21	0	0	0	0	0	12	9	13
MAKE	xx library	5	15	0	0	0	0	0	8	5	4
MAKE	xx life	7	21	0	32	26	8	0	0	0	0
MAKE	xx noise(s)	9	29	0	0	0	0	16	23	10	3
MAKE	xx test(s)	9	20	0	0	0	0	0	24	8	3
MAKE	xx trunk	5	10	0	0	0	0	0	1	6	12
MAKE[a]	xx decision(s)	7	17	0	0	2	0	0	13	4	16
MAKE[a]	xx exception	12	33	0	0	0	0	16	14	22	16
MAKE[a]	xx hollow	23	68	0	0	0	0	16	28	35	41
MAKE[a]	sure	5	16	0	0	0	8	0	4	10	6
MAKE[a]	xx tree(s)	42	103	0	0	0	0	32	55	62	51
TAKE	care	9	22	0	18	10	32	32	2	6	5
TAKE	xx exam(s)	20	47	0	0	6	0	16	30	30	6
TAKE	xx midterm	6	16	0	0	0	0	32	10	6	3
TAKE	part	5	15	0	0	5	0	0	14	3	3

Verb	Postcollocate	Normed frequency per 1,000	Number of texts	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
TAKE	xx test(s)	8	20	0	4	7	0	64	8	8	6
TAKE[a]	xx class(es)	21	58	0	0	0	0	32	18	41	22
TAKE[a]	computer (class)	7	23	0	0	0	0	0	12	10	13
TAKE[a]	xx course(s)	8	23	0	0	0	0	16	15	8	13
TAKE[a]	xx note(s)	8	43	0	0	0	0	0	26	21	23
TAKE[a]	place	5	10	0	0	0	0	16	3	9	3
TAKE[a]	xx time	6	18	0	11	10	16	0	6	5	0

Note. Ind = independent task; int = integrated task.

[a] Collocation occurs in the prompt; xx indicates that the collocate often occurs separated from the verb.

Appendix G

Collocations of Five Light Verbs in Writing

Verb	Postcollocate	Normed frequency per 1,000	Number of texts	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
GET	along	10	15	0	29	17	3	0	0	0	0
GET	better	8	21	0	3	23	8	5	0	0	0
GET	lost	6	12	0	0	0	0	28	4	23	4
GET	xx grade(s)	5	11	0	2	12	8	0	0	0	0
GET	xx job(s)	39	68	50	100	39	44	0	0	0	0
GIVE	xx example	8	18	0	4	4	6	28	8	8	12
HAVE	xx ability(ies)	33	68	130	43	42	16	15	33	8	8
HAVE	xx (dis)advantages	8	21	50	6	6	0	19	8	4	12
HAVE	xx career	5	14	10	4	6	20	0	0	0	0
HAVE	xx chance(s)	10	24	10	12	21	18	0	0	0	0
HAVE	xx choice(s)	5	10	20	8	4	8	0	0	0	0
HAVE	xx effect(s)	7	19	0	6	4	14	19	4	4	16
HAVE	xx fat	19	49	0	0	0	0	33	83	34	53
HAVE	xx friend(s)	5	13	0	14	8	5	0	0	0	0
HAVE	xx interest(s)	14	28	0	43	19	13	0	0	0	0
HAVE	xx job(s)	19	41	30	43	33	16	0	0	0	0
HAVE	xx knowledge(s)	9	19	30	18	10	13	0	4	0	0
HAVE	xx limitations	7	16	0	0	0	0	14	21	16	19
HAVE	xx money	7	15	10	10	9	18	0	0	0	0
HAVE	xx opinion(s)	5	11	0	8	4	0	0	8	11	0
HAVE	xx opportunity(ies)	8	18	0	16	8	13	10	0	0	0
HAVE	xx problem(s)	11	28	10	14	8	6	29	4	23	4
HAVE	xx skills	9	22	0	14	8	16	19	4	4	0

Verb	Postcollocate	Normed frequency per 1,000	Number of texts	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
HAVE	xx time	14	29	40	20	27	13	0	4	0	0
HAVE[a]	crystals	10	26	0	0	0	0	51	25	19	15
MAKE	xx better	5	13	0	14	8	3	0	0	0	4
MAKE	xx decision(s)	9	20	0	12	22	6	5	0	0	4
MAKE	xx happy	7	15	0	16	8	11	0	0	0	0
MAKE	xx money	8	14	10	19	8	8	0	0	0	0
MAKE	xx possible	5	10	0	2	10	8	0	0	4	0
MAKE	xx sense	5	10	0	2	2	6	5	4	15	0
MAKE[a]	xx point(s)	5	13	0	2	0	3	5	12	12	15
TAKE	care	10	19	50	14	0	13	24	0	4	0
TAKE	xx class(es)	12	21	10	24	10	27	0	0	0	0
TAKE	course(s)	7	14	0	6	21	8	0	0	0	0
TAKE	xx example	6	13	0	6	8	18	0	0	0	4
TAKE	xx subject(s)	25	49	20	39	49	29	0	0	0	0

Note. Ind = independent task; int = integrated task.

[a] Collocation occurs in the prompt; xx indicates that the collocate often occurs separated from the verb.

Appendix H

Lexical Bundles in Spoken Responses, Organized by Discourse Function

Bundle	Number of files	Normed per 100,000	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
				Personal/epistemic bundles					
I think my life	98.0	36.9	160.3	110.9	117.1	0.0	0.0	0.0	0.0
I think the best	60.0	21.4	94.7	51.8	50.2	0.0	6.9	2.0	3.2
Think the best way	50.0	18.0	87.4	49.3	41.8	0.0	4.1	0.0	0.0
I think it is	49.0	17.3	36.4	54.2	25.1	0.0	5.5	7.1	3.2
I think that the	35.0	12.1	14.6	27.1	8.4	16.4	15.1	5.1	6.5
I think it's better	32.0	11.1	29.1	39.4	8.4	0.0	4.1	2.0	3.2
so I think that	30.0	10.7	10.9	22.2	0.0	0.0	8.3	12.2	3.2
and I think that	29.0	10.7	25.5	19.7	8.4	0.0	5.5	10.2	3.2
think it's better to	27.0	9.3	29.1	32.0	16.7	0.0	1.4	2.0	3.2
think it is better	26.0	9.3	25.5	27.1	25.1	0.0	1.4	2.0	3.2
so I think it's	18.0	6.2	21.9	19.7	0.0	0.0	1.4	3.1	0.0
and I think it's	17.0	6.6	18.2	9.9	0.0	0.0	5.5	5.1	0.0
I think that my	17.0	6.6	14.6	27.1	33.4	0.0	0.0	0.0	0.0
think that my life	16.0	5.9	14.6	24.7	25.1	0.0	0.0	0.0	0.0
and I think the	15.0	5.2	10.9	9.9	8.4	0.0	6.9	2.0	0.0
I think the most	15.0	4.8	21.9	14.8	16.7	0.0	0.0	0.0	0.0
know what to do	16.0	5.5	3.6	4.9	0.0	0.0	1.4	11.2	3.2
or something like that	24.0	8.6	10.9	9.9	16.7	0.0	6.9	9.2	6.5
Total			630.1	554.7	401.4	16.4	74.3	73.3	38.8

Bundle	Number of files	Normed per 100,000	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
				Attitudinal/evaluative bundles					
the best way for	72.0	27.6	131.1	71.5	92.0	0.0	4.1	0.0	0.0
better to go to	65.0	23.5	109.3	69.0	66.9	0.0	1.4	1.0	0.0
the best way to	52.0	20.0	83.8	69.0	16.7	0.0	1.4	1.0	0.0
best way to relax	42.0	16.6	61.9	66.6	8.4	0.0	0.0	0.0	0.0
it is better to	41.0	15.9	61.9	44.4	58.5	0.0	2.8	1.0	3.2
best way for a	38.0	13.8	58.3	37.0	66.9	0.0	0.0	0.0	0.0
is better to go	35.0	12.8	51.0	41.9	41.8	0.0	1.4	0.0	0.0
it's better to go	33.0	12.4	65.6	32.0	33.4	0.0	0.0	1.0	0.0
is the best way	32.0	11.7	54.6	27.1	25.1	0.0	5.5	0.0	0.0
the problem is that	47.0	16.6	0.0	0.0	0.0	0.0	13.8	33.6	16.2
problem is that he	34.0	12.4	0.0	0.0	0.0	0.0	24.8	14.2	12.9
and the problem is	16.0	5.5	0.0	0.0	0.0	0.0	5.5	10.2	6.5
problem is that the	15.0	4.8	0.0	0.0	0.0	16.4	5.5	9.2	0.0
is the most important	17.0	6.2	18.2	22.2	8.4	0.0	2.8	1.0	0.0
is very important for	17.0	5.9	3.6	9.9	8.4	0.0	8.3	4.1	3.2
it is very important	15.0	4.8	14.6	12.3	0.0	0.0	1.4	3.1	3.2
should go to the	19.0	6.9	7.3	2.5	0.0	16.4	12.4	6.1	3.2
you have to do	15.0	5.2	7.3	12.3	50.2	0.0	0.0	2.0	0.0
have to go to	15.0	5.9	10.9	14.8	0.0	16.4	8.3	1.0	0.0
if you want to	24.0	9.7	21.9	22.2	8.4	0.0	5.5	7.1	3.2
you don't want to	19.0	8.6	0.0	0.0	8.4	0.0	6.9	16.3	9.7
to be able to	24.0	9.7	14.6	12.3	8.4	0.0	6.9	12.2	3.2
not be able to	21.0	8.3	0.0	7.4	0.0	0.0	12.4	7.1	16.2
Total			775.8	574.4	501.7	49.1	130.6	131.2	80.9

Bundle	Number of files	Normed per 100,000	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
Information source/information organizers/discourse organizers									
Information source									
according to the professor	29.0	10.0	0.0	0.0	0.0	16.4	20.6	6.1	22.7
in the listening passage	16.0	5.5	0.0	0.0	0.0	0.0	12.4	7.1	0.0
in my opinion i	21.0	7.3	18.2	12.3	0.0	0.0	6.9	6.1	0.0
I agree with the	19.0	6.9	32.8	14.8	16.7	0.0	0.0	1.0	3.2
in my opinion the	17.0	5.9	7.3	9.9	16.7	0.0	2.8	5.1	3.2
Total			58.3	37.0	33.4	16.4	42.7	25.4	29.1
Information organizers									
the first one is	34.0	11.7	14.6	4.9	0.0	0.0	15.1	12.2	16.2
the second one is	25.0	8.6	7.3	4.9	0.0	16.4	9.6	10.2	9.7
the second reason is	21.0	7.3	21.9	4.9	8.4	16.4	8.3	4.1	3.2
and the second reason	17.0	5.9	18.2	2.5	8.4	0.0	4.1	5.1	6.5
the first reason is	16.0	5.5	18.2	12.3	8.4	16.4	1.4	2.0	3.2
first reason is that	15.0	4.8	25.5	7.4	8.4	16.4	0.0	1.0	3.2
the second solution is	15.0	5.2	0.0	0.0	0.0	0.0	5.5	9.2	6.5
there are two ways	15.0	5.2	0.0	2.5	0.0	0.0	4.1	10.2	3.2
Total			105.7	39.4	33.4	65.5	48.1	53.9	51.8
Discourse organizers									
at the same time	61.0	23.5	3.6	27.1	16.7	0.0	15.1	37.6	19.4
on the other hand	44.0	15.5	0.0	9.9	25.1	32.7	16.5	17.3	22.7
for example if you	26.0	9.7	3.6	7.4	16.7	0.0	8.3	9.2	19.4
and at the same	18.0	6.9	0.0	7.4	8.4	0.0	2.8	13.2	3.2
first of all the	19.0	6.6	0.0	0.0	8.4	0.0	5.5	10.2	12.9
because first of all	16.0	5.5	7.3	14.8	16.7	0.0	0.0	4.1	6.5
Total			14.6	66.6	92	32.7	48.1	91.6	84.1

Note. Ind = independent task; int = integrated task.

Appendix I

Lexical Bundles in Written Responses, Organized by Discourse Function

Bundle	Number of files	Normed per 100,000	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
Personal/epistemic bundles										
think that it is	22.0	9.6	10.1	19.6	9.5	15.3	4.7	0.0	3.9	0.0
I think it is	19.0	9.2	10.1	11.8	17.2	15.3	0.0	0.0	3.9	0.0
I think that it	16.0	6.8	10.1	11.8	9.5	12.8	0.0	0.0	0.0	0.0
it is true that	19.0	7.6	0.0	5.9	0.0	7.7	4.7	4.2	15.5	26.8
to the fact that	16.0	6.4	10.1	2.0	1.9	5.1	9.4	8.5	11.6	15.3
the fact that the	15.0	6.4	10.1	0.0	3.8	2.6	14.1	4.2	19.3	11.5
a matter of fact	13.0	5.6	0.0	7.8	3.8	0.0	9.4	4.2	7.7	11.5
Total			50.6	58.8	45.8	58.7	42.4	21.2	61.9	65.0
Attitudinal/evaluative bundles										
important than it was	93.0	46.5	141.7	82.3	80.2	45.9	0.0	0.0	0.0	0.0
more important than it	86.0	42.9	121.5	76.4	80.2	35.7	0.0	0.0	0.0	0.0
is more important to	82.0	38.5	50.6	80.3	53.5	48.5	4.7	0.0	3.9	3.8
is far more important	77.0	37.3	111.3	74.4	45.8	45.9	4.7	4.2	0.0	0.0
it is more important	76.0	36.1	40.5	70.5	51.6	53.6	4.7	0.0	0.0	3.8
more important to choose	70.0	32.5	40.5	68.6	53.5	35.7	0.0	0.0	0.0	0.0
important to choose to	68.0	32.1	81.0	72.5	42.0	33.2	0.0	0.0	0.0	0.0
far more important than	65.0	30.5	91.1	64.6	42.0	28.1	0.0	4.2	0.0	0.0
it is important to	45.0	21.7	70.9	37.2	24.8	30.6	0.0	0.0	3.9	7.6
is more important than	44.0	18.0	40.5	25.5	36.3	20.4	0.0	4.2	0.0	0.0
it is very important	25.0	12.8	0.0	23.5	19.1	25.5	0.0	0.0	0.0	0.0
others is more important	27.0	12.0	0.0	21.5	24.8	15.3	0.0	0.0	0.0	0.0
is very important to	21.0	9.6	20.2	17.6	13.4	15.3	0.0	0.0	0.0	0.0
is important to choose	20.0	9.6	60.7	19.6	5.7	12.8	0.0	0.0	0.0	0.0

Bundle	Number of files	Normed per 100,000	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
is much more important	19.0	8.0	10.1	5.9	21.0	7.7	0.0	0.0	3.9	3.8
the most important thing	15.0	6.0	10.1	13.7	5.7	5.1	0.0	0.0	3.9	3.8
that it is important	14.0	6.0	30.4	13.7	5.7	2.6	4.7	0.0	0.0	0.0
more important than in	13.0	5.6	0.0	11.8	9.5	7.7	0.0	0.0	0.0	0.0
important than in the	12.0	5.6	0.0	11.8	9.5	7.7	0.0	0.0	0.0	0.0
much more important than	13.0	5.2	10.1	5.9	13.4	0.0	0.0	0.0	3.9	3.8
of the most important	12.0	4.8	10.1	5.9	9.5	5.1	4.7	0.0	0.0	0.0
interested in than to	40.0	18.0	0.0	50.9	24.8	15.3	0.0	0.0	0.0	0.0
we are interested in	25.0	17.6	20.2	58.8	13.4	12.8	0.0	0.0	0.0	0.0
are interested in than	31.0	13.6	0.0	35.3	22.9	10.2	0.0	0.0	0.0	0.0
they are interested in	27.0	12.8	10.1	23.5	26.7	12.8	0.0	0.0	0.0	0.0
I am interested in	38.0	24.1	30.4	70.5	30.6	12.8	0.0	0.0	0.0	0.0
subjects I am interested	15.0	9.2	0.0	33.3	11.5	0.0	0.0	0.0	0.0	0.0
subjects we are interested	12.0	8.8	0.0	31.3	5.7	7.7	0.0	0.0	0.0	0.0
that I am interested	16.0	7.2	10.1	21.5	5.7	7.7	0.0	0.0	0.0	0.0
one is interested in	15.0	7.2	0.0	9.8	7.6	23.0	0.0	0.0	0.0	0.0
agree with the statement	55.0	24.5	60.7	45.0	34.4	33.2	0.0	0.0	3.9	0.0
i agree with the	48.0	20.8	30.4	54.8	28.6	15.3	0.0	0.0	0.0	0.0
agree with this statement	16.0	6.8	20.2	13.7	9.5	7.7	0.0	0.0	0.0	0.0
I agree with this	14.0	6.0	20.2	17.6	3.8	5.1	0.0	0.0	0.0	0.0
agree that it is	13.0	5.2	0.0	9.8	13.4	2.6	0.0	0.0	0.0	0.0
I agree that it	12.0	4.8	0.0	9.8	11.5	2.6	0.0	0.0	0.0	0.0
i disagree with the	12.0	5.2	20.2	7.8	7.6	7.7	0.0	0.0	0.0	0.0
disagree with the statement	10.0	4.8	0.0	7.8	9.5	7.7	0.0	0.0	0.0	0.0
i would like to	24.0	11.2	0.0	21.5	26.7	2.6	0.0	4.2	0.0	3.8
is the best way	12.0	4.8	30.4	7.8	1.9	5.1	4.7	0.0	3.9	0.0
this theory is limited	9.0	4.8	0.0	0.0	0.0	0.0	9.4	8.5	23.2	7.6
Total			1194.3	1333.8	943.3	673.2	37.7	25.4	50.3	38.2

Bundle	Number of files	Normed per 100,000	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
Information source/information organizers/discourse organizers										
Information source										
in the reading passage	80.0	42.1	0.0	0.0	0.0	0.0	113.0	148.1	77.3	99.4
according to the reading	21.0	10.4	0.0	0.0	0.0	0.0	18.8	42.3	19.3	26.8
according to the professor	19.0	10.4	0.0	0.0	0.0	0.0	9.4	42.3	15.5	38.2
made in the reading	17.0	8.8	0.0	0.0	0.0	0.0	9.4	21.2	7.7	49.7
the professor says that	17.0	8.8	0.0	0.0	0.0	0.0	14.1	46.6	7.7	22.9
it is said that	16.0	6.8	10.1	0.0	1.9	0.0	18.8	16.9	11.6	15.3
stated in the reading	12.0	6.4	0.0	0.0	0.0	0.0	9.4	21.2	15.5	19.1
my point of view	14.0	6.0	10.1	13.7	11.5	2.6	0.0	0.0	0.0	0.0
the professor said that	13.0	6.0	0.0	0.0	0.0	0.0	9.4	33.9	15.5	3.8
the professor argues that	12.0	5.6	0.0	0.0	0.0	0.0	0.0	25.4	19.3	11.5
the reading passage says	11.0	5.6	0.0	0.0	0.0	0.0	18.8	12.7	19.3	7.6
the lecture says that	10.0	5.6	0.0	0.0	0.0	0.0	14.1	8.5	27.1	7.6
the speaker says that	10.0	5.6	0.0	0.0	0.0	0.0	4.7	29.6	11.6	11.5
according to this theory	12.0	4.8	0.0	0.0	0.0	0.0	33.0	4.2	3.9	11.5
according to the passage	11.0	4.8	0.0	0.0	0.0	0.0	14.1	21.2	11.6	3.8
as far as I	11.0	4.8	0.0	5.9	15.3	2.6	0.0	0.0	0.0	0.0
Total			20.2	19.6	28.6	5.1	287.2	474.0	263	328.8
Information organizers										
the second theory is	31.0	12.8	0.0	0.0	0.0	0.0	28.2	29.6	42.5	30.6
points made in the	22.0	10.0	0.0	0.0	0.0	0.0	23.5	8.5	23.2	45.9
is one of the	22.0	8.8	10.1	9.8	21.0	7.7	4.7	0.0	3.9	0.0
theory suggests that birds	22.0	8.8	0.0	0.0	0.0	0.0	9.4	21.2	23.2	34.4
in this set of	21.0	8.4	0.0	2.0	0.0	0.0	42.4	25.4	7.7	11.5
there are three theories	20.0	8.0	0.0	0.0	0.0	0.0	23.5	29.6	27.1	3.8
the points made in	19.0	8.0	0.0	0.0	0.0	0.0	18.8	12.7	15.5	34.4
statement that the ability	18.0	7.6	20.2	11.8	17.2	5.1	0.0	0.0	0.0	0.0

Bundle	Number of files	Normed per 100,000	Ind 1	Ind 2	Ind 3	Ind 4	Int 1	Int 2	Int 3	Int 4
the statement that the	18.0	7.6	20.2	9.8	17.2	5.1	0.0	0.0	3.9	0.0
this theory is not	15.0	6.8	0.0	0.0	0.0	0.0	0.0	25.4	23.2	19.1
the first theory is	16.0	6.4	0.0	0.0	0.0	0.0	9.4	16.9	23.2	15.3
first theory suggests that	16.0	6.4	0.0	0.0	0.0	0.0	9.4	16.9	15.5	22.9
theory is that the	14.0	6.0	0.0	0.0	0.0	0.0	14.1	16.9	23.2	7.6
second theory is that	13.0	5.2	0.0	0.0	0.0	0.0	23.5	12.7	15.5	3.8
second theory states that	12.0	4.8	0.0	0.0	0.0	0.0	0.0	16.9	19.3	11.5
the first theory suggests	12.0	4.8	0.0	0.0	0.0	0.0	4.7	8.5	11.6	22.9
Total			50.6	33.3	55.4	17.9	211.8	241.2	278.4	263.8
Discourse organizers										
on the other hand	90.0	36.5	20.2	25.5	32.5	30.6	61.2	76.2	34.8	26.8
at the same time	24.0	10.4	0.0	13.7	19.1	10.2	4.7	4.2	7.7	3.8
as a result of	13.0	6.0	0.0	9.8	5.7	5.1	4.7	4.2	3.9	7.6
for the following reasons	13.0	5.2	10.2	11.8	7.6	5.1	0.0	0.0	0.0	0.0
Total			30.4	60.7	64.9	51.0	70.6	84.6	46.4	38.2

Note. Ind = independent task; int = integrated task.

Appendix J

Descriptive Statistics for 36 Major Grammatical Features

Table J1

Word Length, Present Tense, Past Tense, and Perfect Aspect

Category	*N*	Word length		Present tense		Past tense		Perfect aspect	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	3.8	0.2	124.9	33.9	14.1	28.8	4.9	8.3
SP-ind-3	142	3.9	0.2	103.2	36.8	27.7	33.1	5.0	7.8
SP-ind-4	67	3.9	0.2	98.7	38.6	30.6	41.4	5.5	8.1
SP-int-2	313	4.1	0.3	122.6	29.2	13.0	15.4	2.5	5.6
SP-int-3	654	4.2	0.3	117.8	29.4	14.9	19.5	2.2	5.0
SP-int-4	216	4.3	0.3	116.4	30.6	15.3	19.9	4.0	7.8
WR-ind-1	42	4.3	0.3	124.7	26.2	10.8	11.3	2.2	4.5
WR-ind-2	177	4.3	0.2	112.1	22.7	14.8	13.9	2.4	4.0
WR-ind-3	155	4.4	0.2	103.8	23.3	15.9	14.4	3.6	4.6
WR-ind-4	102	4.5	0.2	100.5	20.3	12.6	10.1	4.9	5.97
WR-int-1	119	4.6	0.2	111.6	22.1	9.1	11.2	1.9	4.2
WR-int-2	118	4.6	0.2	113.0	24.2	8.9	11.1	2.5	4.2
WR-int-3	122	4.6	0.2	110.0	21.0	8.3	10.6	2.5	4.6
WR-int-4	112	4.7	0.2	105.1	21.1	5.4	7.4	3.8	4.4

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J2

Progressive Aspect, Passive Voice, Main Verb BE, and Phrasal Verb

Category	*N*	Progressive aspect		Passive voice		Main verb BE		Phrasal verb	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	6.2	9.0	4.3	7.3	2.9	7.7	2.2	5.2
SP-ind-3	142	9.6	14.4	3.6	7.9	3.0	6.8	1.9	4.8
SP-ind-4	67	9.9	12.6	3.3	5.5	1.9	5.2	2.0	4.9
SP-int-2	313	8.5	10.9	5.1	8.1	3.8	7.0	1.4	4.9
SP-int-3	654	10.4	11.8	7.2	9.7	4.2	6.8	1.4	4.4
SP-int-4	216	10.6	10.4	10.6	10.5	3.8	6.0	1.7	4.8
WR-ind-1	42	11.7	9.6	3.8	5.0	3.1	4.1	0.3	1.2
WR-ind-2	177	12.1	8.8	5.3	5.4	3.6	4.8	0.6	1.5
WR-ind-3	155	12.3	7.6	8.2	6.7	3.1	3.6	1.1	2.4
WR-ind-4	102	14.7	10.1	8.9	6.1	4.2	4.4	0.8	2.2
WR-int-1	119	9.4	9.5	12.2	10.6	2.4	5.0	0.9	2.46
WR-int-2	118	11.1	9.6	16.1	13.6	1.6	3.5	1.0	2.5
WR-int-3	122	12.1	10.3	20.4	13.4	3.4	5.0	1.1	2.4
WR-int-4	112	14.0	9.3	23.9	11.2	2.6	3.5	1.3	2.9

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J3

Possibility Modal, Prediction Modal, Clausal AND, and Adverb

Category	*N*	Possibility modal		Prediction modal		Clausal AND		Adverb	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	17.7	13.8	7.4	13.0	9.5	9.6	54.4	32.2
SP-ind-3	142	11.7	13.1	6.5	10.5	9.3	9.2	52.3	32.3
SP-ind-4	67	9.3	11.9	5.1	11.7	7.8	10.2	55.8	28.3
SP-int-2	313	12.9	14.1	11.8	13.9	11.8	11.2	26.7	17.4
SP-int-3	654	11.4	12.6	12.1	13.6	11.1	9.8	31.0	17.5
SP-int-4	216	8.8	9.8	11.6	12.6	9.5	9.5	35.9	18.9
WR-ind-1	42	15.0	10.6	9.6	8.9	8.6	7.8	40.3	19.6
WR-ind-2	177	12.4	8.9	10.6	10.8	8.8	8.2	41.5	16.7
WR-ind-3	155	11.2	8.7	9.2	7.8	7.2	6.5	45.2	15.7
WR-ind-4	102	9.3	7.0	10.7	8.9	8.2	6.2	46.4	15.2
WR-int-1	119	14.9	11.6	5.0	7.4	9.6	9.0	36.0	16.1
WR-int-2	118	13.4	11.5	2.8	5.2	9.8	9.0	39.8	15.1
WR-int-3	122	14.6	10.9	3.1	5.6	8.2	7.4	39.1	14.8
WR-int-4	112	9.9	7.6	4.4	5.4	6.6	5.5	44.3	13.8

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J4

Split Auxiliaries, Stance Adverbial, First-Person Pronoun and Second-Person Pronoun

Category	*N*	Split auxiliaries		Stance adverbial		First-person pronoun		Second-person pronoun	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	2.7	5.2	4.7	9.1	73.4	52.3	24.6	35.3
SP-ind-3	142	3.4	6.2	10.5	11.9	74.6	49.8	20.2	31.1
SP-ind-4	67	3.1	6.0	9.5	12.3	68.8	45.6	32.1	35.3
SP-int-2	313	1.4	4.4	4.2	7.8	11.7	18.9	14.4	28.8
SP-int-3	654	2.2	4.5	6.4	10.1	10.0	17.8	16.4	30.7
SP-int-4	216	3.0	5.5	6.8	8.7	7.5	13.5	13.9	25.7
WR-ind-1	42	1.3	2.6	3.1	4.0	47.3	36.6	21.2	33.9
WR-ind-2	177	1.6	2.4	5.3	5.5	47.5	32.6	18.2	29.2
WR-ind-3	155	2.9	3.0	5.2	5.0	32.4	25.4	16.1	22.2
WR-ind-4	102	3.8	3.1	5.2	5.4	27.0	22.7	16.4	24.2
WR-int-1	119	1.8	3.4	3.6	4.9	4.8	8.3	0.8	4.3
WR-int-2	118	3.4	4.2	5.5	6.0	3.0	6.2	0.8	3.1
WR-int-3	122	4.2	4.7	5.1	5.4	3.1	5.8	0.6	2.8
WR-int-4	112	5.5	4.7	5.5	5.2	2.1	4.6	1.1	3.9

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J5

Third-Person Pronoun, Linking Adverbial, Noun, and Nominalization

Category	*N*	Third-person pronoun		Linking adverbial		Noun		Nominalization	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	21.5	31.2	16.5	16.0	155.7	29.2	4.3	7.7
SP-ind-3	142	20.2	25.5	17.7	14.1	167.2	37.6	4.5	9.0
SP-ind-4	67	21.7	30.7	18.1	14.0	167.2	33.5	2.8	6.7
SP-int-2	313	58.0	44.6	17.9	13.7	213.3	41.6	2.6	9.1
SP-int-3	654	53.2	40.1	17.8	13.2	204.4	42.5	3.0	9.7
SP-int-4	216	44.2	32.8	17.0	12.8	210.6	39.5	2.5	8.4
WR-ind-1	42	21.7	19.7	15.4	9.5	204.4	40.1	28.0	18.9
WR-ind-2	177	25.7	24.2	15.2	8.8	202.5	38.4	30.1	20.0
WR-ind-3	155	24.3	19.8	13.8	7.7	211.3	36.0	34.8	19.9
WR-ind-4	102	19.8	16.1	12.0	6.3	215.3	39.1	37.6	21.5
WR-int-1	119	22.8	18.0	15.3	10.3	285.9	39.2	40.1	20.4
WR-int-2	118	25.4	22.8	17.7	9.5	274.0	39.2	38.2	17.5
WR-int-3	122	22.0	18.5	14.9	8.1	274.5	39.5	36.5	15.5
WR-int-4	112	23.4	16.7	17.7	8.5	269.7	36.3	38.7	17.5

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J6

Prepositional Phrase, OF Genitive Phrase, Attributive Adjective, and Premodifying Noun

Category	*N*	Prepositional phrase		OF genitive phrase		Attributive adjective		Premodifying noun	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	81.3	24.8	4.9	6.9	19.4	15.8	7.6	11.3
SP-ind-3	142	86.5	28.3	10.5	9.8	24.7	18.2	9.3	11.1
SP-ind-4	67	85.2	25.4	9.8	9.6	23.5	17.0	9.0	10.3
SP-int-2	313	75.6	24.7	9.2	10.9	25.0	20.4	20.6	16.1
SP-int-3	654	78.3	25.9	10.2	11.3	28.1	20.5	18.5	15.6
SP-int-4	216	83.1	25.4	10.9	10.5	32.3	21.6	19.0	15.2
WR-ind-1	42	97.5	24.3	8.0	8.7	24.7	12.8	13.9	12.3
WR-ind-2	177	99.3	21.8	10.6	8.2	27.7	15.9	12.5	10.1
WR-ind-3	155	107.7	22.1	14.1	9.1	32.2	14.8	13.1	8.7
WR-ind-4	102	106.1	20.5	15.6	8.8	39.8	17.7	15.2	7.6
WR-int-1	119	105.4	25.5	20.2	12.7	45.0	18.8	38.2	24.6
WR-int-2	118	99.7	22.7	18.9	12.9	37.6	17.4	32.9	23.5
WR-int-3	122	101.1	21.8	18.4	12.1	44.3	17.1	35.3	21.2
WR-int-4	112	105.6	24.6	21.7	13.9	50.1	17.4	33.1	19.9

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J7

Finite Adverbial Clause, WH Clause, Verb + THAT-Clause, and Adjective + THAT-Clause

Category	*N*	Finite adverbial clause		WH clause		Verb + THAT-clause		Adjective + THAT-clause	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	16.1	13.6	0.7	2.3	5.6	7.0	1.1	3.0
SP-ind-3	142	13.0	11.3	0.7	2.8	4.9	7.4	0.7	2.9
SP-ind-4	67	13.4	11.3	0.8	2.4	3.4	4.6	0.9	2.7
SP-int-2	313	14.0	12.3	0.9	2.8	8.5	9.4	0.2	1.4
SP-int-3	654	13.6	12.2	1.4	4.0	9.7	10.0	0.4	1.9
SP-int-4	216	12.1	10.6	0.8	2.7	9.1	8.5	0.7	2.2
WR-ind-1	42	13.0	7.5	1.0	2.2	5.7	6.5	0.8	2.1
WR-ind-2	177	12.5	9.2	1.1	2.1	5.8	5.3	1.1	2.0
WR-ind-3	155	10.9	7.3	0.8	1.9	6.2	5.1	1.1	2.0
WR-ind-4	102	10.3	7.3	1.0	1.8	6.1	5.1	1.1	1.6
WR-int-1	119	7.5	8.1	1.1	3.1	14.5	12.4	0.4	1.6
WR-int-2	118	8.6	7.8	2.2	4.5	18.2	11.0	0.8	1.9
WR-int-3	122	7.8	7.2	2.1	3.7	18.3	11.4	0.8	2.4
WR-int-4	112	6.9	6.2	1.9	3.5	17.2	10.0	0.7	1.7

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J8

Noun + THAT-Clause, Verb + TO-Clause, Desire Verb + TO-Clause, and Adjective + TO-Clause

Category	*N*	Noun + THAT-clause		Verb + TO-clause		Desire Verb + TO-clause		Adjective + TO-clause	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	0	0	3.1	5.2	6.2	10.2	3.7	6.2
SP-ind-3	142	0.4	2.2	2.2	4.5	4.1	6.8	3.6	7.4
SP-ind-4	67	1.3	3.4	3.2	7.3	2.3	4.7	4.1	6.4
SP-int-2	313	0.5	2.1	5.1	8.3	5.2	8.5	1.5	3.9
SP-int-3	654	0.6	2.6	3.9	6.5	3.8	6.7	1.5	3.9
SP-int-4	216	1.1	3.1	3.9	6.2	3.3	6.0	1.5	3.5
WR-ind-1	42	1.2	3.0	3.4	6.0	7.3	8.6	4.6	4.8
WR-ind-2	177	1.3	2.2	4.9	5.7	8.2	7.9	4.3	4.6
WR-ind-3	155	1.3	2.1	4.1	4.0	5.8	6.0	3.9	3.9
WR-ind-4	102	1.1	1.8	3.3	3.6	5.1	4.6	4.8	4.6
WR-int-1	119	1.5	3.3	1.7	3.5	1.3	3.3	1.4	3.6
WR-int-2	118	2.6	4.0	1.6	3.2	0.8	2.0	2.2	3.7
WR-int-3	122	3.1	4.6	1.4	2.8	0.8	1.9	2.0	4.3
WR-int-4	112	2.8	4.0	1.3	3.1	1.0	2.7	2.6	4.1

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Table J9

Noun + TO-Clause, Verb + ING-Clause, Finite Relative Clause, and ED Nonfinite Relative

Category	*N*	Noun + TO-clause		Verb + ING-clause		Finite relative clause		ED nonfinite relative	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP-ind-2	39	0.4	1.8	0.7	2.4	7.8	11.4	0.0	0.0
SP-ind-3	142	0.6	2.7	0.9	2.8	7.8	11.3	0.4	2.1
SP-ind-4	67	1.0	3.3	0.4	1.8	7.4	9.2	0.1	1.1
SP-int-2	313	0.9	3.3	0.5	2.0	8.8	10.0	0.7	3.0
SP-int-3	654	0.6	2.3	0.7	2.6	10.6	11.4	1.1	3.4
SP-int-4	216	1.1	3.0	0.7	2.3	12.4	9.9	1.8	4.0
WR-ind-1	42	4.5	6.7	0.5	1.7	10.4	10.5	0.8	2.1
WR-ind-2	177	3.7	4.8	0.5	1.4	10.3	8.1	0.4	1.3
WR-ind-3	155	4.1	4.5	0.7	1.7	9.1	6.8	1.5	2.4
WR-ind-4	102	3.1	4.2	0.9	1.6	10.2	7.9	1.8	2.6
WR-int-1	119	0.6	2.0	0.2	1.2	10.5	8.8	2.7	4.3
WR-int-2	118	0.6	1.9	0.2	1.2	9.5	8.2	3.3	4.8
WR-int-3	122	0.2	1.1	0.3	1.2	10.3	7.9	5.1	6.8
WR-int-4	112	0.7	2.0	0.3	1.2	10.3	8.6	6.7	6.5

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task. 1–4 in Category column indicates score level.

Appendix K

Results of the Factor Analysis

Table K1

Rotated Factor Pattern (Standardized Regression Coefficients)

Feature	Factor 1	Factor 2	Factor 3	Factor 4
Word length	0.399	0.199	0.489	-0.091
THAT deletion	-0.484	0.188	0.023	0.017
Present tense	-0.328	-0.038	-0.113	-0.702
Second-person pronoun	-0.100	-0.394	-0.093	-0.231
First-person pronoun	-0.202	-0.333	0.070	0.351
Noun	0.638	0.371	0.083	-0.032
Preposition	0.515	-0.176	0.280	0.073
Past tense	-0.121	0.092	-0.069	0.742
Third-person pronoun	-0.549	0.408	-0.112	0.023
Modal verb	-0.365	0.007	-0.056	-0.238
Finite passive verb	0.405	0.171	0.023	-0.057
Speech verb + *that*-clause	-0.132	0.676	0.076	0.025
Likelihood verb + *that*-clause	-0.451	0.121	0.130	-0.033
Mental noun	-0.075	0.147	0.510	-0.044
Abstract noun	0.090	-0.366	0.376	-0.014
Concrete noun	0.645	0.084	-0.548	-0.005
Place noun	0.098	0.448	0.179	-0.056
Topical adjective	0.399	-0.015	-0.009	-0.020
Activity verb	0.029	-0.162	-0.473	-0.048
Communication verb	-0.241	0.805	0.041	0.056
Mental verb	-0.616	-0.136	0.312	-0.058
Attributive adjective	0.614	-0.070	0.080	-0.030
Premodifying noun	0.389	0.392	-0.124	-0.015
Finite adverbial clause	-0.315	-0.057	-0.154	-0.066
Noun + *of*-phrase	0.472	-0.012	0.145	0.053
Passive *-ed* relative clause	0.321	0.140	0.014	-0.038
Noun + *to* complement clause	-0.073	-0.095	0.331	0.119
Nominalization	0.295	0.117	0.619	-0.031

Table K2

Eigenvalues for the First Four Factors in the Solution

Factor	Eigenvalue	Difference	Proportion	Cumulative
1	5.579	2.899	0.199	0.199
2	2.680	0.392	0.096	0.295
3	2.288	0.506	0.082	0.377
4	1.783	0.321	0.064	0.440

Table K3

Interfactor Correlations

Factor	Factor 1	Factor 2	Factor 3	Factor 4
1	1.000	0.257	0.285	-0.058
2	0.257	1.000	0.083	-0.084
3	0.285	0.083	1.000	0.017
4	-0.058	-0.084	0.017	1.000

Appendix L

Mean Dimension Scores for Each of the Text Categories in the TOEFL iBT Corpus

Mode	Task type	Score	*N*	Factor 1		Factor 2		Factor 3		Factor 4	
				Mean	SD	Mean	SD	Mean	SD	Mean	SD
SP	ind	2	39	-7.87	4.35	-6.22	2.32	-2.21	3.02	1.21	3.42
SP	ind	3	142	-4.24	5.55	-5.52	2.30	-2.32	2.80	2.70	3.73
SP	ind	4	67	-4.00	5.77	-6.09	2.00	-1.91	2.74	2.82	4.23
SP	int	2	313	-4.11	6.94	1.34	4.17	-2.29	3.21	-0.66	1.57
SP	int	3	654	-3.03	7.48	0.61	3.93	-2.07	3.18	-0.46	1.82
SP	int	4	216	-0.31	7.62	0.42	3.41	-1.61	3.20	-0.47	1.86
WR	ind	1	42	-3.29	5.75	-4.98	2.68	5.23	3.08	0.24	1.63
WR	ind	2	177	-2.30	5.96	-4.44	3.00	4.76	2.69	0.89	1.55
WR	ind	3	155	0.93	6.12	-3.73	2.42	4.97	2.61	0.77	1.52
WR	ind	4	102	2.80	6.73	-3.31	2.35	4.54	2.32	0.55	1.28
WR	int	1	119	9.42	7.28	5.23	3.74	1.18	3.79	-0.70	1.11
WR	int	2	118	8.00	8.06	5.46	3.34	1.59	3.45	-0.81	1.20
WR	int	3	122	9.93	8.76	5.21	3.37	1.36	3.56	-0.74	1.07
WR	int	4	112	12.05	8.36	4.80	2.92	1.52	3.60	-0.75	0.93

Note. SP = spoken mode; WR = written mode; ind = independent task; int = integrated task.

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

*America Samoa, Guam, Puerto Rico, and US Virgin Islands