ETS TOEFL

Quality Beyond Measu

Check for updates

TOEFL® Research Report
TOEFL–RR-90
ETS Research Report No. RR–19-45

The Longitudinal Development of Grammatical Complexity at the Phrasal and Clausal Levels in Spoken and Written Responses to the *TOEFL iBT*® Test

Bethany Gray

Joe Geluso

Phuong Nguyen

December 2019

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

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

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Longitudinal Development of Grammatical Complexity at the Phrasal and Clausal Levels in Spoken and Written Responses to the *TOEFL iBT*® Test

Bethany Gray, Joe Geluso, & Phuong Nguyen

Iowa State University, Ames, IA

In the present study, we take a longitudinal, corpus-based perspective to investigate short-term (over 9 months) linguistic change in the language produced for the spoken and written sections of the *TOEFL iBT*® test by a group of English-as-a-foreign-language (EFL) learners in China. The goal of the study is to identify patterns that characterize the trajectory that language learners move through in terms of their use of phrasal and clausal grammatical complexity, as mediated by mode (spoken and written) and task type (independent and integrated). Results of a multidimensional analysis reveal that in many cases, learners developed in expected ways: discourse styles at Time 1 were not always aligned with mode- and task type-specific discourse patterns but developed over time, with discourse styles at Time 2 better approximating expected norms and exhibiting increased task differentiation. These changes were particularly noteworthy for Dimension 1, which is related to phrasal and clausal complexity. Results of a developmental complexity analysis revealed more mixed results, seeming to indicate that the relatively low-proficiency learners represented by the longitudinal corpus may just be beginning on the hypothesized paths of development. The most important developments occurred for independent writing, in which students exhibited increases in the frequency of phrasal features, as well as functional expansion in the use of a range of complexity features.

Keywords task variation; spoken and written variation; multidimensional analysis; complexity analysis; language development

doi:10.1002/ets2.12280

An underlying assumption of much second language development research is that more proficient learners produce more linguistically complex language. Thus, measures of complexity are often used as an index of language development (Bulté & Housen, 2014; Crossley & McNamara, 2014), and a great deal of research has supported this assumption empirically (e.g., Parkinson & Musgrave, 2014; Taguchi, Crawford, & Wetzel, 2013). Although cross-sectional research is common, often due to practicality reasons (Ortega & Byrnes, 2008; Ortega & Iberri-Shea, 2005), longitudinal research designs are needed to explore how learners' language evolves over time. Less is known about the paths and trajectories that learners follow as they move toward increased proficiency. To further complicate the matter, recent comparisons of cross-sectional and longitudinal research designs on related data sets have resulted in contradictory results (e.g., Bestgen & Granger, 2014; Connor-Linton & Polio, 2014).

Knowledge of learner progressions can contribute to a range of goals in applied linguistics, from characterizing acquisition to informing and evaluating educational practices (Ortega & Byrnes, 2008; Yoon, 2018). In language assessment, longitudinal descriptions of learner production can inform the design of language tests and rubrics to fulfill Alderson's (2000) call for assessments to capture gains made by students in English for Academic Purposes (EAP) courses as well as help address the particular challenge of capturing incremental changes in learner language over shorter amounts of time, as is often the case with EAP courses.

In the present study, we took a longitudinal, corpus-based perspective to investigate short-term linguistic change (9 months) in spoken and written responses on the *TOEFL iBT*® test for English-as-a-foreign-language (EFL) learners in China. The goal of the study was to characterize the trajectory that language learners follow in their use of phrasal and clausal grammatical complexity, as mediated by mode (spoken and written) and task type (independent and integrated).

Longitudinal descriptions of learner data also have the potential to inform language assessment validation efforts. Building a validity argument for assessments such as the TOEFL iBT is a multistep process, as illustrated in Chapelle,

Corresponding author: B. Gray, E-mail: begray@iastate.edu

Enright, and Jamieson's (2008) interpretative framework. While the primary goal of this study was to provide detailed linguistic descriptions of learner development as captured by the TOEFL iBT, the results have also been interpreted in light of Chapelle et al.'s (2008) *explanation inference* and the warrant that "expected scores are attributed to a construct of academic language proficiency" (p. 20). This study addressed two assumptions underlying this warrant: (a) that linguistic knowledge (as evidenced by learner language production) varies in expected ways, and (b) that performance varies relative to time and experience spent learning English (Chapelle et al., 2008).

Corpus-based linguistic descriptions are increasingly being used for language assessment validation efforts (e.g., LaFlair & Staples, 2017; see also Xi, 2017). Because learners are expected to produce different types of discourse in spoken versus written and integrated versus independent tasks (Enright & Tyson, 2008), linguistic descriptions of TOEFL responses can provide backing for the first assumption by documenting language variation along these parameters (e.g., Biber & Gray, 2013; Cumming et al., 2005). The present research report extended these studies by also addressing the assumption that test takers' language develops in expected ways relative to time spent learning English, with a focus on grammatical complexity and core discourse characteristics of TOEFL responses.

Background: The Longitudinal Development of Grammatical Complexity

Major Approaches

Longitudinal research on the development of complexity has been carried out using case studies of a few learners (Ferrari, 2012; Gunnarsson, 2012), dynamic systems theory (Polat & Kim, 2014; Spoelman & Verspoor, 2010; Vercellotti, 2017), computational linguistics (Crossley & McNamara, 2014), and corpus linguistic studies of larger data sets (Bulté & Housen, 2014; Friginal & Weigle, 2014; Yoon & Polio, 2017). Most of this research has taken a traditional approach to operationalizing grammatical complexity, relying on ratio- and length-based measures that use the clause as the unit of analysis (e.g., mean length of T-unit, dependent clauses per T-unit[1]). This approach grew out of Hunt's (1966) work on syntactic indices and has become common since Wolfe-Quintero, Inagaki, and Kim's (1998) seminal work (Connor-Linton & Polio, 2014).

In recent years, researchers have advocated for reconceptualizing complexity indices, recognizing that "standard ways of operationalizing syntactic complexity in L2 research have focused on verbal subordination" (Lambert & Kormos, 2014, p. 607). In response, researchers working within this tradition have begun targeting phrasal complexity, proposing measures such as mean length of clause or mean length of noun phrase (e.g., Byrnes, Maxim, & Norris, 2010; Lu, 2011; Norris & Ortega, 2009; Ortega, 2003).

At the same time, the distinctive grammatical characteristics of speech and writing and the differing types of complexity prevalent in those registers have long been acknowledged (Biber, 1988, 1992; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Halliday & Martin, 1996; Wells, 1960). On the basis of documented patterns of grammatical variation across registers, Biber, Gray, and Poonpon (2011) proposed an empirically grounded, register-based approach for measuring complexity that uses rates of occurrence (per *x* words) to capture the frequency of *both* embedded clausal structures (typical in spoken language and leading to elaboration of discourse) and embedded phrasal structures (typical in informational writing and leading to condensed, information-heavy discourse). Because it takes into account the register distributions and functions of complexity features, this approach has been referred to as the *register/functional approach* (Biber, Gray, & Staples, 2016).

Biber et al. (2016, pp. 10 – 11) explained that the two approaches to grammatical complexity are fundamentally different (for a discussion of automated approaches to complexity, see also Lu, 2017):

1. T-unit measures are parsimonious (they result in a single value); they are relatively holistic because they collapse many structural distinctions into one or a small set of measures. As a result, no information is available about the types of linguistic features that result in complex language or their functions relative to the genre/register.
2. The register/functional approach maintains all structural distinctions, allowing for a detailed consideration of individual features and their functions in discourse. However, the approach is not parsimonious, as it results in a separate data point for each feature.

Existing longitudinal studies relying on the T-unit approach have had mixed results. Some studies found very little development, while others found development across multiple complexity measures. The most common measures exhibiting change over time tend to directly or indirectly indicate increased use of phrasal complexity (Bulté & Housen,

2014; Crossley & McNamara, 2014; Ferrari, 2012; Friginal & Weigle, 2014; Polat & Kim, 2014; Spoelman & Verspoor, 2010), while many clausal features did not increase. However, these findings are difficult to interpret because they provide no information regarding what features lead to increased phrasal complexity and thus provide little information about the nature of linguistic developments in learners. Because the goal of the present study was to describe changes in learner language over time, the register/functional approach was adopted.

Several studies have productively applied the register/functional approach to first language (L1) and second language (L2) development research (Parkinson & Musgrave, 2014; Staples, Egbert, Biber, & Gray, 2016; Taguchi et al., 2013). However, this research has largely been restricted to cross-sectional research designs or designs focused on comparing corpora stratified by level—including for TOEFL iBT responses (Biber et al., 2016; Biber & Gray, 2013). In fact, the register/functional approach is well suited to investigating the development of complexity in *TOEFL*® responses. A key feature of the register/functional approach is the idea that the use of features associated with grammatical complexity is related not only to language proficiency but also to the situations (or registers) in which the language is being produced. That is, the approach accounts for (and even expects) different complexity profiles depending on the mode and purpose of the text—differences that are relevant for explaining variation across the modes and task types included in the TOEFL iBT.

A Register-Based Developmental Progression for Grammatical Complexity

The developmental complexity framework used in the present study is designed around the fundamental distinctiveness of speech and writing. Biber et al. (2011) proposed the sequence in response to their desire for an operational definition of developmental complexity that had a strong empirical foundation (based on well-documented register patterns) as well as a strong linguistic basis (able to capture multiple types of complexity). The underlying premises are as follows:

1. Features that are characteristic of speech are acquired earlier, while features occurring more frequently in specialized written registers are acquired later.
2. Grammatical devices are used first in high-frequency lexico-grammatical combinations and then widened to include a wider range of lexico-grammatical patterns and meanings.

Accordingly, a hypothesized set of developmental stages for complexity features was proposed by Biber et al. (2011), represented in Table 1. In the progression, features can be described along two parameters: grammatical structure and grammatical role. Features that are clausal in structure and fulfill syntactic roles within clauses represent clausal complexity features; these are typically much more common in conversation than in writing (Biber et al., 1999; Biber & Gray, 2010) and are thus placed earlier in the hypothesized developmental stages. Frequency effects and lexico-grammar (Premise 2) are further built into the sequence (see Stages 1a and 1b).

In contrast, features that have a phrasal structure and occur as constituents within noun phrases (e.g., nouns as nominal premodifiers) represent phrasal complexity features. These features are much more common in informational writing (Biber et al., 1999; Biber & Gray, 2010) and are thus placed later in the developmental progression. Phrasal features are sequenced according to frequency; for example, attributive adjectives are very common and are thus placed earlier in the sequence than nouns as premodifiers.

Features that are mixed on the two parameters (e.g., relative clauses are clausal in structure but embedded in noun phrases) are placed in intermediate positions in the sequence. In general, the progression moves from finite clauses (which are more common in speech) to nonfinite clauses (which are generally more common in writing; Biber & Gray, 2010; Biber et al., 1999).

Although several language development studies have applied the general premises behind Biber and colleagues' hypothesized developmental stages (e.g., Staples et al., 2016), few have investigated the specific stages summarized here. Parkinson and Musgrave's (2014) study is one exception. Focusing only on the noun phrase complexity features from the framework, Parkinson and Musgrave investigated the extent to which two levels of students (L2 master's students and students enrolled in a lower level EAP course) used the features in their academic writing. Their findings partially supported the developmental sequence, with the higher level students using noun phrase features at higher stages with greater frequency than the EAP students.

Table 1 Proposed Developmental Stages for Phrasal and Clausal Complexity Features Adapted From Biber et al. (2011)

Stage	Complexity features
1	a. Finite complement clauses (*that* and *wh-*) controlled by common verbs (e.g., *think*, *know*, *say*)
2	a. Finite complement clauses controlled by a wider set of verbs b. Finite adverbial clauses c. Nonfinite complement clauses controlled by common verbs (especially *want*) d. Phrasal embedding in the clause: adverbs as adverbials e. Simple phrasal embedding in the noun phrase: attributive adjectives
3	a. Phrasal embedding in the clause: prepositional phrases as adverbials b. Finite complement clauses controlled by adjectives c. Nonfinite complement clauses controlled by a wider set of verbs d. *That* relative clauses, especially with animate head nouns e. Simple phrasal embedding in the noun phrase: nouns as premodifiers f. Possessive nouns as premodifiers g. *of* -phrases as noun postmodifiers h. Simple PPs as postmodifiers, especially with prepositions other than *of* when they have concrete/locative meaning
4	a. Nonfinite complement clauses controlled by adjectives b. Extraposed complement clauses c. Nonfinite relative clauses d. More phrasal embedding in the NP: attributive adjectives and nouns as premodifiers (multiple modifiers) e. Simple prepositional phrases as postmodifiers, especially with prepositions other than *of* when they have abstract meanings
5	a. Preposition + nonfinite complement clause b. Complement clauses controlled by nouns c. Appositive noun phrases d. Extensive phrasal embedding in the NP: multiple prepositional phrases as postmodifiers, with levels of embedding

The Present Study

The goal of this study was to characterize growth patterns in the use of grammatical complexity as exhibited in TOEFL iBT responses produced by EFL learners in China using a longitudinal corpus of their exam responses. It further considered whether their development followed expected paths given (a) existing knowledge about the discourse characteristics of TOEFL iBT responses across mode and task type, (b) hypotheses about the nature and trajectory of complexity development, and (c) expectations that these developments are relative to time and experience in learning English. The following research questions (RQs) were addressed:

1. To what extent are patterns of variation observable in samples of TOEFL responses stratified by level also attested in the longitudinal corpus?
2. How does the use of phrasal and clausal complexity features change over time? Are these changes consistent, or are distinct patterns of change observed as mediated by mode and task type?
3. Considering the features that increase and decrease in use over time, what are patterns in the developmental path for grammatical complexity features?

To answer the first question, the longitudinal corpus was analyzed using Biber and Gray's (2013) multidimensional (MD) analysis model. The purpose of this analysis was to investigate whether test takers' language production developed over time in expected ways. To explore this question, the longitudinal corpus was compared to Biber and Gray's (2013) findings for the much larger TOEFL iBT Public Use data set, which represents the broader TOEFL iBT population, with responses from a wider range of contexts, L1 backgrounds, and proficiency levels. That is, Biber and Gray's analysis established the expected patterns of linguistic variation and, specifically, how that variation was mediated by mode, task type, and score level. To determine whether the learners represented in the longitudinal corpus show speaking and writing development toward expected norms, a direct comparison to the level-stratified corpus was carried out, elucidating how the responses in the longitudinal sample are situated within the expected cline of variation.

The second and third questions were addressed through an analysis of the features included in Biber et al.'s (2011) hypothesized developmental stages for complexity. This analysis was based on the expectation that learners will move from a reliance on features from earlier stages of the progression toward increased use of features from later stages of the

progression. Because of the nature of this framework, the focus of this analysis was on changes in the frequency of use (not in the presence or absence) of these features. Thus, the analysis considered features at each stage and determined to what extent the use of those features increases or decreases over time, as well as how those patterns of change vary according to mode and task type.

It should be noted that the two analyses are complementary but not completely independent. The MD model included some of the complexity features from the developmental stages but also accounts for additional features that have been shown to be important in explaining variation in TOEFL responses. Likewise, the complexity analysis included more specific complexity features and additional features not included in the MD analysis. The complementary perspective enabled a fuller description of the development observed in the longitudinal corpus.

Method

Longitudinal Corpus of Spoken and Written TOEFL iBT Responses

Longitudinal TOEFL iBT Data Set

The present study uses the longitudinal component of Data Set A from Ling, Powers, and Adler's (2014) investigation of the effects of English-language programs and experiences on TOEFL iBT performance. L1 Chinese speakers enrolled in general English and/or TOEFL iBT preparation courses at a high school in China ($N = 90$) took a practice test twice, with 9 months between test administrations (for full details of the longitudinal sample, including test taker characteristics, see Ling et al., 2014, p. 3). The present study focuses on the speaking and writing subsections.

The same prompts were used at both test administrations (Time 1 and Time 2). The speaking section consists of two independent items asking students to provide their opinion on ways for students to relax and whether it is better to stay up late/get up late or go to bed early/get up early. For two integrated speaking items, students read a short passage, listened to a brief audio prompt, and then summarized. The topics included plans for a library renovation and the cause of allergic reactions. The final two integrated speaking tasks asked students to summarize after listening to an audio passage about whether a professor should offer a makeup exam to a student or about generalized versus balanced reciprocity. The writing test included one integrated item in which students read a passage and listened to an audio clip on theories about bird navigation and then produced a written summary. The written independent task asked students to provide their opinion on whether students should choose subjects to study based on interest or job preparation. The full prompts for these items are provided in Biber and Gray (2013, Appendix A, Form 1).

Ling et al. (2014, p. 8) characterized the participant proficiency level as low or intermediate, reporting relatively low initial TOEFL iBT practice test scores in speaking ($M = 15.00$, $SD = 7.68$) and writing ($M = 16.20$, $SD = 7.15$). Their analysis revealed significant gains in writing and speaking proficiency over time as measured by the TOEFL iBT, with moderate gains in writing (effect size $d = .47$) and smaller gains in speaking (effect size $d = .27$; see Ling et al., 2014, p. 8).

While Ling et al.'s (2014) study focused on test scores and their associations with learning contexts, programs, and activities, the present study investigated the linguistic characteristics of the participants' responses over time and by task type. A corpus of the spoken and written TOEFL iBT responses produced by these participants was compiled.

Corpus Compilation and Preparation

ETS provided written responses ($N = 307$; 90 test takers; 4 responses; 53 missing responses) and transcriptions of spoken responses ($N = 1{,}044$; 90 test takers with 12 responses; 36 missing responses). Manual and automated (scripts written in Perl and Python) procedures were developed to reformat/prepare the written and spoken responses for corpus-based linguistic analyses, including converting MS Word files to plain text and addressing formatting issues (making transcription conventions consistent, correcting problematic character encodings, converting unrecognized characters, and reformatting files to UTF-8). To increase the reliability of automatic analysis software, errors due to spelling that were not indicative of developmental patterns (e.g., *thier* instead of *their*; *dameges* for *damages*) were corrected in the written texts, following Biber and Gray (2013); grammatical errors (e.g., subject–verb agreement in **the reading passages provides*) or errors where the intended word was not clear (e.g., *prosaypothity*) were not changed.

The data were then evaluated for missing observations and response length to identify responses that could be analyzed using corpus methodologies. These analyses rely on normalized rates of occurrence (i.e., frequencies that have been standardized as rates per N words), with each text as the "observation" (Biber & Jones, 2009). In general, a minimum text

Table 2 Description of the Longitudinal Corpus of Spoken and Written TOEFL iBT Responses

	Time 1				Time 2			
Mode and task type	No. texts	Total words	Mean length	Max length[a]	No. texts	Total words	Mean length	Max length
Written[b]								
Integrated	39	6,872	176.2	308	39	7,751	198.7	299
Independent	39	10,432	267.5	391	39	12,360	316.9	471
Spoken[c]								
Integrated[d]	42	11,510	274.0	435	42	15,931	379.3	549
Independent[e]	42	6,232	148.4	180	42	7,061	168.1	254
Total	162	35,046	216.3	435	162	43,103	266.1	549

[a]Because text length was restricted, the minimum text length is 100. [b]$N = 39$. [c]$N = 42$. [d]Items 3–6 combined. [e]Items 1 and 2 combined.

length of 100 words is needed to ensure reliable normalized rates of occurrence (Biber, 1990, 1993; see also the discussion in Biber & Gray, 2013, pp. 20–21).

Because of the participants' relatively low proficiency level and technical issues that resulted in some unintelligible audio files, most individual spoken item responses were shorter than 100 words. Thus, spoken independent items (1 and 2) were combined into one observation (i.e., corpus file) and spoken integrated items (3–6) were combined into a second observation per test taker. Note that all six spoken responses were not combined because previous research has shown that grammatical characteristics vary between integrated and independent tasks (e.g., Biber & Gray, 2013; Cumming et al., 2005), which is in line with theoretical expectations that different task types elicit different types of discourse (Bygate, 1999, 2001; Jamieson, Eignor, Grabe, & Kunnan, 2008; Skehan, Foster, & Mehnert, 1998). Each participant's responses were then evaluated according to two criteria to determine whether they would be included in the final corpus: (a) analyzable responses of at least 100 words (with spoken independent and spoken integrated items combined) and (b) the presence of analyzable responses for both task types (within a mode) at both test administrations.

Scoring of Spoken and Written TOEFL iBT Items

Scores for each spoken and written item were provided by ETS. Spoken responses were scored by ETS-trained raters on a scale of 1–4 following standard ETS procedures and rubrics.[2] The speaking rubrics include criteria regarding discourse characteristics (the appropriateness of the response for the task, topic development) and linguistic characteristics (delivery, language use). Scores on the combined spoken items were averaged.

Written responses were scored using the *e-rater*® automated scoring engine, which uses statistical analyses to predict human holistic scores using features related to grammar and usage (including errors), mechanics and style, organization and development, and lexical complexity and prompt-specific vocabulary (Attali & Burstein, 2006; Enright & Quinlan, 2010; Quinlan, Higgins, & Wolff, 2009). *E-rater* produces scores on a scale of 0 to 5.49, which are then rounded to whole integers (i.e., 0, 1, 2, 3, 4, 5). The written responses in Biber and Gray (2013) were rated by human raters on a scale of 0–5 (with scores from two raters averaged, resulting in possible scores of 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5). Biber and Gray converted the writing scores to a scale of 0–4 to make the scores on speaking and writing responses directly comparable (see Biber & Gray, 2013, pp. 11–12). Because the longitudinal corpus in the present study was to be compared to the level-stratified corpus from Biber and Gray, the writing scores for the longitudinal responses were converted to the same 0–4 scale following Biber and Gray (see Appendix A in this report). These converted scores have been used throughout this study for consistency.

Final Corpus Composition

Table 2 summarizes the analyzable subset of participants. Forty-two and 39 participants (out of 90) met the criteria for the spoken and written subcorpora, respectively.[3] Mean text lengths increased in the second test administration (Table 2).

Mean item scores also increased over time (Table 3; Figure 1). A paired-samples *t*-test shows significant increases in the mean scores for each mode and task type, with the most important increases occurring for speaking independent tasks (medium effect of $d = 1.069$) and independent writing (small effect of $d = .722$), following the field-specific recommendations for interpreting effect sizes in L2 research (Plonsky & Oswald, 2014). Although significant, the increases in test

Table 3 Mean Scores by Mode, Task Type, and Test Administration in the Longitudinal TOEFL iBT Corpus

	Time 1		Time 2		Change in scores			
Mode/task	*M*	*SD*	*M*	*SD*	Score Δ	*t*	*p*	Cohen's *d*
Written[a]								
Integrated	1.62	0.91	2.05	0.86	+0.42	3.09	0.004	0.494
Independent	1.85	0.63	2.38	0.85	+0.53	4.45	<0.001	0.722
Spoken[b]								
Integrated	1.82	0.63	2.14	0.65	+0.32	3.95	<0.001	0.496
Independent	2.08	0.47	2.58	0.47	+0.50	5.38	<0.001	1.069

[a]$N = 39$. [b]$N = 42$.

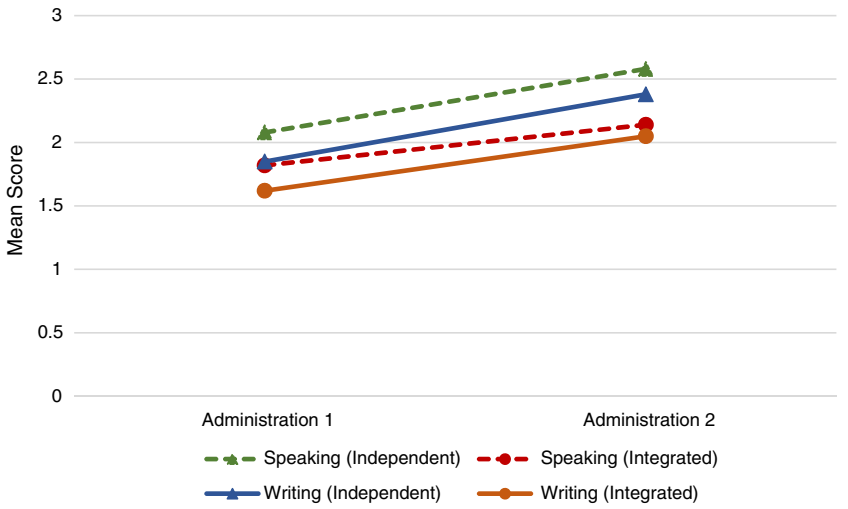

Figure 1 Mean item scores over time by mode and task type.

scores for integrated writing and integrated speaking were smaller in magnitude ($d = .494$ and .496, respectively). Thus the test takers represented in this sample reflect low to intermediate language proficiency (Ling et al., 2014), and their proficiency increased moderately over the time of the study, with the most important gains for independent tasks.

Corpus Annotation

Phase 1: Part of Speech Tagging

All corpus files were part of speech (POS) tagged using the Biber Tagger (Biber, 1988), which uses a combination of contextual rules, large-scale dictionaries, and probabilistic information to annotate each word in a text with morphological, syntactic, and some semantic information (see Biber, 1988; Gray, 2019). A comprehensive accuracy analysis of the tagger was previously carried out on the TOEFL iBT Public Use data set (Biber & Gray, 2013, pp. 15–18) and demonstrated a high degree of accuracy, with most features achieving precision and recall rates greater than .90, and with many greater than .95. Because the TOEFL iBT Public Use data set is similar in nature to the longitudinal corpus used here, accuracy rates and typical errors are assumed to be similar to those reported in Biber and Gray (2013). Thus, the current study followed the same automated and manual procedures to increase the accuracy of grammatical annotation in the longitudinal corpus. After running automated scripts from Biber and Gray, two coders used an interactive program (FixTag) to manually correct tags for three features with lower accuracy rates: present participles, past participles, and all instances of *that*.

While several studies have followed Biber and colleagues' register framework, no studies have specifically used (semi)automated methods to capture the developmental progression proposed by Biber et al. (2011), instead relying on more general features (e.g., Staples et al., 2016) or primarily manual coding (e.g., Parkinson & Musgrave, 2014). This is because the developmental framework makes finer grained distinctions than current tools capture automatically. For example, the Biber Tagger can identify verb *that*-complement clauses but does not distinguish highly frequent verbs (Stage

1a) from a wider set of verbs (Stage 2a). Other features are not tagged (e.g., noun phrases with multiple premodifiers) or require manual analysis (e.g., prepositional phrases as adverbials vs. postnominal modifiers). Phases 2 and 3 of corpus annotation thus targeted the developmental complexity features using manual and automated processes, in which an additional tag for the developmental features was added, resulting in a "complexity tagged" version of the corpus.

Phase 2: Annotation of Prepositional Phrase Features

Six of the features from the developmental complexity framework involve prepositional phrases (PPs; 3a, 3g, 3h, 4e, 5a, and 5d; see Table 1). These stages are distinguished by the grammatical function and meaning of prepositions (e.g., adverbials vs. postnominal modifiers, concrete vs. abstract meanings). The identification of some PP features can be automated: *of*-phrases as postmodifiers (3g) and prepositions followed by nonfinite clauses (5a). Other features such as PPs as adverbials (3a) and PPs as postmodifiers with concrete (3h) and abstract (4e) meanings require manual analysis. In addition, prepositions may fulfill other syntactic roles (e.g., as part of multiword verbs, as adjective complements) that must be excluded. The goal of Phase 2 was to annotate each instance of a preposition for its grammatical/functional role through two steps: (a) preprocessing scripts to assign tags automatically whenever possible and (b) manual coding of prepositions.

Preprocessing scripts were developed to initialize new tagfields for all words and then add tags automatically whenever possible, thus restricting the data set to be annotated manually. All possible preposition tags and operational definitions are included in Appendix B. Coding of the following uses was automated: prepositions in common multiword verbs; prepositions in multiword determiners (e.g., *a lot of*), nouns followed by a postmodifying *of*-phrase (3g), and other specialized uses of prepositions (e.g., *as far as*, *interested in*). The reliability of these scripts were evaluated using a sample of texts,[4] with precision rates of .993 (writing) and .969 (speaking).[5]

To carry out the manual coding process of any unannotated prepositions, the FixTag program ("Phase 1: Part of Speech Tagging" section) was adapted to enable interactive coding in which users could select from the list of possible preposition tags. This ensured consistency in the tags, reduced errors, and increased efficiency. Corpus files were divided into batches, and two coders independently annotated each file. A Python script compared the annotations to assess reliability using percentage of agreement and Cohen's kappa. Because the functions of PPs are not always clear grammatical distinctions (but rather depend on the reader's interpretation of the meaning and/or function of the phrase), reliability ranged between 71% agreement ($\kappa = .577$) and 86% agreement ($\kappa = .757$). Appendix C presents the full reliability results for PP coding. Although 15 of 17 batches achieved "substantial coder agreement" (Cohen's kappa value greater than .6; Galaczi, 2013, p. 7), these rates of reliability did not achieve the target 90% agreement level. Thus, disagreements between the two coders were evaluated by the principal investigator, who assigned the final tag.

Phase 3: Developmental Complexity Tagger

A complexity tagger was developed to identify the remaining developmental complexity features from Biber et al. (2011), with two exceptions: appositive noun phrases and multiple PPs as postmodifiers. These features are particularly characteristic of specialized informational writing (such as published research articles). Given the nature of the TOEFL iBT texts, the relatively low proficiency of the test takers in the corpus, and our observations of the texts, these features were not expected to occur frequently and were excluded. Two features not in the original sequence were added: nominalizations (Stage 4) and nouns with both premodification and postmodification (Stage 5).

The complexity tagger relies on POS tags, contextual algorithms, and frequency-based lexico-grammatical information from the *Longman Grammar of Spoken and Written English* (Biber et al., 1999). For example, Stage 1a (finite complement clauses controlled by common verbs) and Stage 2a (finite complement clauses controlled by a wider set of verbs) are identified using grammatical tags (e.g., *that* tagged as a verb complement) and frequency information from Biber et al.: Verbs occurring with this grammatical structure more than 100 times per million words (Biber et al., 1999, pp. 685–686) were considered as Stage 1a, while all other verbs were considered as Stage 2a. Other features that could not be identified reliably based on tags alone use information from Biber et al. to identify common lexico-grammatical patterns (e.g., extraposed complement clauses typically occur with a restricted set of controlling words). Appendix D gives a description of each feature and its operational definition.

The frequency-based, lexico-grammatical approach used to develop the complexity tagger prioritizes precision (ensuring that an instance tagged as a feature is indeed an instance of that feature) over recall (capturing every instance of a

feature). The accuracy of the complexity tagger was assessed using a 10% sample of the spoken and written subcorpora, including both independent and integrated tasks and a range of score levels (Appendix E; Table E1). A subset of features that occurred infrequently in the sample were further investigated in the full corpus for precision (*wh*-verb complement clauses, *ing*-verb complement clauses, adjective complement clauses, genitive nouns, and nonfinite relative clauses; see Table E2). Since all instances of these features were being manually analyzed in the full corpus, any errors discovered during this process were corrected in the corpus files.

In general, precision and recall rates greater than .90 are considered acceptable, and most features met this criterion. An analysis of the errors revealed that the most common causes of complexity tagger inaccuracies were (a) grammatical errors in the student production; (b) dysfluencies, repetitions, and the lack of punctuation in speech; and (c) incorrect POS tags. Other errors reflected the operational definitions used by the complexity tagger, but these fell within the acceptable range.

One set of features did not achieve the desired level of accuracy: noun phrases with multiple premodifiers. Thus, each instance of the tag hn-4d (head noun with multiple premodifiers) was analyzed in the full corpus, and all associated tags were corrected manually.

Analysis

Multidimensional Analysis Methods

RQ1 compares the patterns of use in the longitudinal corpus with patterns of variation across levels as observed in the TOEFL Public Use data set using MD analysis (Biber, 1988). MD analysis was developed as a means of identifying systematic patterns of variation in a discourse domain by uncovering statistical cooccurrence patterns of linguistic features through factor analysis (Biber, 1988; Conrad & Biber, 2001). These cooccurrence patterns are viewed as "dimensions" of variation that can be interpreted based on the shared communicative functions of the cooccurring features.

There are two types of MD analysis: the generation of new dimensions by conducting an exploratory factor analysis and the application of previously established dimensions to a new corpus (see Conrad & Biber, 2001). The second approach is applied here, as the goal of the analysis was to investigate how the test takers in the longitudinal corpus developed in terms of expected patterns of variation as established for level-stratified data representing a range of proficiency levels in Biber and Gray (2013).

Biber and Gray (2013) identified four dimensions of variation in spoken and written TOEFL iBT responses: (a) literate versus oral responses; (b) information source: text versus personal experience; (c) abstract opinion versus concrete description/summary; and (d) personal narration. Twenty-eight linguistic features contribute to the structure of these four dimensions, listed along with their factor loadings in Table 4. In MD analysis, *positive* and *negative* features refer to sets of cooccurring features that are in complementary distribution: A text with high frequencies of the positive features will typically have low frequencies of the negative features (and vice versa). Each dimension is interpreted based on the underlying shared functions of the cooccurring features and how registers are distributed along the dimension and then provided with a descriptive label. Table 4 summarizes the functional interpretation of the 2013 dimensions (for a fuller description, see Biber & Gray, 2013).

To apply these dimensions to the longitudinal corpus, normalized rates of occurrence per text for these 28 features were generated using the Biber tagcount program. Standardized *z*-scores were calculated for each feature using the means and standard deviations for the TOEFL Public Use data set, the basis of the original MD analysis in Biber and Gray (2013). This method was used to make dimension scores directly comparable to the results across levels from that study. Dimension scores were then calculated for each text in the corpus by summing the positive-loading and subtracting the negative-loading features and serve to characterize the extent to which a text relies on the features associated with each dimension.

Mean dimension scores were calculated for each subcorpus at Times 1 and 2 and plotted graphically alongside Biber and Gray's (2013) results. Biber and Gray's findings are displayed so as to demonstrate the full range of variation along the dimension accounting for variation by mode, task type, and score level. The longitudinal results are plotted as one group, as there was insufficient score variation in the longitudinal sample to divide results into score bands. However, visualizing the changes in dimension scores from Time 1 to Time 2 enables an interpretation of how the linguistic developments do or do not approach established norms (e.g., moving toward task- and mode-specific discourse styles and moving toward discourse styles associated with increased proficiency relative to task and mode).

Table 4 Description of Multidimensional Analysis Model From Biber and Gray (2013)

Model	Positive pole of dimension	Negative pole of dimension
Dimension 1	*Literate responses*	*Oral responses*
Interpretation	Many noun-phrase-based features, long words, and passive voice are associated with written registers, particularly those with an informational purpose, as they package information into complex noun phrases.	Verbs, pronouns, and clausal structures are associated with a more involved (rather than informational) communicative purpose and have been associated with speech-based registers.
Features	Nouns: common nouns (.64) concrete nouns (.64) premodifying nouns (.39) Adjectives: attributive (.61) topical (.40) Prepositions: prepositional phrases (.52) noun + *of* -phrase (.47) Passives: finite (.41) postnominal (.32) Other: word length (.40)	Verbs: present tense (−.33) mental verbs (−.62) modal verbs (−.36) Pronouns: third person (−.55) That-clauses: controlled by likelihood verbs (−.45) that-omission (−.48) Other: finite adverbial clauses (−.31)
Dimension 2	*Information source: text*	*Information source: personal experience*
Interpretation	Nouns, third person pronouns, and communication verbs are used to describe and summarize information from a source (rather than personal information).	First and second person pronouns, along with abstract nouns, are used in conjunction to discuss the speaker/writer's personal experiences.
Features	Nouns: nouns (.37) place nouns (.45) premodifying nouns (.39) Pronouns: third person pronouns (.41) Other: communication verbs (.80) *that*-clauses controlled by communication verbs (.68)	Nouns: abstract nouns (−.37) Pronouns: first person (−.33) second person (−.39)
Dimension 3	*Abstract opinion*	*Concrete description/summary*
Interpretation	Nouns (and verbs) are used to refer to processes and abstract entities create discourse that focuses on abstract concepts.	In juxtaposition to the positive features, language to refer to concrete entities and their actions is used to focus on concrete descriptions or summaries.
Features	Nouns: nominalizations (.62) mental nouns (.51) abstract nouns (.38) noun + to-complement clause (.33) Verbs: mental verbs (.31) Other: word length (.49)	Nouns: concrete nouns (−.38) Verbs: activity verbs (−.47)
Dimension 4	*Personal narration*	(*no label*)
Interpretation	Speakers/writers narrate their stories and experiences using first person pronouns and past tense.	–
Features	First person pronouns (.35) Past tense verbs (.74)	Present tense verbs (−.70)

Paired-samples *t*-tests are used to identify significant differences in dimension scores over time for each of the four subcorpora (with a Bonferroni adjustment for multiple comparisons; α is set at .0125). We follow Plonsky and Oswald's (2014) recommendation for field-specific effect interpretations for effect sizes in L2 research using Cohen's *d* in a within-group research design, with $d \geq 1.40$ interpreted as a large effect size, values from 1.00 to 1.39 as a medium effect, and values from 0.60 to 0.99 as a small effect size. However, Plonsky and Oswald also acknowledged that effect size interpretations can be better contextualized by considering the relative effect sizes rather than viewing effect size interpretation along a rigid scale. Thus, we report all effect sizes as relative markers of the importance of the significant differences.

Dimension scores are also descriptively linked to scores. However, per the advice of a statistics consultant, no inferential statistics to link dimension scores (or dimension score change) to test scores is employed at this time (concerns included a lack of statistical power due to small sample sizes and the narrow score range represented in the longitudinal corpus, resulting in categorical data unsuitable for correlational analysis).

Complexity Analysis Methods

A complexity tagcount program was developed to generate per text normalized rates of occurrence (per 100 words) for each individual complexity feature. Related features that occur in the same stage are summed (e.g., *to*- and *ing*- complement clauses are summed to represent nonfinite verb complements with common verbs in Stage 2c), particularly if any individual feature has a low frequency in the corpus (e.g., genitives as premodifiers are combined with nouns as premodifiers). To summarize the data, rates of occurrence for each feature within a stage are combined to characterize the extent to which the hypothesized stage is used and to analyze how the reliance on earlier and later stages fluctuates over time. That is, this analysis enables an analysis of the shifting frequencies as features associated with earlier stages in the developmental progression are used less frequently from Time 1 to Time 2, and features associated with later stages in the progression are used more frequently from Time 1 to Time 2.

Paired-samples *t*-tests investigate change over time in the use of the complexity stages, with a Bonferroni adjustment of $\alpha = .01$ (.05/5 for each stage). The distribution of features within each stage is then explored descriptively through mean normalized rates of occurrence for each subcorpus and qualitatively through discourse analysis of the texts. As with the MD results, no inferential statistics linking the use of features to test scores is employed at this time, instead focusing on the descriptive patterns of use.

Results

Multidimensional Analysis

The first RQ examines to what extent development in the longitudinal corpus follows expected paths based on the broader patterns of variation in TOEFL responses across mode, task type, and score level, based on the results of an MD analysis. Thus, in this section, the general cline of variation for TOEFL iBT responses is described and established first (based on Biber & Gray, 2013), followed by an exploration of how longitudinal development occurs along that cline. The direct comparison of the patterns of use observed in the longitudinal corpus to Biber and Gray is a key step in answering the first RQ. Biber and Gray's findings provide an empirically based set of target norms for TOEFL iBT responses, establishing how language use is expected to vary in TOEFL iBT responses across mode, task type, and proficiency level. By first establishing how we expect language to vary across mode and task type at different proficiency levels, we can then evaluate whether the changes in language use we observe in the longitudinal sample reflect those expectations. Thus, throughout this section, we present an overview of the expected patterns of variation, followed by the findings for the longitudinal corpus.

While the four dimensions in the MD model established in Biber and Gray (2013) were strongly associated with mode and task type, and to a lesser extent score differences in the larger level-stratified data set (Biber & Gray, 2013, p. 54, Table 16), Dimension 1 had the strongest relationship to score level, and is the most strongly associated with phrasal and clausal complexity features.

Development in the Use of Dimension 1: Literate Versus Oral Responses

Figure 2 shows the distribution of the level-stratified and longitudinal subcorpora along Dimension 1: oral versus literate. The positive-loading features on Dimension 1 include multiple features related to phrasal complexity, including nouns,

Table 5 Dimension 1 Scores Over Time: Literate Versus Oral Responses

	Time 1		Time 2					
Corpus	*M*	*SD*	*M*	*SD*	Δ[a]	*t*	*p*	Cohen's *d*
Writing								
Integrated	4.59	5.33	2.37	4.90	(−)	2.09	0.043	0.335
Independent	−9.49	5.25	−2.73	5.82	+	7.39	< 0.001*	1.184
Speaking								
Integrated	−7.99	4.51	−5.47	4.13	+	2.81	0.008*	0.434
Independent	−11.83	4.33	−9.45	3.87	(+)	2.55	0.015	0.394

[a]Direction of the change, with parentheses indicating a nonsignificant change. *$p < .0125$.

premodifying nouns, noun + *of* -phrases, longer words, and attributive adjectives (see Table 4 for a full listing). As Biber and Gray (2013) pointed out, this constellation of features has been consistently identified in MD analyses and can be associated with written texts with an informational purpose, in which information is packaged into a phrasal discourse style. Three trends can be observed for spoken and written iBT responses in the level-stratified corpus (see discussion in Biber & Gray, 2013; Biber et al., 2016):

1. Written responses consistently use more of the "literate" features than spoken responses (evidenced by the higher mean dimension scores on Dimension 1).
2. Regardless of mode, integrated responses are higher on the "literate" scale than independent responses within the same mode.
3. There is a general trend that higher scoring responses are more "literate" than lower scoring responses when comparing within a mode and task type (e.g., written independent responses with an item score of 4 have higher Dimension 1 scores than written independent responses with lower scores).

Biber and Gray (2013) linked these trends to the situational characteristics of TOEFL iBT 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. (p. 55)

Like the positive features, the negative-loading features on Dimension 1 have also been identified in previous MD studies. They include verbal and clausal features along with third person pronouns, which have been associated with the spoken mode and registers with a communicative purpose that is more personal and "involved." They thus represent many features associated with clausal complexity. In TOEFL iBT responses, spoken responses, independent tasks (which have a personal, argumentative purpose), and lower scoring responses (when compared to the same mode and task type) tend to have a greater reliance on these verbal and clausal features, and thus lower Dimension 1 scores (see Figure 2).

Figure 2 also displays the Dimension 1 scores for each longitudinal subcorpus at both test administrations.[6] With the exception of integrated writing, Dimension 1 scores increase from the first test administration to the second. Paired-samples *t*-tests (Table 5) show that these increases are significant for independent writing ($d = 1.184$) and integrated speaking ($d = 0.434$).

The increases in Dimension 1 scores for integrated and independent speaking and independent writing are aligned with the expected patterns of development: over time, responses became more literate and less oral. In addition, it is important to note that the language of these responses remains consistent with the norms of the mode- and task type-specific patterns: integrated writing has the highest positive Dimension 1 score, followed by independent writing, integrated speaking, and independent speaking. These patterns became clearer and more consistent with expectations over time (with responses at Time 2 better reflecting the expected patterns of use across mode and task type).

Furthermore, Figure 2 shows that at Time 1, spoken and written responses in the longitudinal corpus had negative dimension scores well below the responses of the same mode/task type in the level-stratified corpus. However, by Time 2, independent writing, integrated speaking, and independent speaking all increase their Dimension 1 scores to the point that they are nearly on par with responses from the level-stratified corpus that received similar scores. For example, integrated

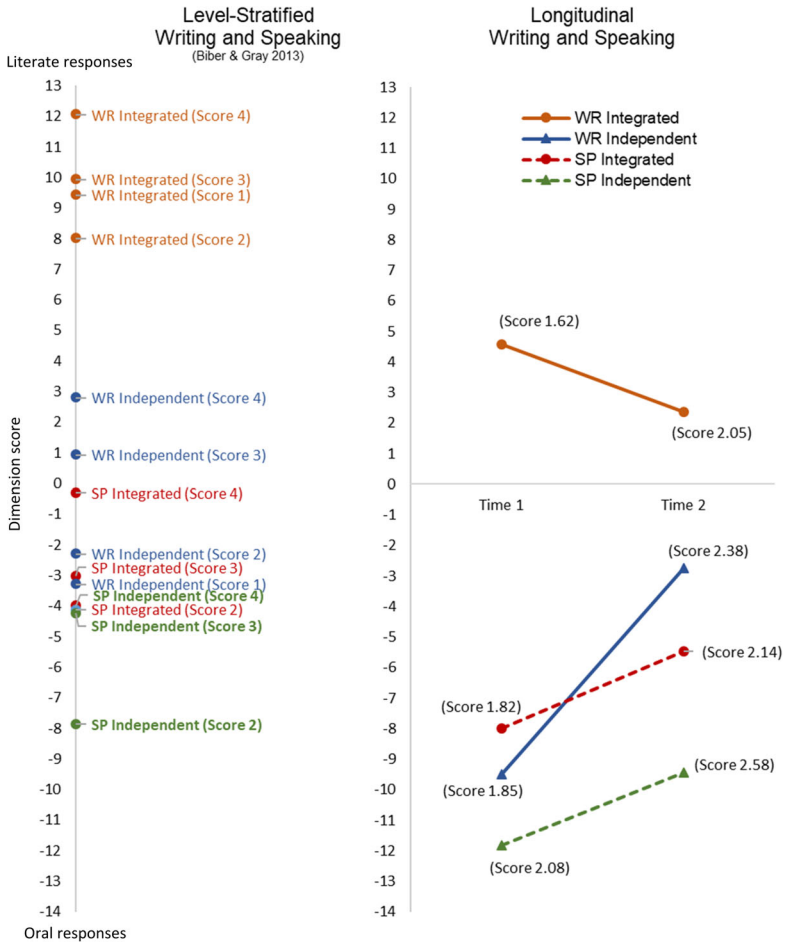

Figure 2 Distribution of level-stratified and longitudinal subcorpora on Dimension 1, literate versus oral (see table). TOEFL item scores are provided in parentheses, based on Biber and Gray's (2013) score groupings. The longitudinal scores represent the mean for all texts in that mode/task type subcorpus.

speaking responses exhibited a mean Dimension 1 score of −7.99 at Time 1, which is well below the dimension score of −4.1 for the lowest level (score 2) for integrated speaking from Biber and Gray (2013). However, by Time 2, the mean Dimension 1 score had increased to −5.47, which is similar to the texts from the level-stratified corpus receiving the same score (2). This pattern is also reflected in independent speaking and writing, with each approximating mode- and task type-specific norms for responses with similar scores. This pattern of development might suggest that the learner production at Time 1 was particularly misaligned with the expectations of TOEFL iBT responses at the start of the EAP courses, but that they were better able to meet these expectations over time (which may be reflected in the increase of TOEFL scores over time documented in "Scoring of Spoken and Written TOEFL iBT Items" section).

Examples 1 and 2 illustrate the shifts in discourse style of one test taker's written independent responses, in which we would expect a more mediated use of positive features (due to its written mode) and negative features (due to the independent task type). Example 1 reflects a text from the first test administration with a very large negative Dimension 1 score (−11.35), thus representing a text that does not use the features in the expected way (and that received a test score of 2). Example 1 demonstrates how the frequent use of negative features (**bolded**) on Dimension 1 create a personal, involved discourse style:

(1) *Written Independent*, *Time 1*, *Score 2*, *Dim 1 Score − 11.35* (*A079*)
First, you **can** learn the subjects you **like** easier and better. Such **is** human nature, everyone **attracts** by the things **they like**. **It will** help you a lot to **learn it**. You **will** be willing to spend much time on **it** and **will** have more energy to learn **it** deeply.

In contrast, this same test taker's written independent response at Time 2 maintains a negative dimension score but is more mediated in its use of positive and negative features (and receives a test score of 3). It exhibits fewer negative

Dimension 1 features (**bolded**) than Example 1 and more of the positive, informational features (underlined), which shifts the argument toward objective information and away from the personal:

(2) *Written Independent*, *Time 2*, *Score 3*, *Dim 1 Score − 3.7* (*A079*)
First, because of the tons of population nowadays, it**'s** not easy for the undergraduates to find job. The unemployed problem **is** crucial in many of the countrys in the world. Students **have to** prepare for their future carefully.

Integrated speaking responses also exhibited significant and meaningful increases on Dimension 1. Examples 3 and 4 illustrate this development. At Time 1, the speaker uses many negative Dimension 1 features (**bolded** and underlined) to produce a response that is focused on an example; there is no indication that this information comes from an external text, and there is no informational component (i.e., mentioning of specific concepts or terms, definitions, etc.), making the response appear very personal:

(3) *Spoken Integrated*, *Time 1*, *Score 1*, *Dim 1 Score − 11.48* (*A026*)
When your brother just, **have** just moved into a new house and **have** no money to buy, a furniture, **buy** furniture, then you **will**, you **will** provide **him** a bed, but when your neighbor, when your neighbor **have** no money to buy furniture, you **will** not provide **him** bed. Um, and this **is** the example from the talk.

In contrast, at Time 2, the response contains fewer negative features (**bolded** and underlined):

(4a) *Spoken Integrated*, *Time 2*, *Score 2*, *Dim 1 Score − 6.44* (*A026*)
The professor, the professor was talking about, reciprocity, and there **is** two types. The first one **is** generalized reciprocity. And it **is Ø** you **give** to, to somebody and **doesn't** expect, and **don't** expect to return immediately. for example, your brother just moved to a new house and **doesn't** have new furniture, and so you **give him** furniture and **doesn't expect** in return, and you, you **know** that **he have to** help you [inaudible]. And the second type **is** balanced, balanced, balanced reciprocity.

The same example story is utilized, but it is framed more explicitly around informational features and indicators that the information is being provided based on an external text. Example 4b highlights the positive features (*italicized* and underlined), demonstrating how the use of positive features creates this informational base:

(4b) *Spoken Integrated*, *Time 2*, *Score 2*, *Dim 1 Score − 6.44* (*A026*)
The *professor*, the *professor* was talking about, *reciprocity*, and there is two *types*. The first one is *generalized reciprocity*. And it is you give to, to somebody and doesn't expect, and don't expect to return immediately. for *example*, your *brother* just moved to a *new house* and doesn't have *new furniture*, and so you give him *furniture* and doesn't expect in *return*, and you, you know that he have to help you [inaudible]. And the second *type* is *balanced, balanced, balanced reciprocity*. it is, you give to, to somebody and return, and expect to return. for *example*, your *neighbor* who have, want to have a *new bed* and you have it and you give to them.

Example 5 illustrates a higher scoring item, demonstrating how spoken integrated responses rely on an even more balanced use of positive and negative features to fulfill the demands of a task that is both spoken (leading to the use of involved/negative features; **bolded** and underlined) and integrated (leading to the use of informational features; *italicized* and double-underlined). The personal story created using third person pronouns, modal verbs, and present tense verbs is still present, while passives, PPs, and complex noun phrases convey specific information:

(5) *Spoken Integrated, Time 2, Item 6, Score 2.75, Dim 1 Score − 2.13* (*A027*)
And the *term reciprocity* is defined by the *anthropology*, is used by the *anthropology* defined the *interrelationship* between *people*. And this **can** be divided into two *type*. For *instance*, *generalized* one. Um, in this *type* and *people* **can** give something to others which **have** a more *closer*, **have** a *closer social relationship*, and they **know** that the others **will** return later. For *example*, if my brother **have**, just moved in *new house* and **he don't** have *enough money* now, and I bought **him** *new bed* and I **know he will** return me, **he will** help me when I get in *trouble*. And the second *type* **is** [inaudible] **type**. This **is** a *type* that, between the *people* who **have** a, whose *social distance* **is**, **is** greater than in the first *type*.

Written integrated responses in the longitudinal corpus begin at Time 1 with a positive dimension score (which is expected). As with the other subcorpora, the mean dimension score is substantially lower than the norms in the level-stratified corpus (regardless of level). However, unlike other subcorpora, integrated writing does not fit the expected pattern of increasing Dimension 1 scores over time (although it should be noted that the observed decrease is not statistically significant). Thus, while test takers in this corpus seem to make up some ground in speaking (both task types)

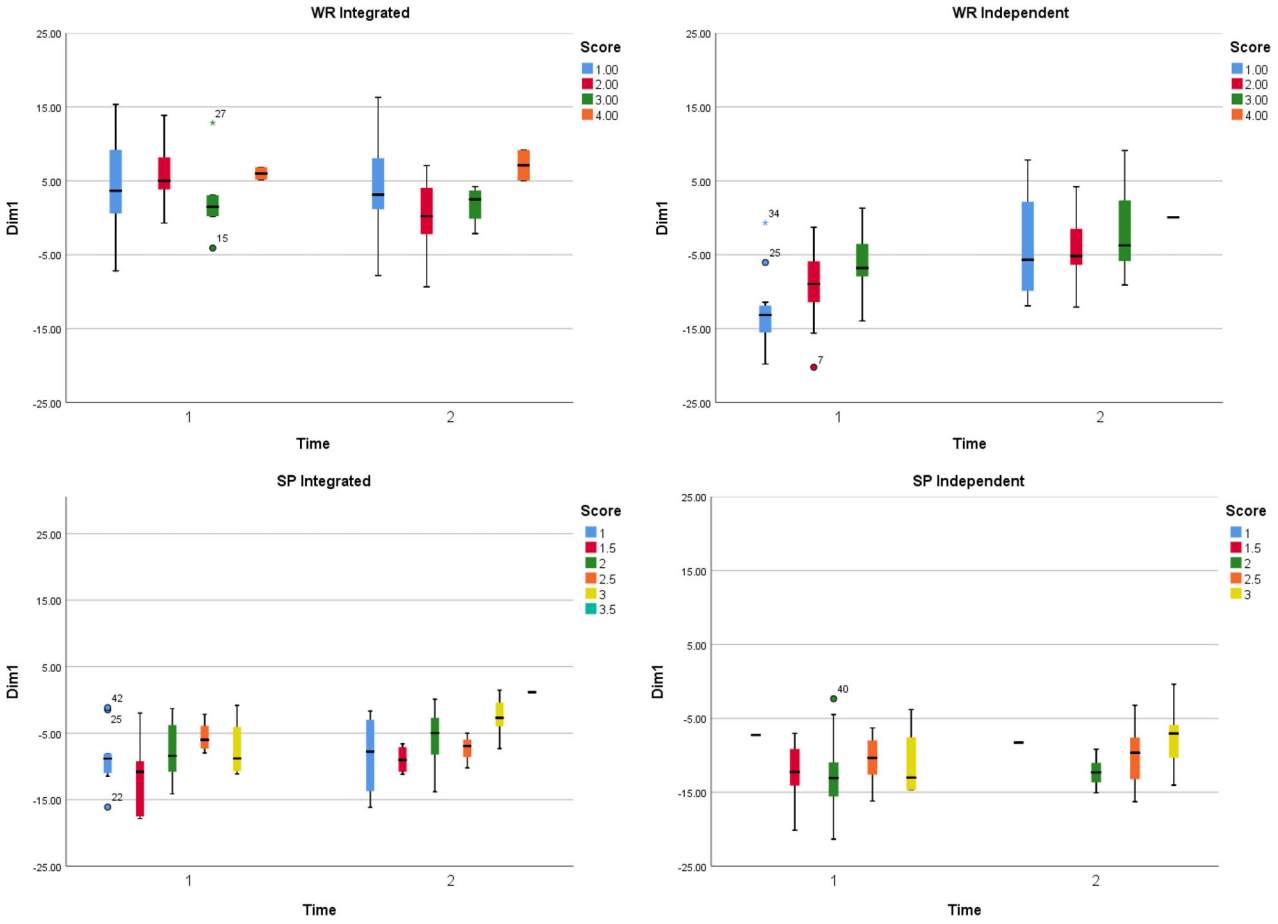

Figure 3 Descriptive relationship between Dimension 1 scores and item score over time. Because multiple spoken items are combined, the mean score of all items is used, resulting in score intervals of 1, 1.5, 2, and so on. Speaking independent scores represent the mean of two items; because four tasks are averaged for integrated tasks, scores were rounded to the nearest .5.

and independent writing, in that their language develops to more closely reflect the expected discourse styles for different modes and task types, this type of development is not observed for integrated writing tasks (note that integrated writing tasks also had the smallest effect size for TOEFL score changes). The apparent lack of development in integrated writing is subsequently discussed in further detail.

Figure 3 descriptively links Dimension 1 scores with score levels across the modes and task types. Although it was not possible to test this relationship statistically, Figure 3 shows fairly consistent patterns which link higher item scores with higher Dimension 1 scores, at least for the second test administration. This pattern is especially clear for independent writing. It is interesting to note that in both speaking task types, the patterns are less consistent at Time 1 and clearer at Time 2. One possible interpretation for this is that less knowledge or ability to produce task type-specific language resulted in fewer distinctions between different types of responses at Time 1, which then became more differentiated over time and as the learners developed. Although a connection between higher item scores and higher Dimension 1 scores should be interpreted with caution, as it has not been statistically established, it appears that this longitudinal data set further supports Biber and Gray's (2013) observation that higher TOEFL scores are associated with more literate discourse styles when considered within mode- and task type-specific norms.

Development in the Use of Dimension 2: Information Source

Whereas Dimension 1 primarily served to make a distinction between spoken and written texts, Dimension 2 cuts across mode to make a primary distinction with respect to task type based on the source of information for the response (see Table 4 for all features associated with Dimension 2). Biber and Gray (2013) linked the positive features of Dimension 2

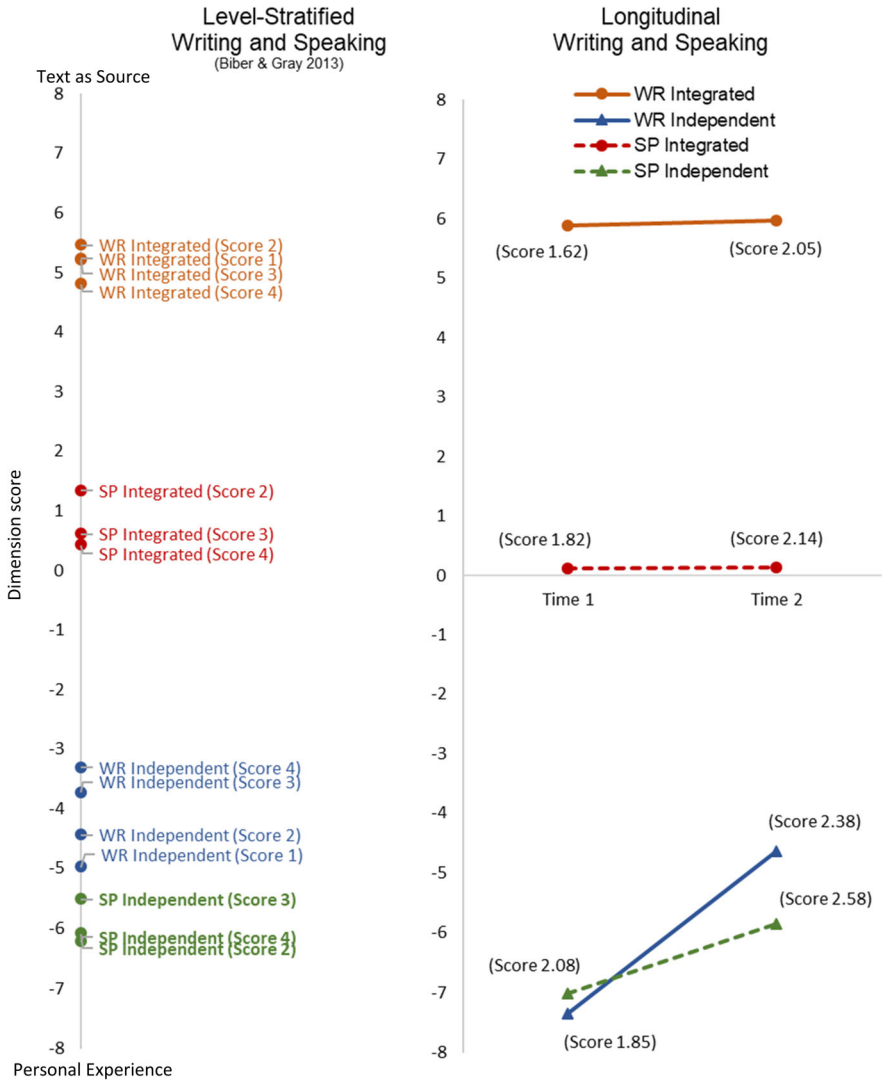

Figure 4 Distribution of level-stratified and longitudinal subcorpora along Dimension 2, information source: Text versus personal experience (see Table 4).

to the goal of describing and summarizing information that is characteristic of integrated tasks: third person pronouns to refer to the writer/speaker of the texts and communication verbs (often with *that*-verb complement clauses) to report what they said or wrote. They further characterize the negative features as functioning to focus on "typical events and consequences based on the speaker/writer's own personal experience" (pp. 57–58). First person pronouns refer to the test takers themselves (e.g., *I think that the students* …), while second person pronouns are used with generic meanings to refer to impersonal/third person meanings (e.g., *if you stay up late at night and then go to sleep late, you will be tired*).

Figure 4 shows the distribution of the TOEFL subcorpora along Dimension 2, showing this primary division between integrated and independent tasks. Biber and Gray (2013) pointed out the inverse relationship between TOEFL score and Dimension scores depending on task type. That is, Figure 4 shows that lower scoring responses have larger positive and negative dimension scores for the dimension associated with the respective task type. Thus lower scoring responses on integrated tasks tend to have larger positive Dimension 2 scores, while lower scoring responses on independent tasks have larger negative Dimension 2 scores. In other words, lower scoring integrated tasks tend to rely more heavily on text-based discourse, while lower scoring independent responses rely heavily on opinions and anecdotes that are more overtly personal.

Paired-samples *t*-tests (Table 6) show significant changes over time in Dimension 2 in independent writing ($d = 0.839$) and independent speaking ($d = 0.450$). There are no significant changes for Dimension 2 scores for either integrated task;

Table 6 Dimension 2 Scores Over Time—Information Source: Text Versus Personal Experience

	Time 1		Time 2					
Corpus	*M*	*SD*	*M*	*SD*	Δ[a]	*t*	*p*	Cohen's *d*
Writing								
Integrated	5.89	3.16	5.97	2.71	(+)	0.13	0.901	0.020
Independent	−7.34	2.13	−4.64	2.99	+	5.24	<0.001*	0.839
Speaking								
Integrated	0.11	2.24	0.13	1.65	(+)	0.05	0.957	0.008
Independent	−7.01	1.81	−5.85	2.15	+	2.92	0.006*	0.450

[a]Direction of the change, with parentheses indicating a nonsignificant change. *$p < .0125$.

however, it should be noted that the mean Dimension 2 scores for these subcorpora are generally aligned with the norm for their respective mode and task type. As with Dimension 1, both subcorpora showing significant changes in the use of Dimension 2 features start out at Time 1 with larger positive and negative dimension scores than expected, with dimension scores that are more extreme than the lowest scoring responses in the respective modes in the level-stratified corpus. Significant increases in the Dimension 2 scores for independent writing and speaking then bring them in line with the level-stratified observations.

Examples 6–9 illustrate how both types of independent responses begin with more extreme negative scores (negative features are **bolded**) and move toward a more mediated use at Time 2. While all responses have negative dimension scores (as expected for independent tasks), what is particularly noteworthy about the differences between Time 1 (Examples 6 and 8) and Time 2 (Examples 7 and 9) is the extent to which the response is framed in personal terms (using first and second person pronouns, underlined) versus the comparative lack of that focus at Time 2:

(6) *Written Independent, Time 1, Score 2, Dim 2 Score − 4.22 (A002)*
Interest is the best teacher when **you** are doing something, if **you** are actually not interested in learning the **subject** **you** choose, **you** will absolutely pay no **attention** on it, also, every **subject** leads to a nice **career** once **you** want to learn it well. […]

(7) *Written Independent, Time 2, Score 3, Dim 2 Score − 1.82 (A002)*
Since the financial **crisis** in 2008, an increasing number of people are unemployed and especially in developing countries. It is predictable that people majored in high-**technology**, **science**, medical, **engineering** and etc. will play a main **role** in the **future** keep-developing **society**.

(8) *Spoken Independent, Time 1, Score 2.5, Dim 2 Score − 8.28 (A024)*
And if **you** stay up late, it will, it will cut down, cut down your, the **quality** of your sleep and **you** will feel very tired tomorrow. However, if **you** go to bed early at night and wake up early, **you** will, **you** will, **you** will, **you** will be very **energy**, energetic and **you** will, **you** will get the most fresh air in the morning, so in my opinion, is go to bed early at night and wake up early is a good idea.

(9) *Spoken Independent, Time 2, Score 3, Dim 2 Score 3.65 (A024)*
I think that it is better to go to bed early at night and wake up early in the morning because of these three **reasons**. The first **reason** is go to bed early and wake up early is good for people's physical **health**, and this, it can help people to have a good body and they can - and this can lead to a good, good **study**. And the second **reason** is in this, in this **plan**, it can make a good **habit** for people.

Figure 5 descriptively links Dimension 2 scores with score levels across the modes and task types. No clear relationship to score is apparent for integrated task types regardless of mode, which is not unexpected given the lack of change in Dimension 2 scores in these subcorpora. A general trend of higher Dimension 2 scores with higher scoring responses is observed for independent tasks (with the exception of independent speaking at time 1). These results are not surprising, given that Biber and Gray (2013, Table 16) did not reveal a significant main effect for score for Dimension 2, although there was a significant interaction effect between task and score (likely reflecting the apparent relationship to score for independent tasks but not integrated tasks).

Development in the Use of Dimension 3: Abstract Versus Concrete

As Figure 6 shows, Dimension 3 identifies clear divisions between independent writing, integrated writing, and speaking (both task types). Biber and Gray (2013) interpreted the positive features of Dimension 3 (mental and abstract nouns,

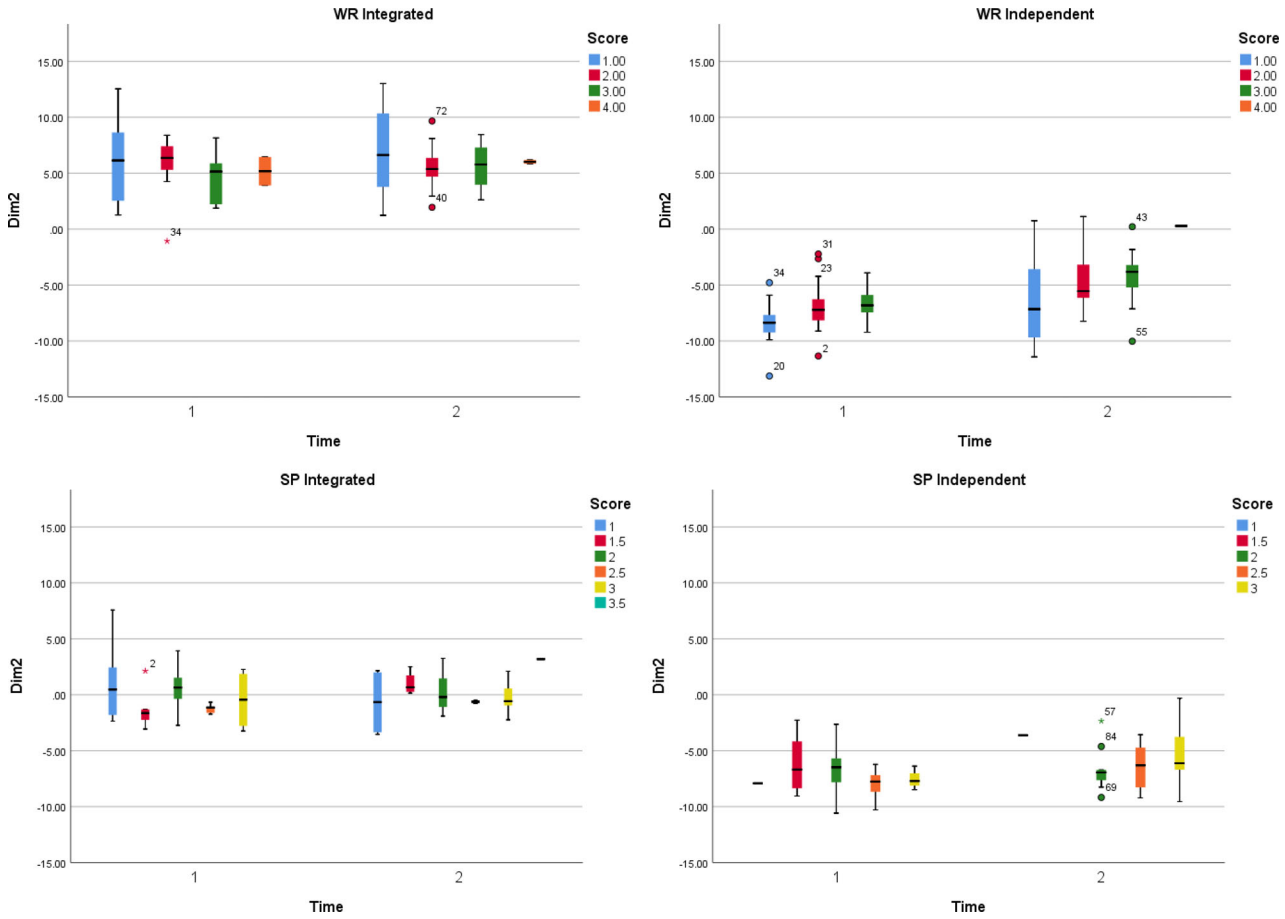

Figure 5 Descriptive relationship between Dimension 2 scores and item score over time. Because multiple spoken items are combined, the mean score of all items is used, resulting in score intervals of 1, 1.5, 2, and so on. Speaking independent scores represent the mean of two items; because four tasks are averaged for integrated tasks, scores were rounded to the nearest .5.

mental verbs, long words, and nominalizations) as functioning to convey abstract concepts, while the negative features (concrete nouns and activity verbs) are used to discuss more concrete entities (see Table 4 for a full accounting of all features associated with Dimension 3). All spoken responses fall on the concrete end of this continuum because of the prompts' focus on "life choices and normal everyday practices" (Biber & Gray, 2013, p. 60), such as when to go to bed and ways to relax. In contrast, written independent tasks are the most abstract, which they attribute to prompt topics about "larger personal/societal issues" (p. 60), such as how students should select what subjects to study. Finally, written integrated tasks fall in the middle, having more abstract language than speech due to the ability to plan, but being more concrete than independent writing because of the topics (e.g., bird navigation) and the reliance on external texts which contain concrete examples that students then use in their responses (p. 60). While this dimension reveals clear mode/task type distinctions, Biber and Gray found no statistical relationship with scores in TOEFL iBT data. Thus, the analysis here will focus on changes in Dimension 3 scores without linking these to score data.

Table 7 presents paired-samples *t*-tests for Dimension 3 scores in the longitudinal corpus, while Figure 6 displays the development over time graphically. The change in Dimension 3 scores is significant for two of the subcorpora: Scores *decrease* significantly in integrated writing ($d = 0.701$) and *increase* in independent speaking ($d = 0.825$).

At Time 1, integrated writing responses tended to have large positive Dimension 3 scores, indicating a more abstract orientation. In Examples 10 and 11, abstract features are **bolded** or underlined, while concrete features are *italicized*. Example 10 shows a more frequent use of abstract features at Time 1, particularly mental nouns and verbs and abstract nouns. Fewer of these features occur in Example 11 (Time 2); instead, there are more nouns and activity verbs, which describe the actions of the birds concretely:

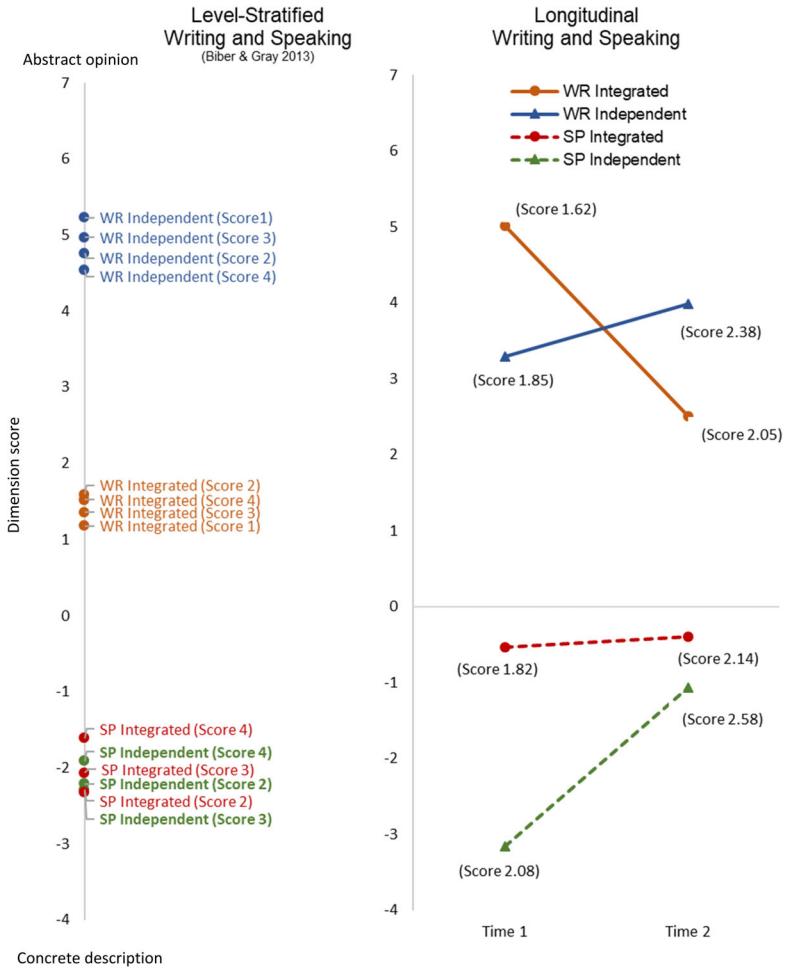

Figure 6 Distribution of level-stratified and longitudinal subcorpora along Dimension 3, abstract opinion versus concrete description (see Table 4).

Table 7 Change in Dimension 3 Scores Over Time—Abstract Opinion Versus Concrete Description

	Time 1		Time 2					
Corpus	*M*	*SD*	*M*	*SD*	Δ[a]	*t*	*p*	Cohen's *d*
Writing								
Integrated	5.01	2.97	2.50	1.56	−	4.37	<0.001*	0.701
Independent	3.29	2.16	3.99	1.71	(+)	1.86	0.071	0.298
Speaking								
Integrated	−0.53	1.46	−0.40	1.41	(+)	0.42	0.680	0.064
Independent	−3.16	1.81	−1.07	2.63	+	5.35	<0.001*	0.825

[a]Direction of the change, with parentheses indicating a nonsignificant change. *$p < .0125$.

(10) *Integrated Writing, Time 1, Score 1, Dimension 3 Score 7.89 (A052)*
From *listening* to the material. we **know** that *bird* has three **ways** to **find** the correct **direction** they **want** to *go* to. One **theory** suggests that *birds* navigate in **reference** to celestial *objects* like the *sun* and the *stars*. For **example**, some *birds* migrate by day *stay* on **course** by orienting their **flight** relative to the *sun's* across the *sky*. […]

(11) *Integrated Writing, Time 2, Score 2, Dimension 3 Score 2.79 (A052)*
In the reading material the speaker **indicates** three **possibilities** that may lead the *birds* go back home, however, in the listening material, the *professor* directly **deem** that these **interpretations** are all not complete. In reading material, the *birds* can be *guided* by the celestial *objects* such as *sun* and *stars*. Actually, it **ignore** the **circumstance** when it was night with *cloud* and there is neither *stars* nor *moon*. *Birds* have no **access** to *see* the *objects* in the *sky* for there are too much *clouds*.

Independent speaking responses in the longitudinal corpus have a large negative dimension score compared to their counterparts in the level-stratified corpus, with increases over time to near zero. Examples 12 and 13 illustrate how these responses become more mediated in terms of their use of positive and negative features over time. Positive features are **bolded**, while negative features are *italicized*. In Example 12, the focus is on specific activities, such as *playing* and *running*. In contrast, in Example 13, abstract and mental nouns and verbs are used to focus on the more abstract aspects of what can be done to help students to relax:

(12) *Independent Speaking, Time 1, Score 2, Dimension 3 Score − 3.18 (A038)*
The best **way** I **think** for the *student* to *relax* after *working* hard is to *play*, um, tennis. I **think** tennis is a very interesting sport. It, um, can *make* us *relax* because its need us to *run* fast and need us to, um, *play* it with **happiness**, and where I *play* the tennis, I feel very interesting.

(13) *Independent Speaking, Time 2, Score 2.5, Dimension 3 Score 0.41 (A038)*
In my **opinion**, um, in my **opinion**, the best **way** for a *student* to *relax* after *working* hard is *listening* music, and there are two **reasons** I *give*. The first one is that *listening* music can help, *relax* our—can help, *make* our **mood** be more, um, comfortable, because *listening* to music can *take* us different [unintelligible] about other things, and the second, *listening* music in—in our spare **time** […]

Results for Dimension 4 (personal narration) are omitted here but are available from the corresponding author upon request. Dimension 4 was the least important dimension in Biber and Gray (2013), accounting for the least amount of variance and made up of only a few features. In addition, the analysis reveals no significant change in dimension scores for the longitudinal corpus on Dimension 4.

Discussion

Comparing the longitudinal corpus to established patterns of use through MD analysis has revealed several consistent observations and has placed language development over time within the broader cline of variation in TOEFL iBT responses. One general trend is that these dimensions of variation appear to be quite stable in differentiating mode- and task type-specific language on TOEFL iBT responses, even for very different populations of test takers. Across the three dimensions examined here, test takers utilized discourse styles that were generally consistent with expected patterns, indicating that the test items elicited similar discourse styles from different populations. While the test takers represented in the longitudinal corpus were all L1 Chinese speakers with relatively low English proficiency enrolled in EAP/TOEFL preparation courses in an EFL setting, the test takers in the level-stratified corpus represented a range of L1s, backgrounds, and score levels. Despite this, the longitudinal results show that the Chinese EFL students maintained mode- and task type-specific patterns of use across the dimensions (e.g., on Dimension 1, written integrated responses were the most literate, followed by written independent, then spoken integrated and spoken independent). This occurred consistently across all three dimensions examined here, particularly at the second test administration.

Learners also exhibited increased task type differentiation. That is, over time, the test takers in the longitudinal corpus began to produce more specialized language for specific modes and task types. Support is provided for this claim by the somewhat mixed patterns of use at the first test administration (e.g., larger positive and negative dimension scores than expected, and responses from different modes/task types which were similar linguistically), which then resolved into clearer mode- and task type-specific patterns of use at the second test administration (for some modes and task types).

It is important to note that most changes in patterns of language use observed in the longitudinal corpus, and all significant changes, reflected developments toward the mode- and task type-specific norms established in the larger level-stratified corpus. Table 8 summarizes the significant changes. The test takers in the longitudinal corpus showed the most development in independent responses. In writing, independent responses became more literate (utilizing noun phrase complexity features to a greater extent over time) and oriented toward more impersonal rather than personal sources of information. Spoken independent responses also showed substantial development, becoming less personally oriented and more abstract.

Integrated tasks, which saw smaller gains in TOEFL scores, also saw fewer linguistic developments. However, like independent tasks, some developments more closely aligned those tasks with the expected language patterns. One possibility for the differential findings between independent and integrated tasks is that integrated tasks are more cognitively

Table 8 Summary of Score and Dimension Score Changes in the Longitudinal Corpus

Mode/task	TOEFL score	Dimension 1 (literate vs. oral)	Dimension 2 (information source)	Dimension 3 (abstract vs. concrete)
Written integrated	(+)	(−)		−
Written independent	+	+ +	+	(+)
Spoken integrated	(+)	(+)		
Spoken independent	+ +	(+)	(+)	+

Note: For significant results only; + indicates an increased dimension score, while − indicates a decreased dimension score). + + + or − − − effect size > 1.4 (large). + + or − − effect size > .1.0 (medium). + or − effect size > .6 (small). (+) or (−) effect size > .2.

demanding (Enright et al., 2008). As the learners represented in the longitudinal corpus had relatively low proficiency levels, it may be expected that developments would occur first in the more familiar and less demanding independent tasks, with developments in integrated tasks occurring at a later stage of development or after a longer period of EAP training.

Although it was not possible to statistically assess the link between TOEFL item scores and discourse patterns (i.e., dimension scores) in this study, the analysis presented here has offered preliminary support for the consistency of scores across contexts for responses exhibiting similar language patterns. It has also shown that there are descriptive relationships between scores and language use, particularly for Dimension 1. It is hoped that the preliminary findings from this study can be further explored with larger longitudinal data sets representing a wider range of proficiency levels to provide statistical evidence with respect to these observations.

Developmental Complexity Analysis

The second and third RQs investigate to what extent the use of phrasal and clausal complexity features changes over time in mode- and task type-specific ways to identify patterns in the developmental paths for complexity features. Because the patterns of development are expected to be different in speech versus writing, the results for each mode are presented separately. Within these analyses, results are presented by first examining the shifting frequencies of use over time for features grouped by stage. Unlike other developmental sequences, where learners are expected to move from producing one stage to the next (e.g., in terms of question formation), the register/functional developmental progression focuses on shifting frequencies of use that reflect fundamentally different ways of packaging information (clausal vs. phrasal). That is, learners do not abandon Stage 1 features and move on to Stage 2 features; rather, at earlier levels, the expectation is that learners will rely most on, for example, verb *that*-complement clauses with the most common verbs (Stage 1). As they develop, they still use those common verbs, but less frequently, as they shift to use other, less common verbs with greater frequency. Thus, the focus of this analysis is to look at each stage over time to see whether increases or decreases are observed over time.[7]

Development in Writing

The register basis of the developmental complexity framework leads to the hypothesis that in writing, features related to clausal complexity decrease over time, while phrasal features increase. It is also expected that these changes would be mediated by task type: Given the more informational nature of integrated tasks, we would expect these developments to occur more markedly in integrated writing. These expectations are further supported by the results for Dimension 1 in the MD analysis, in which the large positive dimension scores for integrated writing reflect the more frequent use of some of the phrasal complexity features under investigation in the fuller complexity analysis (however, integrated did not significantly change in its Dimension 1 score).

Tables 9 and 10 present the mean rates of occurrence for the complexity features grouped by stage at Time 1 and Time 2 in integrated and independent writing, respectively. Note that because the different stages contain different numbers of features, and because some features are generally more common than others, the focus of this analysis is not on the absolute frequency differences between the stages (i.e., that Stage 1 features are less common than Stage 2 features), but

Table 9 Change in the Use of Complexity Features (Per 100 Words) by Stage Over Time in Integrated Writing

	Time 1		Time 2					
Stage	*M*	*SD*	*M*	*SD*	Δ[a]	*t*	*p*	Cohen's *d*
Stage 1	0.92	0.72	1.21	0.82	(+)	1.546	0.130	0.248
Stage 2	8.58	2.43	10.37	2.14	+	3.855	0.000*	0.617
Stage 3	11.67	3.11	10.70	2.34	(−)	1.513	0.139	0.242
Stage 4	5.93	2.25	4.64	1.68	−	2.752	0.009*	0.441
Stage 5	1.52	0.86	1.57	0.95	(+)	0.182	0.856	0.029

[a]Direction of the change, with parentheses indicating a nonsignificant change. *$p < .01$.

Table 10 Change in Use of Complexity Features (Per 100 Words) by Stage Over Time in Independent Writing

	Time 1		Time 2					
Stage	*M*	*SD*	*M*	*SD*	Δ[a]	*t*	*p*	Cohen's *d*
Stage 1	0.67	0.50	0.33	0.33	−	3.897	0.000*	0.624
Stage 2	6.81	2.29	8.18	1.53	+	3.448	0.001*	0.552
Stage 3	7.67	2.23	9.40	2.22	+	3.631	0.001*	0.581
Stage 4	2.12	1.16	3.32	1.55	+	4.376	0.000*	0.701
Stage 5	0.80	0.78	1.30	0.91	+	2.954	0.005*	0.473

[a]Direction of the change, with parentheses indicating a nonsignificant change. *$p < .01$.

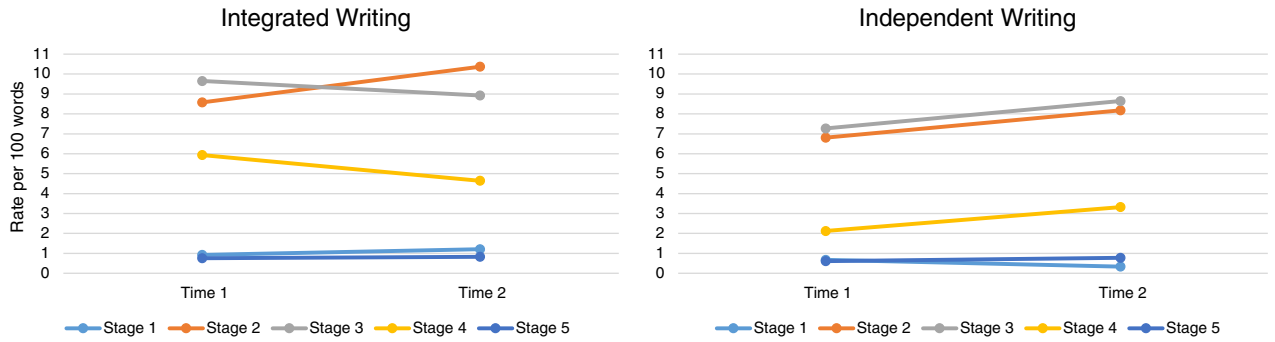

Figure 7 Rates of occurrence for complexity features by developmental stage in integrated and independent writing.

rather on the shifts in the use of features across Times 1 and 2 and the relative frequency of the same stage between integrated and independent writing.

Table 9 shows a significant increase for Stage 2 features ($d = .617$) and a significant decrease for Stage 4 features ($d = .441$) in integrated writing. Increases in Stages 1 and 5 and the decrease in Stage 3 are small and not significant, (also see Figure 7). Similar to the MD analysis results for Dimension 1, integrated writing shows few significant and important changes, and not all changes follow the expected paths.

In contrast, Table 10 reveals significant changes in all five stages in independent writing. Furthermore, these shifts reflect expected patterns of development (see Figure 7), with a decrease in Stage 1 clausal features ($d = .624$) and increases in Stages 2–5 ($d = .473$–$.701$). This pattern of development also reflects the hypothesized paths, with Stage 1 features decreasing over time and the strongest increases in Stage 4, which is primarily focused on the development of nonfinite clauses and phrasal noun modifiers (nominalizations, nouns with multiple premodifiers, and PPs as abstract postmodifiers).

Figure 7 also enables a meaningful comparison of the use of complexity features across task types in writing, demonstrating its mediating effect. Despite the lack of increases in phrasal features (captured especially in Stages 3 and 4) in integrated writing and the presence of increases in independent writing, Figure 7 shows that Stages 3 and 4 are consistently

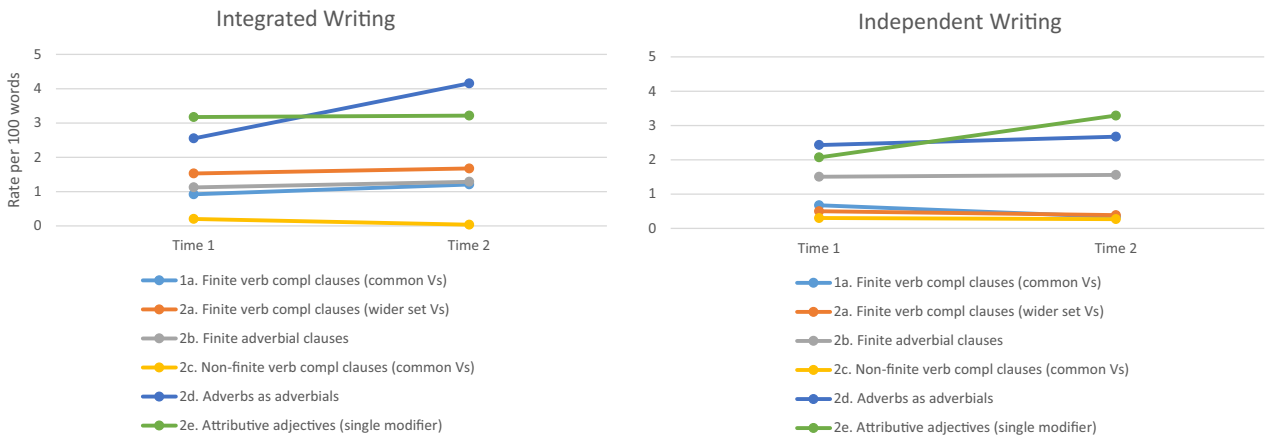

Figure 8 Mean rates of occurrence for Stages 1 and 2 features in writing.

more frequent in integrated rather than independent tasks. Thus, consistent with the MD analysis, more development appears to be occurring in independent writing, but integrated writing does retain its expected discourse style relative to independent writing.

In the following, a closer look at the individual features within each developmental stage is undertaken. Descriptive statistics for each feature are provided in Appendix F. Figure 8 displays the mean normalized rates of occurrence for Stage 1 and 2 features. Adverbs as adverbials contributed to the increase in Stage 2 features in integrated writing, with much smaller increases in finite verb complement clauses and adverbial clauses. Functionally, the increases in adverbs in integrated writing (and the smaller increase in independent writing) are largely due to the increased use of linking adverbials at Time 2. Examples 14 and 15 illustrate how one test taker used linking adverbials (including two phrasal/clausal structures) frequently and with a range of meanings (resultative, additive, contrastive) at Time 2 to explicitly mark relationships between ideas and organize discourse, while rarely using this feature at Time 1.

(14) *Integrated Writing, Time 1, Score 2 (A082)*
Another point is that birds navigate by landmarks. They have ability to the hippocampal region. And its ability to navigate is impaired as well. **Thus** migrating birds must be using memorization skills to navigate. The third theory proposes that birds use a type of internal compass that responds to Earth's magnetic field.

(15) *Integrated Writing Time 2, Score 2 (A082)*
Firstly, the lecturer said that it is unreal that the birds can identify the directs by celestial objects such as the Sun or the stars. It is often full of clouds in the sky and they will disrupt the eyesights of the birds. **As a result**, the birds cannot see the stars or even the Sun in the cloudy. **Secondly**, from the reading passage, it insisted that the birds navigate by the landmarks like the mountains, rivers and so forth. The professor, **however**, think that it is also limited in this theories. The birds just can remember the matters a few minines. **Thus** they cannot find the way. **What's more**, the birds always migrate to the new locations that are unknown to them.

In independent writing, Figure 8 shows a decrease in finite verb complement clauses from Stages 1 and 2 and slight increases in Stage 2 features, such as adverbs (again, due to an increased use of linking adverbials) and finite adverbial clauses. However, independent writing is particularly marked by its more substantial increases in one of the main phrasal complexity features: attributive adjectives as a single modifier in a noun phrase. Attributive adjectives are the lowest + phrasal, + noun phrase constituent feature in the sequence and are thus one of the first features we would expect to see increase.

In addition to increases in frequency, writers of independent tasks at Time 2 used a wider range of adjectives (154 types) than at Time 1 (91 types). Common adjectives at Time 1 included words such as *best*, *better*, *future*, *good*, *great*,

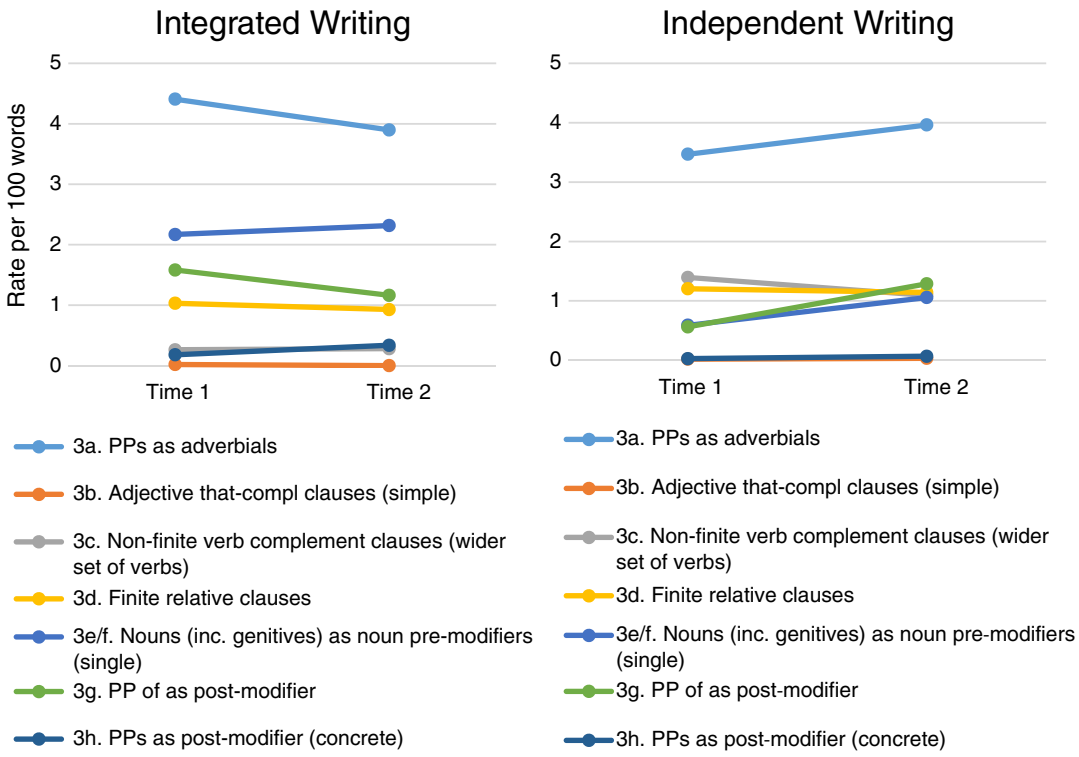

Figure 9 Mean rates of occurrence for Stage 3 features in writing.

happy, *high*, *human*, *important*, *interested*/*interesting*, *useful*, and *wonderful*, all adjectives that are generally morphologically fairly simple and/or highly frequent words in English. Common adjectives at Time 2 included some of these same adjectives, along with additional, less frequent adjectives (*basic*, *boring*, *certain*, *common*, *different*, *famous*, *favorite*, *main*, *major*, *simple*). However, a number of lower frequency, morphologically complex adjectives also appear at Time 2 (e.g., *complicated*, *comprehensive*, *consequential*, *contemporary*, *disastrous*, *distinguished*, *increasing*, *outstanding*, *practical*, *promising*, *successful*). Thus the increased frequency appears to be accompanied by a higher degree of diversity and functional expansion in the use of adjectives.

It should be noted that in integrated writing, the frequency of this feature was relatively high at Time 1 and remained stable over time. However, an investigation of the attributive adjectives used in integrated tasks at both Time 1 and Time 2 shows a frequent use of topic-specific adjective–noun combinations from the source texts (*celestial objects*, *important role*, *internal compass*, *magnetic crystals*, *magnetic field*, *migrating birds*, *navigational ability*, *spatial ability*), indicating a high reliance on input language in the integrated responses. This is in part expected, as the adjective + noun combinations are a mechanism for conveying the precise and specific information that the integrated task requires test takers to summarize. However, the reliance on prompt language may also explain why there is a high frequency of this feature (writers do not have to come up with the language patterns on their own) and why the frequency does not change over time (the task requires writers to convey precise, given information and does not require development to increase the frequency, because many of the sequences are provided by the prompt).

Figure 9 focuses on Stage 3 features, which did not change significantly in integrated writing. There is a slight decline in PPs as adverbials and a slight increase in nouns as single premodifiers—both changes which would be predicted by the developmental sequence as writers move toward packaging information in noun phrases.

In contrast, the changes for Stage 3 in independent writing are significant ($d = .581$). The largest increases occur for PPs as adverbials, nouns as premodifiers, and *of*-phrases as postmodifiers. It is particularly noteworthy that two clausal features in Stage 3 decrease in independent writing: nonfinite verb complement clauses and finite relative clauses; this rise in phrasal features and accompanying decline in clausal features is consistent with the expected

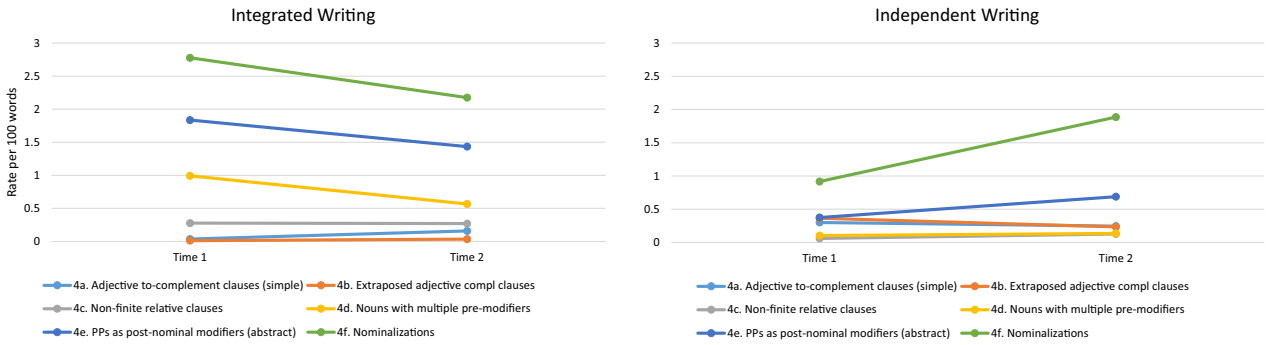

Figure 10 Mean rates of occurrence for Stage 4 features in writing.

path of development. However, their placement in the same stage of development may be an indication that the developmental sequence would benefit from further refinement in separating these clausal and phrasal features into different stages.

Examples 16 and 17 illustrate these three changes in one student's independent writing. PPs as adverbials are underlined, nouns as premodifiers are **bolded**, and postmodifying *of*-phrases are double-underlined. These examples illustrate how the use of adverbial PPs functions to add details or link discourse, while postnominal *of*-phrases and nouns as premodifiers specify the head noun, adding information in a condensed manner. Example 17, which uses these features with higher density, conveys more specific information, rather than retaining the more general and personal orientation of Example 16.

(16) *Independent Writing, Time 1, Score 1 (A043)*
Many students have a problem about which **study** subjects to choose. A great number of them think it is more important to choose the subjects they are interested in that choose subjects to prepare for a job or career. I agree with the opinion.

(17) *Independent Writing, Time 2, Score 1 (A082)*
With the more attention of education, the problem of whether the students should option the subjects to prepare for a job or career or not have also attract **people'** focus. In my mind, the students should choose the subjects that they are interested in. To begin with, the interest is one of the best teachers that can spur your motivation. The interest of one thing can help you focus on the thing and you are willing to pay more energy on it. […] As me, I decide to do to the university to study the **cloth** design. I want to be a well-known **cloth** designer.

Stage 4 features also changed significantly in independent writing ($d = .701$) and integrated writing ($d = .441$). However, Figure 10 shows that three features exhibit very different behaviors in the two writing tasks. In integrated writing, nominalizations, abstract PPs as postmodifiers, and nouns with multiple premodifiers *decrease* in frequency, while those same features *increase* moderately in independent writing. Adjective complement clauses (*to*-clauses, extraposed clauses) and nonfinite relative clauses do not occur frequently in these corpora.

The lack of development toward an increased use of these informational features in integrated writing (and instead an observed decrease) is difficult to interpret (although it is consistent with the MD analysis results). On one hand, it should be noted that all three features start out with higher frequencies in integrated writing and maintain that higher level of frequency despite the decreases from Time 1 to Time 2. That is, writers still use these features more frequently in their integrated writing than in independent writing. Examples 18–20 demonstrate this for one learner; Example 18 reflects a frequent use of these features even at Time 1 in integrated writing, while Example 19 shows the lack of these features at Time 1 in independent writing. Example 20 shows an increased used of these features at Time 2 in independent writing. Like other noun-based and phrasal features from the framework, these features function to position information in nominal structures (nominalizations are **bolded**, abstract PPs as postmodifiers are underlined, and nouns with multiple premodifiers (*italicized*) are in SMALL CAPS). (Note that Example 18 also exhibits many of the Stage 3c informational features.)

(18) *Integrated Writing, Time 1, Score 4 (A061)*
They use rivers, coastline, and so on for **navigation**. In contrast, the speaker is opposed to this theory [… This **situation** can fight against the theory supported by the passage. Third, the essay says that the birds use a type of internal compass that responds to *Earth's magnetic* FIELD. In this theory, birds can sense the way *Earth's magnetic* FIELD pulls on the magnetite crystals. This is used to direct the **navigation direction**. Contrastly, the man in the listening material says that the compass is not enough for the birds to find their ways in some specific location.

(19) *Independent Writing, Time 1, Score 2 (A067)*
I live under the **impression** that to choose to study subject to prepare for a job or career is better than which you are interested in. Why some people choose the first one not the second one is beyond me. The subject which you are interested in you may study good.

(20) *Independent Writing, Time 2, Score 3 (A067)*
An issue, whether it is more important to choose to study subjects you are interested in that to choose subjects to prepare for a job or career, has arouse a heated **discussion** among people. As far as I am concerned that choosing a subject you are interested in is more important. What is **priority** is that a subject you are interested in can evoke you to study better and do well in it.

Nouns with multiple premodifiers are generally rare in these writing corpora, but they are more frequent in the integrated responses. However, upon further examination, it appears that nearly all of the sequences of multiple modifiers used at Time 1 reflect prompt language: *birds' navigational abilities*, *migrating birds' ability*, *Earth's magnetic field*, *bird's hippocampal region*, *sun's east/west path*. However, it appears that at Time 2, novel combinations are just beginning to appear. The same prompt sequences are present, but some learners begin to use novel combinations, such as *important memory formation*, *different new locations*, *predominant migration method*, *serious counter arguments*, *bird's precise navigation*. However, these sequences are rare. One possible explanation is that the learners represented in the longitudinal corpus are at the beginning stages of developing phrasal complexity features. They are beginning to use novel complex expressions in integrated writing by Time 2. The decreases in phrasal complexity features at Time 2 for integrated tasks thus may reflect development in a different area: less reliance on prompt language. That is, some of these phrasal structures decrease at Time 2 because learners are not using prompt language as frequently.

To investigate this further, consider PPs as abstract postmodifiers. Their frequency also decreases over time in integrated writing. At Time 1, many of these structures represent prompt language (or near-prompt language):

theories about X
landmarks like rivers, coastlines, and mountains
celestial objects like the sun or stars
in reference to celestial objects
in relation to the North Star
the pull on the crystals

At Time 2, these examples also occur, but they begin to have variations. For example, instead of only *theories about*, a range of other nouns are used with *about* as a postmodifier: *the problem about*, *debatable aspects about*, *the limitation about*, *an example about*, *the passage about*, *the reading material about*, *a different view about*. Other attested novel abstract PPs as postmodifiers at Time 2 include *possible explanations on*, *an argument on*, *their views on*, *three assumptions on*, *different evidence regarding*, *the distance to the destination*, *the route to the destination*, *the days without celestial objects*. These sequences are not frequent, nor are they always accurate, but they represent an emergence of creativity largely not seen at Time 1 in integrated writing.

In independent writing, the largest increases are for nominalizations and abstract PPs as postmodifiers. While nouns with multiple premodifiers showed a slight increase, this feature is overall very rare in independent writing. This may be because the task type does not elicit this sort of language or because the learners represented here are not yet to the stage of producing these structures independently without prompt language. Nominalizations, however, are more frequent. At Time 1, the learners represented here utilize 41 distinct nominalizations, none of which appeared in the prompt language for the independent response. At Time 2, nominalizations became more diverse, with 82 distinct forms. Abstract PPs as postmodifiers span a range of prepositions (*about*, *after*, *against*, *among*, *around*, *between*, *except*, *for*, *from*, *in*, *like*, *on*, *such as*, *to*, *with*, *without*) at both Time 1 and Time 2.

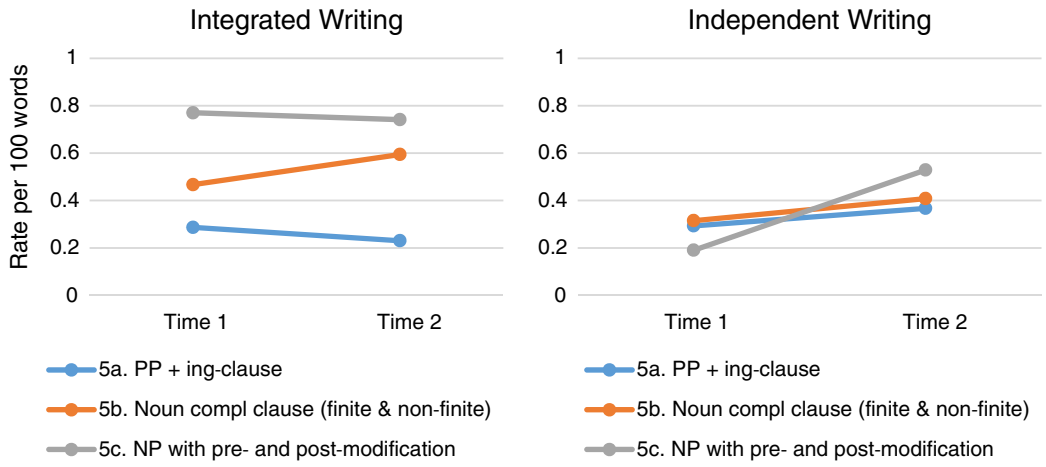

Figure 11 Mean rates of occurrence for Stage 5 features in writing.

Stage 5 features (Figure 11) are relatively infrequent in these corpora; this is not unexpected, as Stage 5 features are characteristic of highly informational texts written for a specialized audience, such as academic research articles (Biber et al., 2011; Biber & Gray, 2010). However, Stage 5 features do increase significantly in independent writing. While all three features increase in frequency, the increase for noun phrases with both premodification and postmodification is noteworthy, as it reflects multiple meaning expansions on the same noun phrase.

To illustrate this type of development, Examples 21 and 22 come from a test taker who produced no noun phrases with both premodification (*italicized*) and postmodification (underlined) at Time 1 (head nouns with modification are **bold**). At Time 2, the writer's language contains not only more frequent noun modification but also nouns with both premodifiers and postmodifiers (*increasing number of barriers*, *people's way to success*, and *enough knowledge regarding our future job*):

(21) *Independent Writing, Time 1, Score 1 (A080)*
Apparently, many people always consider the realities and they argue that choosing the subjects to prepare for a job or career is the *proper* and *fittest* **way**. But there is also many **people** prefer interests to facts. In effect, I am just a **supporter** of the latter.

(22) *Independent Writing, Time 2, Score 3 (A080)*
What is more, nowadays, there are an *increasing* **number** of barriers on *people's* **way** to success, such as *technology* **development** and *economic* **revolution**. So, we have no alternative but to make preparation as soon as possible to keep in steps with the times and for us students, the only **thing** we can do is to accumulate *enough* **knowledge** regarding our *future* **job**. Besides, students can develop interests during their *leisure* **time** to relax themselves from *busy studying* **life**.

Development of Complexity in Speaking

The register basis of the developmental complexity framework leads to different hypotheses for speech. Because clausal complexity is predominant in spoken registers, stages of the framework with clausal structures are of primary interest for spoken TOEFL iBT responses. Thus, the expected patterns of development include moving from clausal structures with very common lexico-grammatical patterns to wider sets of lexis–grammar combinations and from finite to nonfinite clausal constructions. Again, however, these hypotheses are mediated by task type. As seen with the MD analysis, it is expected that the informational purpose of integrated tasks (as well as the literate input from reading passages) will lead to the use of phrasal complexity features. However, without the planning and editing available in the written mode, these features are expected to be less frequent than in writing.

Tables 11 and 12 present the mean rates of occurrence for the complexity features grouped by stage at Time 1 and Time 2 in integrated and independent speaking, respectively. The most surprising finding from these tables is that no significance was found for any of the stages in either speaking task type.

However, a visual inspection of the means for these features across Time 1 and Time 2 (Appendix F; graphical displays also available upon request from the corresponding author) reveals a few noteworthy changes in the frequencies

Table 11 Change in Use of Complexity Features (Per 100 Words) by Stage Over Time in Integrated Speaking

	Time 1		Time 2					
Stage	*M*	*SD*	*M*	*SD*	Δ[a]	*t*	*p*	Cohen's *d*
Stage 1	1.06	0.73	1.01	0.59	(−)	0.361	0.720	0.056
Stage 2	7.84	1.89	7.70	1.83	(−)	0.360	0.721	0.056
Stage 3	6.57	2.15	7.38	1.66	(+)	2.136	0.039	0.330
Stage 4	3.23	1.35	3.39	1.34	(+)	0.534	0.596	0.082
Stage 5	0.34	0.39	0.49	0.35	(+)	2.244	0.030	0.346

[a]Direction of the change, with parentheses indicating a nonsignificant change. *$p < .01$.

Table 12 Change in Use of Complexity Features (Per 100 Words) by Stage Over Time in Independent Speaking

	Time 1		Time 2					
Stage	*M*	*SD*	*M*	*SD*	Δ[a]	*t*	*p*	Cohen's *d*
Stage 1	1.20	0.84	0.83	0.57	(−)	2.093	0.043	0.323
Stage 2	6.34	1.61	6.93	2.13	(+)	1.300	0.201	0.201
Stage 3	7.11	2.09	7.44	1.98	(+)	0.739	0.464	0.114
Stage 4	1.61	1.05	2.04	0.99	(+)	2.09	0.043	0.322
Stage 5	1.01	0.81	0.93	0.74	(−)	0.465	0.644	0.072

[a]Direction of the change, with parentheses indicating a nonsignificant change. *$p < .01$.

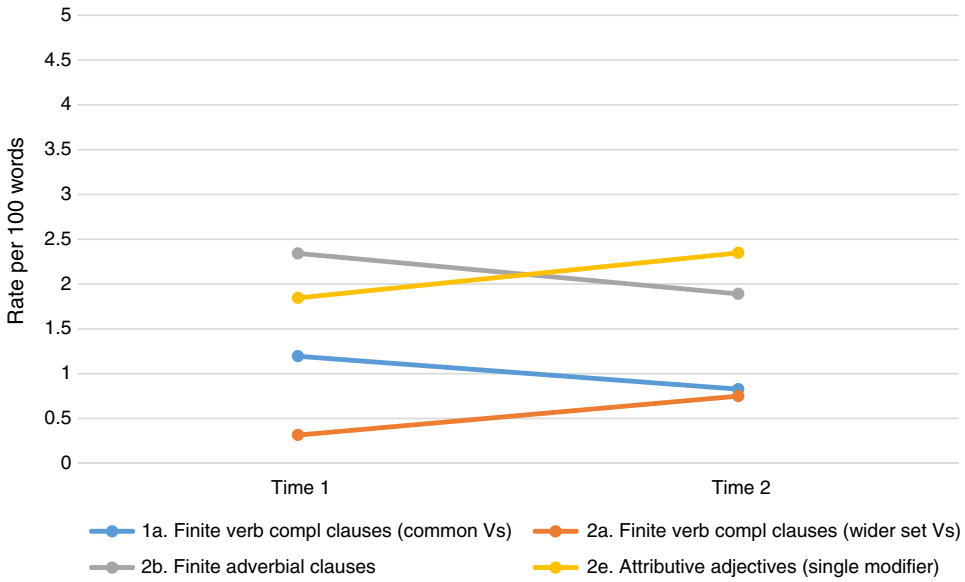

Figure 12 Select complexity features in independent speaking.

of individual complexity features and several differences in the rates of occurrence between integrated and independent speaking. Figure 12 displays the change over time for select features in independent speaking. Two features from the lowest levels of the framework, finite complement clauses controlled by common verbs and finite adverbial clauses, decrease slightly. In contrast, finite complement clauses with a wider set of verbs increase. While at first this appears to support the hypothesized path, it turns out that at Time 2, nearly all instances of these finite verb complements are structures with *be*:

The first reason is that …
The second reason is that …

In fact, 64% of all instances of Stage 2a ($N = 50$) at Time 2 are made up of this structure. The pattern *reason is that* may be a formulaic chunk learned over the course of the study.

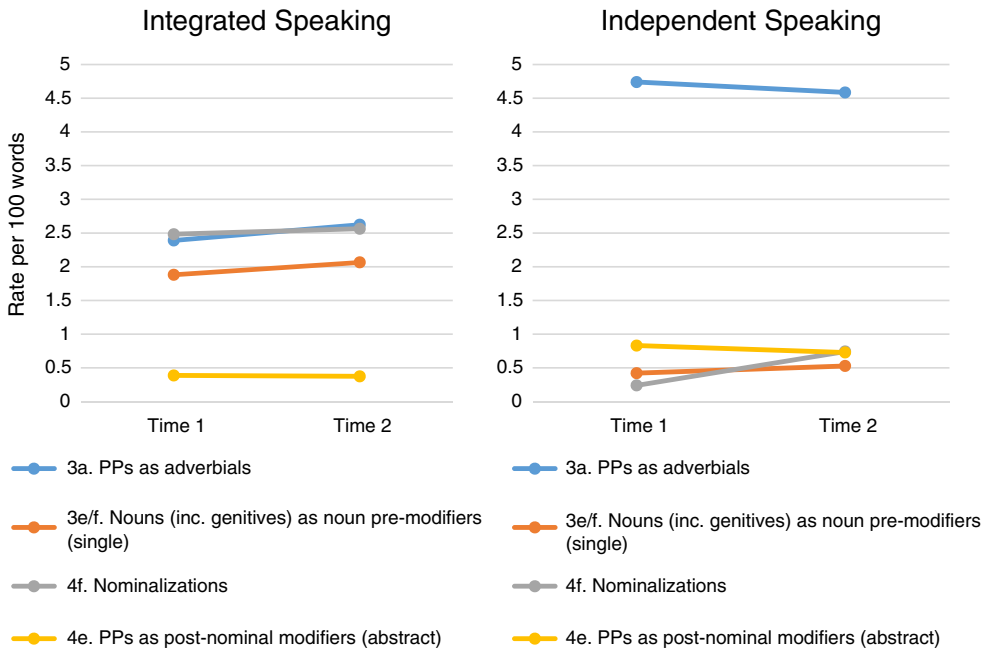

Figure 13 Select complexity features in integrated and independent speaking.

Attributive adjectives also increase from Time 1 to Time 2. However, analysis of the adjectives used reveals that at Time 1, 43% of these adjectives were accounted for by the two phrases *best way* and *good way*, which are primed by the prompt, as it asked about the best way for students to relax. At Time 2, the proportion of attributive adjectives accounted for by these two phrases dropped to 25%; these phrases are still quite common, but as the frequency has increased, it has also resulted in an expanded use of a range of attributive adjectives, such as *beautiful scenery*, *best friend*, *excellent behavior*, *favorite movies*, *fresh air*, *good habit*, *good health*, *outdoor activities*, *personal opinion*, *physical health*, *regular schedule*, and *spare time*. Thus, a meaning and functional expansion accompanied the increase in frequency in independent speaking.

The final set of features to examine here are features that seem to be differentiated, not by frequency changes over time, but by their relative frequencies across integrated and independent task types in speaking. Figure 13 displays four features across the two subcorpora: PPs as adverbials, nouns as premodifiers, nominalizations, and abstract PPs as post-modifiers. PPs as adverbials are more frequent in independent speaking than in integrated speaking. In part, this higher frequency may be due again to prompt effects, as many instances of adverbial PPs deal with the topic of going to bed (e.g., *at night*, *in the morning/evening*, *to bed*). A few, however, are task specific, helping to orient the point of view of the speakers while they provide their opinions in the dependent task (*for me*, *in my opinion*) or organize discourse (*for example*).

In contrast, two features are more common in integrated tasks than in independent tasks: nouns as premodifiers and nominalizations. The higher frequency of these features can be attributed to the informational purpose of integrated tasks, as both represent informational features. It should be noted that most of the nominalizations used in integrated speaking were used on the prompts: *ability*, *assignment*, *competition*, *exception*, *reference*, *reaction*, *reciprocity*, and *solution*.

Likewise, many of the nouns as premodifiers at both times were adopted (e.g., *dust mite*, *midterm exam*, *reference section*, *swim competition*, *swim team*, *wood floor*, *school library*) or adapted (e.g., *woman professor*, *man professor*, *listening material*, *reading passage*). Although there does not appear to be a functional expansion as seen with other phrasal features, students increase their use of this feature over time.

Discussion

The developmental complexity analysis has revealed distinctive developmental patterns for different modes and task types, with the most evidence for a developmental trajectory occurring in independent writing. Table 13 summarizes the significant changes in each of the complexity stages in all four corpora. However, as the preceding qualitative analyses showed,

Table 13 Summary of Complexity Stage Changes in the Longitudinal Corpus

Mode/task	TOEFL score	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Written integrated	+		++		−	
Written independent	++	−−	++	++	++	+
Spoken integrated	+					
Spoken independent	+++					

Note. For significant results only; + indicates an increased dimension score, while − indicates a decreased dimension score. +++ or −−− effect size > .8 (large). ++ or −− effect size > .5 (medium). + or − effect size > .2 (small).

these quantitative trends (both increases in the use of features and stable uses of features) must be interpreted relative to the functional interpretations of the language patterns observed as they occur in the texts.

Independent writing responses showed the clearest pattern of development, with significant increases in Stages 2–4 (and a decrease in Stage 1, which we would expect writers to move away from as proficiency increases). Within these stages, attributive adjectives represent one of the most important changes; not only did this informational feature increase in frequency, it exhibited an expansion in meaning and complexity in independent writing. Given the relatively low proficiency of the students in this corpus, these developments are expected for this phrasal feature in earlier stages (e.g., Stage 2a).

Two qualitative changes also demonstrated the need for analyses of developmental complexity to investigate more than just frequency changes in the use of features. Two features that increased seemed to be learned strategies, based on the consistency with which these patterns occurred: the increased use of adverbs as linking adverbials in writing and the use of the construction *the reason is that …*

It should be noted here that the complexity analysis did not function particularly well in capturing complexity in spoken tasks. However, there are several potential reasons for this finding. One possibility is that the speaking responses in the longitudinal corpus did not develop in terms of complexity. However, given several significant findings for speaking tasks in the MD analysis, this seems unlikely to account for these results in full. Another possibility is that the developmental complexity framework is not well suited to capturing linguistic development in speech (which could be explained by the fact that the hypothesized sequence was originally motivated by a desire to better capture development in writing). For example, while Stages 1 and 2 do focus predominantly on clausal structures or phrasal structures that are clause constituents, the remaining stages have a heavy focus on phrasal complexity features. Yet a third possibility is that the framework is not well suited to capturing short-term developments in speech or developments in speakers with relatively low proficiency levels and/or a narrow range of levels. These possibilities require further exploration in larger longitudinal speaking corpora.

Summary and Conclusion

The MD analysis and developmental complexity analysis revealed complementary results. For this group of learners, it appears that independent tasks, and independent writing in particular, underwent the most substantial development. Integrated writing tasks showed the least development and exhibited some patterns not consistent with expectations. However, it is possible that the presence of the source texts constrains students' creativity in adopting new features in a novel way. The substantial reliance on prompt language may inhibit creativity, whereas independent tasks promote it, because students must come up with language on their own. However, it iss also possible that the mixed results for integrated writing reflect the fact that integrated writing is a more difficult task. That is, at earlier stages, students rely more on the prompt language, thus using many of the structures studied here because they are used in the prompts or input. Then, as students develop over time and begin relying less on the prompt language, they compensate for the increased cognitive demands of the integrated writing task by using less complex grammatical structures (resulting in lower frequencies for these features). However, this hypothesis requires further exploration.

The first RQ asked to what extent patterns of variation observable in the level-stratified TOEFL corpus are also attested in the longitudinal corpus. The results here are relatively clear. The MD analysis results are remarkably consistent: Speakers

and writers often started out at Time 1 with discourse styles further away from the expected mode- and task type-specific discourse styles, and generally followed development paths that brought them more in line with similar scoring responses in the level-stratified corpus. Furthermore, responses in the longitudinal corpus, especially by Time 2, maintained those mode- and task type-specific patterns of use (relative to other modes and task types).

A major implication of these findings is that MD analysis as a methodology may be particularly well suited to capturing the more nuanced language development that occurs over relatively short periods of time, which has long been a major challenge for applied linguists working in the areas of language assessment, but also for those investigating the effects of particular teaching methods, evaluating the effectiveness of training programs and courses, or even studying the effects of immersive language learning experiences like study abroad. The utility of the MD analysis approach in capturing these short-term changes in learners' speaking and writing may be attributed to the multivariate nature of the method — the fact that it focuses on the linguistic co-occurrence of multiple features rather than any individual linguistic feature (a benefit also reported by Biber et al., 2016). Because the methodology uncovers underlying patterns within discourse structure, it has the potential to capture incremental changes in discourse style and language production, yet the method has not been widely adopted in research focused on short-term language development.

At the same time, the results for RQ1 also raise an important question that this study has not been able to answer. Across dimensions, learners produced language that was highly divergent from expectations at Time 1 but that quickly began to converge into observable and expected norms by Time 2. Because all the students represented in the longitudinal corpus were enrolled in EAP/TOEFL preparation courses, the current study could not tease apart whether the changes in their linguistic production were due to true language development and increased proficiency, or whether those developments might be attributed to increased knowledge of the test, question prompts, and test-taking strategies, or the effects of task repetition (because the same test form was used at both administrations). Further longitudinal research is needed to tease apart the potential effects of these factors, for example, by comparing development over the same period of time by learners in different learning contexts and receiving different types of language instruction. It would also be possible to pair corpus-based analyses with more qualitative methods that are able to tap into students' perceptions of their writing and speaking abilities, the learning that resulted from educational opportunities, and how they approached the testing situation.

The second RQ asked to what extent the phrasal and clausal complexity features changed over time and whether they were mediated by mode and task type. The complexity analysis revealed more moderate results in terms of development, but it clearly showed that what developments did occur were mediated by mode and task type, with particularly strong patterns in independent writing. In fact, the results for independent tasks in general, and in writing specifically, seem to indicate that development may occur first in independent tasks.

In general, these findings support the use of multiple task types in language assessments, as multiple tasks may be required to observe language ability and L2 development for learners at different stages. In addition, the tasks clearly elicit language that varies in systematic ways regardless of language proficiency (i.e., mode and task type variation was observable even within this group of learners, who represented a fairly low baseline proficiency level).

The third RQ on patterns in the developmental paths is harder to answer based on this complexity analysis data. As there were relatively few significant changes in the use of the various stages (especially in the speaking corpora), individual quantitative and functional shifts can be explained (as presented earlier). However, it does not seem that the amount of evidence is sufficient to make claims about developmental paths at this time. In particular, the patterns of development may be muted in these data due to the low proficiency level of the group of learners as well as their relative homogeneity.

However, an important finding from the developmental complexity analysis has been the further evidence that corpus-based analyses of language development cannot focus on quantitative patterns of use alone. On one hand, increases in the frequencies of particular features may be attributed to increased accuracy in those features, or such increases may occur because learners begin to expand those features functionally (i.e., using them for a wider range of meanings and functions). On the other hand, a lack of change in frequency of use also does not necessarily reflect a lack of development: The quantitative use of particular features may remain stable, with development occurring in the variability and range of meanings, functions, or lexico-grammatical associations. Several features in the developmental complexity analysis exhibited patterns such as these. It is important to note that the register/functional approach to complexity analysis offers one methodology for capturing these more qualitative and functional language developments, as it enables the

consideration of the specific constructions that contribute to complexity in terms of frequency, function, and meaning. Thus qualitative discourse analysis should be a key component of corpus-based research aimed at describing language development.

Returning to the two warrants that function as part of the explanation inference, this study sought to provide evidence toward the warrant that linguistic knowledge varies in expected ways, particularly with respect to mode and task type. The MD analysis, in particular, seemed to support this warrant, as most changes reflected developments toward the expected norms. In addition, most significant changes in the developmental complexity analysis were also in the hypothesized direction. In particular, the importance of mode and task type as fundamental parameters underlying variation points to the need for language assessment validation research to consider theoretical expectations specific to those parameters, rather than having a primary focus on general language proficiency.

With respect to the second warrant, that time and experience learning English are related to variations in performance, has also received partial support. Importantly, the evidence for this warrant is strengthened when considering the functional expansions and shifts that were observed in particular subcorpora, recognizing that development is not only about frequency or accuracy. It was not possible to link these changes statistically to scores, but this type of analysis will be particularly important for more fully exploring these warrants and the inferences to which they contribute. However, further research is needed to explore this warrant more fully. For example, research spanning a longer period of development may reveal more pronounced patterns of development. In addition, controlling more closely for type of experience in learning English may provide evidence for whether linguistic changes are the result of educational content and experiences or simply time.

Several limitations remain for the present study. The primary limitation of the study is that it was not possible to statistically link TOEFL scores with the linguistic changes, based on advice from a statistics consultant. Several characteristics of the data resulted in these issues, including the amount of missing data due to technical difficulties with audio and very short responses that could not be analyzed using corpus-based methods. In addition, as Ling et al. (2014) noted, the results of a study based on this longitudinal data set must be interpreted "in the context of the students' English proficiency level" (p. 14). The students included in this study had low or intermediate English-language proficiency. The relatively low proficiency of the learners represented in the longitudinal corpus, along with the lack of score variation within the group, may be impacting the ability to see the range of development that is possible in TOEFL iBT responses. Many features were quite infrequent in the corpora used here, which makes it harder to see the range of possible variability.

Two other limitations inherent to this data set warrant mentioning. Because students took the same form of the TOEFL at both test administrations, and all responses are based on a single form, it is possible that there is an effect of repeating the test. In addition, it is recommended that more than two data points be collected for longitudinal research designs. A replication of this study on a larger, more diverse longitudinal corpus (including additional test forms, learners from a range of contexts and with a broader range of proficiency levels) may be able to explore the developmental paths more fully (including through inferential statistical analyses like cluster analysis), either to validate the findings of the present study or to uncover additional developments not observed here. These types of statistical procedures were not possible with the current corpus due to power issues and the number of linguistic variables being investigated.

A further consideration is overlap between the various analyses, including in future analyses linking TOEFL scores to linguistic patterns in the use of grammatical complexity features. For example, the MD analysis and complexity analysis presented here are not completely independent. Indeed, Dimension 1 included many of the complexity features included in the developmental framework (often in more general terms). However, it did not include all of these features, and it additionally captured features not included in the complexity framework but which were nonetheless important for explaining the variation in language use that occurs across modes and task types in TOEFL iBT responses. In addition, it is possible that e-rater may be attending to some of the same features being analyzed in either the MD or the complexity analysis. Thus, statistical links between those scores and the linguistic analyses must be interpreted with caution.

Notes

1 Several different bases have been proposed depending on the register being investigated, such as *T-units*, *CS-units*, and *AS-units*. We use the term *T-unit* to refer to all of these measures.

2 Available at https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

3 If restricted to test takers who have analyzable responses for both task types and both modes at each test administration, the sample size reduces to 29. To maximize the data for these analyses, we included test takers who had analyzable responses across task types and test administrations in at least one mode. This necessitates analyzing spoken and written responses separately rather than jointly.

4 Twenty speaking files and 26 writing files were used for this analysis; to ensure that the program worked equally well on a range of texts types, six files containing 500-word samples from research articles in a range of disciplines were also included in this analysis.

5 Precision was calculated to ensure that all instances tagged automatically indeed represented instances of that feature. Recall was not calculated for the preprocessing scripts, because any instance missed by the preprocessing scripts would be manually tagged in subsequent steps.

6 The comparison between the longitudinal corpus and Biber and Gray's (2013) findings differ in that the former is presented as one group, while the latter are divided into score bands. This is because there were not enough observations or enough score variation in the longitudinal sample to report the data in the exact same way. However, this comparison is useful, as it demonstrates that the development did occur over the 9 months, and that this development is what we would expect given the mean test scores for the learners represented in the longitudinal corpus.

7 In other words, the developmental component occurs when the extent to which learners use or rely on those features changes. That is, at earlier levels, we would expect language learners to produce verb + that-complement clauses with only the most common verbs (Stage 1), and as they develop, they will still use those common verbs, but less frequently, as they increase in the number of other, less common verbs in this structure. Because our concern is in the changing frequencies of features, we are interested in seeing if we observe those decreases in frequency of the features associated with earlier stages in the progression and increases in the frequency of features associated with later stages in the hypothesized progression. This is why we compare the use of features within each stage over time.

References

Alderson, J. (2000). Testing in EAP: Progress? Achievement? Proficiency? In G. M. Blue, J. Milton, & J. Saville (Eds.), *Assessing English for academic purposes* (pp. 21–47). Bern, Switzerland: Peter Lang.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V. 2.0. *Journal of Technology, Learning, and Assessment*, *4*(3), 13–18.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, *5*, 257–269. https://doi.org/10.1093/llc/5.4.257

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, *15*, 133–163. https://doi.org/10.1080/01638539209544806

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*, 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*, 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis* (Research Report No. RR-13-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*, 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*, 639–668. https://doi.org/10.1093/applin/amu059

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

Biber, D., & Jones, J. K. (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.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research*, *3*, 185–214. https://doi.org/10.1177/136216889900300302

Bygate, M. (2001). Effect of task repetition on the structure and control oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning teaching and testing* (pp. 23–48). Harlow, England: Pearson Education.

Byrnes, H., Maxim, H., & Norris, J. (2010). Realizing advanced L2 writing development in a collegiate curriculum: Curricular design, pedagogy, assessment. *Modern Language Journal*, *94*(Monogr), 1–221. https://doi.org/10.1111/j.1540-4781.2010.01148.x

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

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, *26*, 1–9. https://doi.org/10.1016/j.jslw.2014.09.002

Conrad, S., & Biber, D., (2001). Multi-dimensional methodology and the dimensions of register variation in English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 13–42). Harlow, England: Pearson.

Crossley, S., & McNamara, D. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.002

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*, 5–32. https://doi.org/10.1016/j.asw.2005.02.001

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R. N., Mollaun, P., Nissan, S., … Schell, M. (2008). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a Foreign Language* (pp. 97–143). New York, NY: Routledge.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing*, *27*, 317–334. https://doi.org/10.1177/0265532210363144

Enright, M., & Tyson, E. (2008). *Validity evidence supporting the interpretation and use of TOEFL iBT scores* (TOEFL iBT Research Insight, Series I v4). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277–297). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/lllt.32.12fer

Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing*, *26*, 80–95. https://doi.org/10.1016/j.jslw.2014.09.002

Galaczi, E.D. (2013). Content analysis. In A.J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1323–1339). New York: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla063

Gray, B. (2019). Tagging and counting linguistic features for multi-dimensional analysis. In T. Berber-Sardinha & M. Veirano (Eds.), *Multidimensional analysis* (pp. 43–66). New York, NY: Continuum/Bloomsbury.

Gunnarsson, C. (2012). The development of complexity, accuracy and fluency in the written production of L2 French. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 247–276). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/lllt.32.11gun

Halliday, M. A. K., & Martin, J. R. (1996). *Writing science: Literacy and discursive power*. London, England: Falmer Press. (Work first published 1993)

Hunt, K. (1966). Recent measures in syntactic development. *Elementary English*, *43*, 732–739.

Jamieson, J., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a Foreign Language* (pp. 55–95). New York, NY: Routledge.

LaFlair, G., & Staples, S. (2017). Using corpus linguistics examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*, 451–475. https://doi.org/10.1177/0265532217713951

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*, 607–614. https://doi.org/10.1093/applin/amu047

Ling, G., Powers, D. E., & Adler, R. M. (2014). *Do TOEFL iBT scores reflect improvement in English-language proficiency? Extending the TOEFL iBT validity argument* (Research Report No. RR-14-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12007

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writer's language development. *TESOL Quarterly*, *45*, 36–61. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*, 493–511. https://doi.org/10.1177/0265532217710675

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*, 555–578. https://doi.org/10.1093/applin/amp044

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518. https://doi.org/10.1093/applin/24.4.492

Ortega, L., & Byrnes, H. (2008). *The longitudinal study of advanced L2 capacities*. New York, NY: Routledge.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, 26, 26–45. https://doi.org/10.1017/S0267190505000024, *25*

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, *14*, 48–59. https://doi.org/10.1016/j.jeap.2013.12.001

Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Polat, B., & Kim, Y. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics*, *35*, 184–207. https://doi.org/10.1093/applin/amt013

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of e-rater scoring engine* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02158.x

Skehan, P., Foster, P., & Mehnert, U. (1998). Assessing and using tasks. In W. Renandya & G. Jacobs (Eds.), *Learners and language learning* (pp. 227–248). Singapore: SEAMEO Regional Language Center. https://doi.org/10.1037/e614082009-001

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*, *31*, 532–553. https://doi.org/10.1093/applin/amq001

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*, 149–183. https://doi.org/10.1177/0741088316631527

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*, 420–430. https://doi.org/10.1002/tesq.91

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*, 90–111. https://doi.org/10.1093/applin/amv002

Wells, R. (1960). Nominal and verbal style. In T. A. Sebeok (Ed.), *Style in language* (pp. 213–220). Cambridge, MA: MIT Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawai'i Press.

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing*, *34*, 565–577. https://doi.org/10.1177/0265532217720956

Yoon, H. (2018). The development of ESL writing quality and lexical proficiency: Suggestions for assessing writing achievement. *Language Assessment Quarterly*, *15*, 387–405. https://doi.org/10.1080/15434303.2018.1536756

Yoon, H., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, *51*, 275–301. https://doi.org/10.1002/tesq.296

Appendix A

Transformation of Scores for Individual Written TOEFL iBT Responses

Following Biber and Gray (2013, p. 12), the following score transformations were applied to each written response:

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

Appendix B

Annotation of Prepositional Phrases

Table B1 lists tags assigned by the preprocessing scripts and indicates the contexts (i.e., operational definitions) in which tagging was automated. Operational definitions in Table B1 reflect only those contexts that could be automated reliably; other operational definitions are possible and applied at the hand-coding stage (see Table B2). Prepositions not matching the contexts described in Table B1 were tagged *pp?* to be manually coded.

Table B1 Operational Definitions and Tags Used in Prepositional Phrase Preprocessing Scripts (Automatic Tags)

Tag	Description and operational definition
ppadvl-3a	*Preposition functioning as an adverbial* Specific combinations of words in which the PP is always adverbial in nature (based on common words/phrases observed in the TOEFL iBT Longitudinal corpus during testing and program development): *for example*, *prepare for*, *in reference to*, *because of*
ppnof-3g	*Preposition* of *functioning as postnominal modifier* Noun followed by preposition *of* (including instances of three-word complex prepositions in which the second preposition is *of*). Excludes instances of the following: multiword determiners (e.g., *a lot of*) N + of + *ing*-clause patterns (these are included under ppning-5a) instances of two-word complex prepositions with *of*
ppna-4e	*Prepositional phrases functioning as noun postmodifiers with abstract meanings* Specific phrase *such as*, which is typically postnominal Three-word complex prepositions (Biber et al., 1999, p. 75) are analyzed compositionally. The second preposition in a three-word complex preposition is automatically tagged as a noun postmodifier when the second word is a noun (e.g., *in exchange for*, *in return for*) The first preposition in three-word complex prepositions is tagged pp? Combinations with *of* as the second preposition (e.g., *in light of*, *by way of*) are already captured by ppnof-3g.
ppning-5a	*Preposition before an* ing*-complement clause when functioning as a noun modifier/complement* Preposition occurring in the following pattern: Word tagged as noun + preposition + word tagged as nonfinite *ing*-form
ppxxing-5a	*Preposition before an* ing*-complement clause when not functioning as a noun modifier/complement* Preposition occurring in the following pattern: Word not tagged as noun + preposition + word tagged as nonfinite *ing*-form
ppxof-0x	*Preposition* of *in other functions (not postnominal modifiers, multiword verbs, or with nonfinite complement clauses)* *Of* occurring as part of multiword determiners: *a lot of*, *all of*, *some of*, *half of*, *both of*, *lots of*, *few of*, *many of* *Of* when it occurs in contexts not specified under tag ppadvl-3a, ppnof-3g, ppning-5a, and ppxing-5a
pmwv-0x	*Preposition as part of a multiword verb* Words tagged as prepositions when preceded by a verb and when the verb and preposition combination matches one of the common multiword verbs. The list of common multiword verbs was generated based on Biber et al. (1999, pp. 410–422) and analyzed by the research team to identify verb + preposition combinations that are typically multiword verbs when they occur together (with consultation to the longitudinal TOEFL iBT corpus). List of multiword verbs selected for automatic tagging available upon request. *Note*: Some multiword verbs are more variable, and the combination of a verb and preposition could occur either as a multiword verb or as a single lexical verb with an adverbial prepositional phrase. Instances of these verbs were not automatically tagged but rather could be tagged as either structure during the manual coding process.
ppxx-0x	*Other specialized uses of prepositions* Prepositions occurring in contexts not included in the framework (i.e., contexts other than adverbial, noun postmodifier, or multiword verb constructions). Automatic tags added for the following: the common combination *interested in* (i.e., an adjective complement) prepositions tagged as pied piping (e.g., *in which*) constructions the second preposition in three-word complex prepositions, when the second word is an adjective or adverb (i.e., *as far as*, *as well as*, *as distinct from*, *as opposed to*)
pp?	*Uncoded preposition* Tag assigned if none of the above contexts are met, indicating that the preposition should be manually annotated

Table B2 provides the operational definitions and coding notes used during the manual annotation for PPs, including all possible prepositional phrase tags.

Table B2 Operational Definitions and Coding Notes for the Manual Coding of Prepositional Phrases

Stage/tag	Feature and operational definitions
3A	Phrasal embedding in the clause: Prepositional phrases as adverbials
ppadvl-3a	Manual coding of prepositions when answering questions of *where*, *when*, *how*, or *why* an action occurred. These include the following typical contexts (not exhaustive): 1. PPs directly following a verb (but distinguished from multiword verbs) 2. PPs directly following a *be* verb 3. PPs indicating a causative meaning (e.g., *because of X*) 4. *by*-phrases indicating the agent in passives (e.g., *was taken by the student*) 5. PPs that function as linking adverbials (e.g., *on the other hand*, *in addition*)
3G	*Of*-phrases as noun postmodifiers
ppnof-3g	All instances tagged during preprocessing scripts; see Table B1
3H	Simple prepositional phrases as postmodifiers, especially with prepositions other than *of* when they have concrete/locative meaning
ppnc-3h	Prepositional phrase occurring after a noun, in which the prepositional phrase identifies the referent of the head noun or adds descriptive information about the head noun. PPs as postnominals can often be rephrased with a relative clause. Restricted to instances in which the PP carries a concrete or locative meaning, including textual location. Examples: *something in the dust* (cf. *something which is located in the dust*) *everyone around you* (cf. *everyone who is around you*) *the correct way to their home* (cf. *the correct way that leads to their home*) *the theory in the passage* (cf. *the theory that appears in the passage*)
4E	Simple prepositional phrases as noun postmodifiers, especially with prepositions other than *of* when they have abstract meanings
ppna-4e	Prepositional phrase occurring after a noun, in which the prepositional phrase identifies the referent of the head noun or adds descriptive information about the head noun. PPs as postnominals can often be rephrased with a relative clause. Restricted to instances in which the PP carries an abstract meaning. Many types of abstract meaning are possible; a few example meanings observed in the longitudinal iBT corpus include the following: Topic: *three theories about bird's navigational abilities*, *information about the job* Time: *the air in the morning*, *a period in the past* Other: *all the responses for them*, *a conclusion from an experiment*, *landmarks like rivers and coastlines*, *heated discussion among the students*, *a dilemma between a useful job and an interested job* Prepositions (especially *in*, *on*, and *to*) are coded as abstract if there is not a literal, locative meaning. Thus the following examples would be coded as abstract: *your major in the university*, *the pull on the crystals*, *influence on their future job*, *their way to success* Instances of *for* are tagged as abstract postnominals in the construction N + for NP + to-clause: *three ways for birds to navigate*
5A	Prepositions with nonfinite complement clauses
ppning-5a	Postnominal preposition + ing-complement clause constructions Instances of the pattern *noun + preposition + ing-clause*, where the function of the PP is postnominal: *the best way for insuring that* *research about choosing subjects* *effort in studying this subject*
ppxing-5a	Preposition + ing-complement clauses when not following a noun Instances of the pattern *preposition + ing-clause* when not functioning as a noun postmodifier: *… feels happy through teaching* *afraid of going into the class* *came into the human body by breathing* *the professor rebutted it by mentioning that …*
5D	Extensive phrasal embedding in the NP: Multiple prepositional phrases as postmodifiers, with levels of embedding **Not included in this analysis; this type of extensive embedding is not common in these types of texts or for speakers/writers with this level of proficiency.*

Table B2 Continued

Stage/tag	Feature and operational definitions
Other	Additional tags assigned during manual coding and/or preprocessing to exclude instances of prepositions from the developmental complexity analysis
ppamb-??	Prepositional phrase with an ambiguous function Prepositions that could be equally interpreted as adverbial or postnominal. While other contexts are possible, the following are some common contexts in which ambiguity occurs: 1. *be* + NP + PP: *Money is the most important thing in their lives* can be interpreted as when/in what circumstances (adverbial) money is important (cf. *In their lives*, *money is the most important thing*) or an identifier (postnominal) for *thing* (cf. *Money is the most important thing that exists in their lives*) 2. have + NP + PP: *Birds have the type of compass in the brain* can be interpreted as where (adverbial) birds have something (cf. *In the brain*, *birds have the type of compass*) or an identifier (postnominal) limiting the referent of the head noun to those in the brain (cf. *Birds have the type of compass which is located in the brain*) 3. Other verbs + NP + PP: *He has to earn money for a better life* can be interpreted as the reason (adverbial) to make money (cf. *For a better life*, *he has to earn money*) or to identify (postnominal) the referent of the noun (cf. *He has to earn money which is for a better life*). *The region plays an important role in memory formation* can be interpreted as where/in what circumstances (adverbial) the region plays a role (cf. *In memory formation*, *the region plays an important role*) or as an abstract identifier (postnominal) for *role* (cf. *The region plays an important role which contributes to memory formation*) 4. Existential *there* construction: *There are two solutions for her* can be interpreted as either an adverbial telling how the proposition applies to (cf. *For her*, *there are two solutions*) or as a postnominal that restricts the referent of *solutions* (cf. *There are two solutions which are for her*) *Note*: Some existential *there* constructions are less ambiguous and were thus coded accordingly: *There is a debate on the campus* is adverbial, since *on the campus* explains where the debate exists (and cannot be interpreted in this context to mean a debate which is about the campus). *There is a section in the brain that can help the birds remember* can be interpreted as postnominal, since it would be unusual for there to be an adverbial PP intervening between a head noun and another postnominal modifier (the relative clause *that can help …*)
ppxx-??	Indeterminate function Prepositional phrase with an indeterminate function, often due to errors by the speaker/writer: *I prefer to give her over test because …* *The to the stay up to students* *He is rellargic about in dust* *Joe has got an antibody fight to the invader*
pmwv-0x	Preposition as part of a multiword verb Words tagged as prepositions when preceded by a verb and when the verb and preposition combination matches one of the common multiword verbs listed in Biber et al. (1999, pp. 410–422). See Table B1 for combinations tagged automatically. All others evaluated during manual coding.
ppxof-0x	Preposition *of* in other functions (not postnominal modifiers, multiword verbs, or with nonfinite complement clauses) *Of* occurring as part of multiword determiners: *a lot of*, *all of*, *some of*, *half of*, *both of*, *lots of*, *few of*, *many of* *Of* when it occurs in contexts not specified under tag ppadvl-3a, ppnof-3g, ppning-5a, and ppxing-5a *Note*: All instances tagged during preprocessing scripts
ppxx-0x	Other specialized uses of prepositions Prepositions occurring in contexts not included in the framework (i.e., contexts other than adverbial, noun postmodifier, or multiword verb constructions). These include (but are not limited to) the following: Adjective complements (e.g., *better than hers*, *satisfied with his choice*, *people close to you*) Pied piping constructions (e.g., *in which*) Prepositions are part of hyphenated words (e.g., *make-up exam*) Repeated prepositions due to dysfluencies (e.g., *you give to*, *to somebody*) See also Table B1.

Appendix C
Interrater Reliability for Prepositional Phrase Coding

	Written subcorpus		Spoken subcorpus	
File batch	Agreement (%)	Cohen's κ	Agreement (%)	Cohen's κ
Set 1	79	0.631	78	0.608
Set 2	71	0.577	78	0.612
Set 3	75	0.586	77	0.622
Set 4	75	0.644	81	0.710
Set 5	73	0.600	75	0.628
Set 6	78	0.657	82	0.709
Set 7	80	0.677	86	0.757
Set 8	79	0.663	81	0.688
Set 9			79	0.635
Mean	76.25	0.629	79.67	0.663

Appendix D
Operational Definitions for Developmental Complexity Tagger

The complexity tagger is based on texts that have been tagged for POS by the Biber Tagger and have been subjected to automatic scripts to correct common errors. Additional preprocessing steps required for some features are indicated in the following table. The following information lists the possible tags and explains how each feature was operationally defined.

Stage/tag	Feature and operational definitions
1A	**Finite complement clauses (*that*, *wh-*) controlled by common verbs (e.g., *think*, *know*, *say*)**
	Preprocessing steps
	FixTagging (all instances of that)
vcmpth-1a	*That-complement clauses*
	Instances of *that* tagged as verb complement when preceded by very common verbs occurring in this structure (>100 time per million words; Biber et al., 1999, pp. 661–666). Occurring in the following patterns:
	V + *that*-clause: *believe*, *feel*, *find*, *guess*, *know*, *see*, *think*, *say*, *show*, *suggest*
	V + NP + *that*-clause, allowing up to three intervening words for NP: *show*
	V + to NP + clause, allowing up to two intervening words for NP: *say*, *suggest*
	Instances of these very common verbs tagged as containing a zero complementizer
vcmpwh-1a	*Wh-complement clauses*
	Instances of *wh*-words (*where*, *when*, *who*, *whom*, *which*, *why*, *whose*, *whatever*, *whoever*, *what*, *whichever*, *how*) following very common verbs occurring in this structure (>50 times per million words; Biber et al., 1999, pp. 685–686). Occurring in the following patterns:
	V + *wh*-clause: *tell*, *know*, *wonder*, *see*
	V + NP + *wh*-clause, allowing up to three intervening words for NP: *tell*
	Instances of *if* preceded by very common verbs (Biber et al., 1999, pp. 691–693): *know*, *see*, *wonder*
	Instances of *whether* preceded by very common verbs (p. 692): *know*
	Note: Patterns in which the *wh*-complementizer is followed by a *to*-clause (e.g., *I don't know where to put this*) are tagged as verb *to*-complement clauses (see Biber et al., 1999, p. 685)
2A	**Finite complement clauses (*that*, *wh-*) controlled by a wider set of verbs**
	Preprocessing steps
	FixTagging (all instances of *that*)
vcmpth-2a	*That-complement clauses*
	Instances of *that* tagged as verb complement when not preceded by one of the very common verbs listed in 1A (Biber et al., 1999, pp. 685–686). Occurring in the following patterns:

Stage/tag	Feature and operational definitions
	V + that-clause V + NP + that clause V + to NP + clause Instances of verbs tagged as containing a zero complementizer (excluding the verbs from 1A)
vcmpwh-2a	*Wh-complement clauses* Instances of *wh*-words (*where*, *when*, *who*, *whom*, *which*, *why*, *whose*, *whatever*, *whoever*, *what*, *whichever*, *how*) following other common verbs occurring in this structure (>20 times per million words and "other attested verbs"; Biber et al., 1999, pp. 685–686) Occurring in the following patterns (lists of verbs for each patter available upon request): V + *wh*-clause V + NP + *wh*-clause (allowing up to three intervening words for NP) V + prep + *wh*-clause Instances of *if* preceded by other verbs (p. 693) Instances of *whether* preceded by other verbs (p. 692)
2B	**Finite adverbial clauses**
	Preprocessing steps FixTagging (all instances of *that*)
fadvl-2b	*Single- and multiword subordinators* Lexical, tag, and/or contextual matches for common circumstance adverbial subordinators (Biber et al., 1999, pp. 841–844) Lexical match (all occurrences): *although*, *because* (not followed by *of*), *unless*, *whenever*, *whereas*, *wherever* Lexical match when tagged as a subordinator: *for*, *once* Lexical match when (a) tagged as a subordinator and (b) not followed by an *ing*-form: *after*, *before*, *like*, *since*, (*even/as*) *though*, *until*, *while*, *whilst* Lexical match when (a) tagged as a subordinator and (b) not followed by an *ed*-clause: *as* Lexical match for words when they (a) do not meet criteria for wh-complement clauses (see 1A, 2A), (b) are not followed by an *ing*-form, and (c) are not preceded by an adjective or a preposition: *when*, *where*, *whatever* Lexical match for common multiword subordinators when *that* is tagged as a subordinator (not demonstrative): *except that*, *now that*, *so that*, *such that* Instances of *if* when not preceded by a common verb controlling *if*-clauses (see 1A, 2A) or common adjectives controlling if-clauses (*sure*, *unsure*, *clear*, *unclear*, *certain*, *uncertain*); instances of *as if*
2C	**Nonfinite (*to*-, *ing*-) complement clauses controlled by common verbs**
	Preprocessing steps FixTagging (present participles)
vcmpto-2c	*To-clauses* Instances of *to* tagged as an infinitive marker when preceded by very common verbs occurring in this structure (>100 time per million words; Biber et al., 1999, pp. 699–705). Occurring in the following patterns: V + *to*-clause: *attempt*, *begin*, *like*, *seem*, *tend*, *try*, *want* *Note*: Excludes bare infinitive clauses
vcmping-2c	*Ing-clauses* *Ing*-verb forms tagged as nonfinite when preceded by very common verbs occurring in this structure (>40 time per million words; Biber et al., 1999, pp. 740–741). Occurring in the following patterns: V + *ing*-clause: *begin*, *go* (*around/on*), *keep* (*on*), *start*, *stop* V + NP + *ing*-clause, allowing up to three intervening words for NP: *see*
2D	**Phrasal embedding in the clause: Adverbs as adverbials**
adv-2d	*Circumstance adverbials* Most common single-word circumstance adverbs (Biber et al., 1999, pp. 795–798); single words typically functioning as adverbials, when tagged as an adverb: *again*, *already*, *also*, *always*, *ever*, *Friday*, *here*, *Monday*, *never*, *now*, *often*, *Saturday*, *sometimes*, *still*, *Sunday*, *then*, *there*, *Thursday*, *today*, *Tuesday*, *usually*, *Wednesday*, *yesterday* Single words that can be adverbials or modifiers when followed by a verb (all other instances excluded): *even*, *just*, *only* *Note*: Single words that can be adverbials or modifiers that are excluded: *too*

Stage/tag	Feature and operational definitions
adv-2d	*Stance adverbials* Most common single-word stance adverbs (Biber et al., 1999, pp. 869–879) Single words typically functioning as adverbials, when tagged as an adverb: *actually*, *certainly*, *definitely*, *generally*, *maybe*, *perhaps*, *probably* Single words that can be adverbials or modifiers/other parts of speech, when followed by a verb (all other instances excluded): *really*, *totally* *Note*: Single words that can be adverbials or modifiers that are excluded: *like*
adv-2d	*Linking adverbials* Most common single-word linking adverbials (Biber et al., 1999, p. 887) Single words typically functioning as adverbials, when tagged as an adverb: *anyway*, *finally*, *first*(*ly*), *furthermore*, *hence*, *however*, *nevertheless*, *second*(*ly*), *then*, *therefore*, *third(ly)*, *though*, *thus*, *yet* *Note*: Single words that can be adverbials or modifiers/other parts of speech are excluded: *rather, so*
2E	**Simple phrasal embedding in the noun phrase: Attributive adjectives**
jatrb-2e	*Attributive adjective as nominal premodifier* Adjectives tagged "attributive" by the Biber Tagger only when 1. the following word is not tagged as an adverbial noun (e.g., … *feel very tired tomorrow*) 2. there are no additional adjectives, nouns, or genitive nouns in the phrase
3A	**Phrasal embedding in the clause: Prepositional phrases as adverbials**
	Preprocessing steps Prepositional phrase preprocessing scripts Manual coding
ppadvl-3a	*Prepositional phrases functioning as adverbials in the clause* See Appendix B.
3B	**Finite complement clauses controlled by adjectives**
	Preprocessing steps FixTagging (all instances of *that*)
jcmpth-3b	*That-complement clauses controlled by adjectives, simple (*i.e.*, nonextraposed)* Instances of *that* tagged as an adjective complement when no extraposed patterns (see 4B) are found
3C	**Nonfinite complement clauses controlled by a wider set of verbs**
	Preprocessing steps FixTagging (present participles)
vcmpto-3c	*To-clauses* Instances of *to* tagged as an infinitive marker when preceded by other verbs occurring in this structure (>20–50 times per million words, other attested verbs; Biber et al., 1999, pp. 700–705). Occurring in the following patterns (list of verbs included for each pattern available upon request): V + *to*-clause V + NP + to-clause, with up to two intervening words for NP Instances of *to* tagged as an infinitive marker when preceded by a *wh*-complementizer (*which*, *who*, *whom*, *whose*, *where*, *when*, *why*, *what*, *how*, *whether*) *Note*: Excludes bare infinitive clauses
vcmping-3c	*Ing-clauses* *Ing*-verb forms tagged as nonfinite when preceded by verbs occurring in this structure (excluding very common verbs in 2C; Biber et al., 1999, pp. 740–741). Occurring in the following patterns (list of verbs included for each pattern available upon request): V + *ing*-clause *be* + V*ed* + prep + *ing*-clause V + prep + *ing*-clause
3D	**Finite relative clauses**
	Preprocessing steps FixTagging (all instances of *that*)
finrel-3d	*Finite relative clauses with* that All instances of *that* tagged as "rel" *Note*: Does not count instances of zero relativizer
finrel-3d	*Finite relative clauses with wh-relative pronouns and adverbs* All instances of relative pronouns or determiners (*who*, *whom*, *whose*, *which*) tagged as "rel" Instances of relative adverbs (*where*, *when*, *why*) when preceded by common nouns heading relative clauses with adverbial gaps (Biber et al., 1999, p. 627–628):

Stage/tag	Feature and operational definitions
	nouns preceding relative adverb *where: area*, *bit*, *case*, *condition*, *country*, *example*, *hospital*, *house*, *place*, *point*, *room*, *situation*, *spot* nouns preceding relative adverb *when: bit*, *case*, *day*, *moment*, *occasion*, *period*, *season*, *time* nouns preceding relative adverb *why*: *reason* *Note*: Does not count instances of zero relativizer
3E	**Simple phrasal embedding in the noun phrase: Nouns as noun premodifiers**
npnm-3e	*Nouns as noun premodifier* Word tagged as a noun followed by another word tagged as a noun, excluding 1. when N2 is tagged as an adverbial noun, e.g., *our dreams today*) 2. when there are no additional adjectives, nouns, or genitive nouns in the phrase
3F	**Possessive nouns as noun premodifiers**
npnmgen-3f	Word tagged as a noun followed by a word tagged as a possessive marker (^$) followed by another noun (i.e., noun + possessive + noun). Excludes instances when there are additional adjectives, nouns, or genitive nouns in the phrase.
3G	***Of*-phrases as noun postmodifiers**
	Preprocessing steps Prepositional phrase preprocessing scripts
ppnof-3g	*Of*-phrases as noun postmodifiers See Appendix B.
3H	**Simple prepositional phrases as postmodifiers, especially with prepositions other than *of* when they have concrete/locative meaning**
	Preprocessing steps Prepositional phrase preprocessing scripts Manual coding
ppnc-3h	All prepositions (excluding *of*) as noun postmodifier, when the PP itself has a concrete and/or locative meaning See Appendix B.
3I	**Nominalizations**
	Preprocessing steps Manual analysis of word list of all words tagged as "noun" with specified derivational suffixes
nom [2nd tagfield]	*Derivationally derived nominalizations (from verbs and adjectives)* All words tagged as nouns that end in derivational suffixes (*age*, *al*, *an/ian*, *ance/ence*, *ant/ent*, *cy*, *dom*, *ee*, *er/or*, *ery/ry*, *ese*, *ess*, *ette*, *ful*, *hood*, *ician*, *Ie/y*, *ing*, *ism*, *ist*, *ite*, *ity*, *let*, *ment*, *ness*, *ship*, *tion*, *ure*) were extracted from the corpus and subjected to manual analysis to determine whether they would be considered nominalizations. Words that derived from verbs (e.g., *attraction* from *attract*) or adjectives (e.g., *ability* from *able*) were considered nominalizations. Words derived from other nouns (e.g., *friendship from friend*) were not considered nominalizations. The list of words occurring in the corpus considered nominalizations is available upon request. The complexity tagger added/retained the nominalization tag for these words and removed the nominalization tag if present on the original Biber tags for words not on this list. *Note*: This list includes several nonwords (e.g., *recopristy*). These are words used by the test takers that were not changed during the spell-checking process because (a) the intended form could not be determined (e.g., *prosaypothity*) and it was thus considered an error beyond spelling or (b) the word represented a grammar/morphology error rather than spelling (e.g., *unhappiness*). Thus these forms were retained because they represent grammatical errors.
4A	**Nonfinite complement clauses controlled by adjectives (simple)**
]jcmpto-4a	*To-complement clauses controlled by adjectives, simple* (i.e., *nonextraposed)* Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that typically occur only with nonextraposed, postpredicative *to*-clauses (Biber et al., 1999, pp. 718–721). List of adjectives included available upon request. Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that can occur with either postpredicate or extraposed complements, when not preceded by *it* and a form of *be/seem/become* (Biber et al., 1999, pp. 718–720; see patterns listed in 4B). List of adjectives included available upon request. *Note*: Does not include patterns adj + for NP + *to*-clause (e.g., *too difficult for them to remember*)

Stage/tag	Feature and operational definitions
4B	**Extraposed complement clauses**
	Preprocessing steps
	FixTagging (all instances of *that*)
jcmpxtra-4b	*Extraposed that-complement clauses controlled by adjectives*
	Instances of *that* tagged as an adjective complement and preceded by the set of adjectives that typically occur only with extraposed complements (Biber et al., 1999, p. 671–674). List of adjectives included available upon request.
	Instances of *that* tagged as an adjective complement and preceded by the set of adjectives that can occur with either postpredicate or extraposed complements, when preceded by *it* and a form of *be/seem/become*:
	it + (word) + (word) + *be/seem/become* + adj + *that* (e.g., *it is likely that*)
	it + (word) + *be/seem/become* + (word) + adj + *that* (e.g., *it seems highly unlikely that*, *it is not clear that*)
	Adjectives: *certain*, *good*, *important*, *impossible*, *likely*, *possible*, *right*, *sad*, *sensible*, *unlikely*
jcmpxtra-4b	*Extraposed to-complement clauses controlled by adjectives*
	Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that typically occur only with extraposed to-complements (Biber et al., 1999, pp. 618–621). List of adjectives included available upon request.
	Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that can occur with either postpredicate or extraposed complements, when not preceded by *it* and a form of *be/seem/become* (Biber et al., 1999, pp. 618–621):
	it + (word) + (word) + *be/seem/become* + adj + *to*
	it + (word) + *be/seem/become* + (word) + adj + *to*
	Adjectives: *awkward*, *bad*, *better*, *difficult*, *due*, *easier*, *easy*, *good*, *hard*, *harder*, *impossible*, *likely*, *nice*, *possible*, *right*, *smart*, *tough*, *unlikely*, *unwise*, *wise*, *wrong*
	Note: Does not capture pattern adj + for NP + *to*-clause (e.g., *it is difficult for them to choose*)
4C	**Nonfinite relative clauses**
	Preprocessing steps
	FixTagging (for past and present participle forms)
nfrel-4c	*Ing- and –ed clauses as postnominal modifiers*
	present participle forms tagged as postnominal modifier (tag "vwbg")
	past participle forms tagged as postnominal modifier (tag "vwbn")
	Note: to-clauses as postnominals are not included in this analysis
4D	**More phrasal embedding in the NP: Attributive adjectives and nouns as premodifiers (multiple modifiers)**
	Note: The tagger works by processing all features prior to this feature; after all features are tagged, the tagger processes the text again to look for instances of multiple modifiers. Tags are appended with "4d" when multiple modifiers are found, so that single modifiers can be counted under Features 2e, 3d, and 3f. To obtain a count for this feature (i.e., a noun phrase with multiple premodifiers), a tag is then added to the head noun so that there is one tag per noun phrase.
jatrb-2e4d	*Attributive adjectives occurring with other premodifiers*
	Adjectives tagged as *jatrb-2e* when they are preceded or followed by another word tagged as a noun premodifier (i.e., *jatrb-2e*, *npnm-3e*, *npnmgen-3f*)
npnm-3e4d	*Nouns as noun premodifier with other premodifiers*
	Nouns tagged as *npnm-3e* when they are preceded or followed by another word tagged as a noun premodifier (i.e., *jatrb-2e*, *npnm-3e*, *npnmgen-3f*)
npnmgen-3f4d	*Possessive nouns as noun premodifier with other premodifiers*
	Possessive nouns tagged as *npnmgen-3f* when preceded or followed by another word tagged as a noun premodifier (i.e., *jatrb-2e*, *npnm-3e*, *npnmgen-3f*)
hn-4d	*Head noun modified by multiple premodifiers*
	Head noun of phrase (i.e., a noun not tagged as a noun premodifier *npnm-3e*) when preceded by a word tagged as a multiple modifier (i.e., *jatrb-2e4d*, *npnm-3d4d*, *npnmgen-3f4d*)
4E	**Simple prepositional phrases as noun postmodifiers, especially with prepositions other than *of* when they have abstract meanings**
	Preprocessing steps
	Prepositional phrase preprocessing scripts
	Manual coding (see Appendix B)

Stage/tag	Feature and operational definitions
ppna-4e	*Prepositional phrases (other than of) as postnominal modifiers* See Appendix B.
5A	**Prepositions with nonfinite complement clauses** *Preprocessing steps* FixTagging (for present participle forms) Prepositional phrase preprocessing scripts Manual coding (see Appendix B)
ppning-5a	Postnominal preposition + ing-complement clause constructions Instances of the pattern *noun + preposition + ing-clause*, where the function of the PP is postnominal: *the best way for insuring that* *research about choosing subjects* *effort in studying this subject*
ppxing-5a	Preposition + ing-complement clauses when not following a noun Instances of the pattern *preposition + ing-clause* when not functioning as a noun postmodifier: … *feels happy through teaching* *afraid of going into the class* *came into the human body by breathing* *the professor rebutted it by mentioning that* …
5B	**Complement clauses controlled by nouns** *Preprocessing steps* FixTagging (all instances of *that*, present participles)
ncmpth-5b	*That-complement clauses controlled by nouns* Instances of *that* tagged as a noun complement
ncmpto-5b	*To-complement clauses controlled by nouns* Instances of *to* tagged as an infinitive marker and preceded by a common (>10 times per million words) and less common noun controlling *to*-complement clauses (Biber et al., 1999, p. 652). List of nouns included available upon request.
ncmping-5b	*Noun + of + ing-complement clauses* Instances of an *ing*- form tagged as nonfinite and preceded by common (>5 times per million words) nouns in the following pattern (Biber et al., 1999, pp. 653–654). List of nouns included available upon request. Pattern: noun + *of* + present participle
5C	**Appositive noun phrases** *Not included in this analysis*
5D	**Extensive phrasal embedding in the NP: Multiple prepositional phrases as postmodifiers, with levels of embedding** *Not included in this analysis*

Appendix E

Precision and Recall for Developmental Complexity Tagger

Table E1 contains precision and recall rates for features analyzed based on the 10% sample, while Table E2 presents those analyzed in the whole corpus. The data presented in the tables reflect final precision and recall rates. For example, some features had low precision rates, and hence all tags relevant to such features were manually analyzed to improve precision. Complete information about such cases and steps taken to improve the precision and recall rates is available from the corresponding author upon request. When remediation steps were taken, final accuracy rates are provided in parentheses.

Table E1 Precision and Recall Rates for Common Features (Based on 10% Sample)

	Written corpus			Spoken corpus		
Tag	*N*	Precision	Recall	*N*	Precision	Recall
vcmpth-1a	31	.97	.91 (1.00)	39	.95	.77 (0.97)
vcmpth-2a	30	.83 (1.00)	1.00	31	.94	.91 (1.00)
fadvl-2b	66	.97	.96	80	.94	.99
vcmpto-2c	14	1.00	.93	9	1.00	1.00
adv-2d	127	1.00	1.00	73	.99	1.00
jatrb-2e	105	.97	.95	102	.99	.97
vcmpto-3c	45	.96	1.00	36	.97	.92
finrel-3d	55	.95	.96	20	1.00	.95
npnm-3e	59	.92	.96	49	.94	.92
jatrb-2e4d	25	.96 (1.00)	.89	6	.67 (1.00)	IS
npnm-3e4d	17	.65 (1.00)	.92	12	.67 (1.00)	1.00
npnmgen-3f4d	15	1.00	1.00	3	.33 (1.00)	IS
hn-4d[d]	25	.92 (1.00)	.92	8	.88 (1.00)	.88
ncmpth/to/ing-5b	24	.92	1.00	10	1.00	.83

Note: IS, insufficient sample.

Table E2 Precision Rates for Less Common Features (Based on Whole Corpus)

	Written corpus			Spoken corpus		
Tag	*N*	Precision	Recall	*N*	Precision	Recall
vcmpwh-1a	24	1.00	n/a	4	1.00	n/a
vcmpwh-2a	31	1.00	n/a	34	.88[a] (1.00)	n/a
vcmping-2c	4	.75 (1.00)	n/a	42	.98	n/a
jcmpth-3b[b]	7	1.00	.78 (1.00)	7	1.00	1.00
vcmping-3c	15	.80 (1.00)	n/a	9	.89 (1.00)	n/a
npnmgen-3f	47	.89 (1.00)	n/a	86	.86 (1.00)	n/a
jcmpto-4a	82	.92 (1.00)	.85 (1.00)	31	.71 (1.00)	n/a
jcmpxtra-4b (with *that*)	8	.75 (1.00)	1.00	1	1.00	1.00
jcmpxtra-4b (with *to*)	75	.80 (1.00)	n/a	68	.88 (1.00)	n/a
nfrel-4c	69	.88 (1.00)	n/a	43	.88 (1.00)	n/a

Appendix F

Descriptive Statistics for Developmental Complexity Features

Table F1 Descriptive Statistics for Developmental Complexity Features in Written Tasks Over Time

	Time 1		Time 2	
Complexity features	*M*	*SD*	*M*	*SD*
Written integrated				
vcmpfin_1a_total	0.92	0.73	1.21	0.82
vcmpth-1a	0.78	0.66	1.05	0.73
vcmpwh-1a	0.15	0.28	0.16	0.26
vcmpfin_2a_total	1.53	0.79	1.68	0.74
vcmpth-2a	1.44	0.81	1.59	0.76
vcmpwh-2a	0.09	0.29	0.08	0.21
fadvl-2b	1.12	0.95	1.29	0.90
vcmpnf_2c_total	0.20	0.33	0.04	0.12
vcmping-2c	0.00	0.00	0.01	0.07
vcmpto-2c	0.20	0.33	0.02	0.11
adv-2d	2.55	1.72	4.16	1.31
jatrb-2e	3.17	1.01	3.22	1.00
ppadvl-3a	4.41	1.54	3.90	1.24
jcmpth-3b	0.02	0.08	0.00	0.00
vcmpnf_3c_total	0.26	0.50	0.28	0.36
vcmping-3c	0.00	0.00	0.00	0.00
vcmpto-3c	0.26	0.50	0.28	0.36
finrel-3d	1.03	0.70	0.93	0.53
npnm_total	2.17	1.18	2.32	0.97
npnm-3e	1.93	1.27	2.05	0.94
npnmgen-3f	0.24	0.38	0.26	0.46
ppnof-3g	1.58	0.85	1.16	0.70
ppnc-3h	0.18	0.36	0.34	0.45
jcmpto-4a	0.04	0.13	0.16	0.35
jcmpxtra-4b	0.01	0.08	0.04	0.13
nfrel-4c	0.28	0.34	0.27	0.44
jatrb-2e4d	0.76	0.64	0.59	0.55
npnm-3e4d	0.33	0.49	0.10	0.28
npnmgen-3f4d	0.96	0.75	0.46	0.44
hn-4d	0.99	0.74	0.57	0.48
ppna-4e	1.84	1.00	1.43	0.81
ppning_5a_total	0.29	0.46	0.23	0.40
ppning-5a	0.01	0.06	0.03	0.13
ppxing-5a	0.28	0.46	0.20	0.34
ncmp_5b_total	0.47	0.61	0.59	0.62
ncmping-5b	0.00	0.00	0.01	0.09
ncmpth-5b	0.16	0.37	0.43	0.54
ncmpto-5b	0.31	0.43	0.13	0.30
nom	2.78	1.21	2.18	1.06
ppnca_total	2.01	1.02	1.77	0.82
prepostN	0.77	0.53	0.74	0.50
ppxx-??	0.18	0.35	0.18	0.36
Written independent				
vcmpfin_1a_total	0.67	0.50	0.33	0.33
vcmpth-1a	0.67	0.50	0.33	0.33
vcmpwh-1a	0.01	0.05	0.01	0.04
vcmpfin_2a_total	0.50	0.41	0.39	0.38
vcmpth-2a	0.41	0.40	0.31	0.33
vcmpwh-2a	0.09	0.18	0.07	0.13
fadvl-2b	1.51	0.75	1.56	0.92
vcmpnf_2c_total	0.30	0.42	0.27	0.34
vcmping-2c	0.00	0.00	0.02	0.12

Table F1 Continued

	Time 1		Time 2	
Complexity features	*M*	*SD*	*M*	*SD*
vcmpto-2c	0.30	0.42	0.25	0.33
adv-2d	2.43	1.15	2.67	0.90
jatrb-2e	2.07	1.62	3.29	1.39
ppadvl-3a	3.47	1.27	3.96	1.22
jcmpth-3b	0.02	0.08	0.03	0.09
vcmpnf_3c_total	1.40	0.96	1.09	0.95
vcmping-3c	0.09	0.20	0.01	0.06
vcmpto-3c	1.30	0.93	1.08	0.94
finrel-3d	1.20	0.74	1.14	0.70
npnm_total	0.59	0.65	1.06	0.84
npnm-3e	0.56	0.61	1.01	0.84
npnmgen-3f	0.03	0.15	0.05	0.13
ppnof-3g	0.56	0.54	1.29	1.02
ppnc-3h	0.03	0.10	0.07	0.14
jcmpto-4a	0.30	0.38	0.25	0.27
jcmpxtra-4b	0.36	0.40	0.23	0.27
nfrel-4c	0.06	0.15	0.13	0.24
jatrb-2e4d	0.13	0.48	0.13	0.20
npnm-3e4d	0.09	0.28	0.14	0.33
npnmgen-3f4d	0.01	0.05	0.01	0.04
hn-4d	0.10	0.37	0.14	0.25
ppna-4e	0.37	0.45	0.69	0.65
ppning_5a_total	0.29	0.39	0.37	0.39
ppning-5a	0.11	0.23	0.17	0.26
ppxing-5a	0.18	0.28	0.20	0.24
ncmp_5b_total	0.32	0.43	0.41	0.40
ncmping-5b	0.05	0.14	0.03	0.10
ncmpth-5b	0.21	0.31	0.31	0.35
ncmpto-5b	0.06	0.21	0.07	0.14
nom	0.92	0.71	1.89	1.18
ppnca_total	0.40	0.48	0.76	0.68
prepostN	0.19	0.29	0.53	0.55
ppxx-??	0.14	0.30	0.19	0.29

Table F2 Descriptive Statistics for Developmental Complexity Features in Spoken Tasks Over Time

	Time 1		Time 2	
Complexity features	*M*	*SD*	*M*	*SD*
Speaking integrated				
vcmpfin_1a_total	1.06	0.73	1.01	0.59
vcmpth-1a	1.03	0.71	1.01	0.59
vcmpwh-1a	0.03	0.13	0.00	0.00
vcmpfin_2a_total	0.94	0.83	1.02	0.52
vcmpth-2a	0.82	0.81	0.90	0.56
vcmpwh-2a	0.12	0.26	0.12	0.21
fadvl-2b	2.01	0.96	1.67	0.72
vcmpnf_2c_total	0.59	0.62	0.44	0.43
vcmping-2c	0.12	0.17	0.14	0.20
vcmpto-2c	0.47	0.61	0.30	0.42
adv-2d	1.40	0.82	1.75	0.97
jatrb-2e	2.90	1.03	2.82	1.10
ppadvl-3a	2.39	1.18	2.62	0.87
jcmpth-3b	0.01	0.07	0.02	0.08
vcmpnf_3c_total	0.77	0.60	0.56	0.40
vcmping-3c	0.00	0.00	0.01	0.05

Table F2 continued

Complexity features	Time 1		Time 2	
	M	*SD*	*M*	*SD*
vcmpto-3c	0.77	0.60	0.55	0.40
finrel-3d	0.41	0.41	0.63	0.56
npnm_total	1.88	1.14	2.07	0.89
npnm-3e	1.62	0.99	1.78	0.78
npnmgen-3f	0.26	0.34	0.28	0.35
ppnof-3g	0.43	0.51	0.59	0.45
ppnc-3h	0.14	0.23	0.25	0.29
jcmpto-4a	0.05	0.17	0.06	0.13
jcmpxtra-4b	0.04	0.17	0.03	0.09
nfrel-4c	0.09	0.23	0.16	0.23
jatrb-2e4d	0.16	0.25	0.22	0.32
npnm-3e4d	0.23	0.38	0.14	0.26
npnmgen-3f4d	0.00	0.00	0.07	0.15
hn-4d	0.18	0.25	0.20	0.26
ppna-4e	0.39	0.38	0.37	0.26
ppning_5a_total	0.05	0.18	0.08	0.15
ppning-5a	0.02	0.15	0.02	0.06
ppxing-5a	0.03	0.10	0.06	0.14
ncmp_5b_total	0.12	0.20	0.19	0.19
ncmping-5b	0.00	0.00	0.01	0.04
ncmpth-5b	0.08	0.15	0.16	0.19
ncmpto-5b	0.04	0.11	0.02	0.07
nom	2.48	1.22	2.56	1.20
ppnca_total	0.53	0.42	0.63	0.39
prepostN	0.17	0.32	0.23	0.27
ppxx-??	0.41	0.49	0.13	0.20
Speaking independent				
vcmpfin_1a_total	1.20	0.84	0.83	0.57
vcmpth-1a	1.20	0.84	0.83	0.57
vcmpwh-1a	0.00	0.00	0.00	0.00
vcmpfin_2a_total	0.32	0.68	0.75	0.86
vcmpth-2a	0.25	0.51	0.69	0.86
vcmpwh-2a	0.06	0.32	0.06	0.25
fadvl-2b	2.34	1.12	1.89	1.57
vcmpnf_2c_total	0.10	0.25	0.11	0.29
vcmping-2c	0.00	0.00	0.02	0.11
vcmpto-2c	0.10	0.25	0.09	0.22
adv-2d	1.74	1.42	1.83	1.25
jatrb-2e	1.85	1.10	2.35	1.38
ppadvl-3a	4.74	1.53	4.59	1.65
jcmpth-3b	0.02	0.10	0.01	0.09
vcmpnf_3c_total	0.67	0.71	0.95	0.65
vcmping-3c	0.04	0.16	0.07	0.32
vcmpto-3c	0.63	0.71	0.88	0.66
finrel-3d	0.14	0.54	0.22	0.45
npnm_total	0.42	0.68	0.53	0.60
npnm-3e	0.39	0.67	0.40	0.46
npnmgen-3f	0.03	0.15	0.13	0.31
ppnof-3g	0.24	0.56	0.37	0.44
ppnc-3h	0.02	0.15	0.03	0.16
jcmpto-4a	0.04	0.16	0.07	0.22
jcmpxtra-4b	0.47	0.51	0.38	0.41
nfrel-4c	0.01	0.09	0.04	0.17
jatrb-2e4d	0.01	0.09	0.11	0.33
npnm-3e4d	0.00	0.00	0.06	0.22
npnmgen-3f4d	0.01	0.09	0.00	0.00
hn-4d	0.01	0.09	0.09	0.21

Table F2 Continued

Complexity features	Time 1		Time 2	
	M	*SD*	*M*	*SD*
ppna-4e	0.83	0.71	0.73	0.66
ppning_5a_total	0.30	0.43	0.23	0.38
ppning-5a	0.04	0.16	0.06	0.18
ppxing-5a	0.26	0.40	0.18	0.33
ncmp_5b_total	0.10	0.24	0.12	0.24
ncmping-5b	0.00	0.00	0.00	0.00
ncmpth-5b	0.04	0.16	0.07	0.19
ncmpto-5b	0.05	0.19	0.04	0.16
nom	0.24	0.43	0.74	0.75
ppnca_total	0.85	0.76	0.75	0.67
prepostN	0.62	0.68	0.58	0.64
ppxx-??	0.16	0.29	0.20	0.36

Suggested citation

Gray, B., Geluso, J., & Nguyen, P. (2019). *The longitudinal development of grammatical complexity at the phrasal and clausal levels in spoken and written responses to the TOEFL iBT® test* (TOEFL Research Report No. RR-90). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12280

Action Editor: John Norris

Reviewers: This report was reviewed by the Research Subcommittee of the TOEFL Committee of Examiners.