Language Teaching Research Quarterly **EUROKD**

2025 (51), 353–375
https://doi.org/10.32038/ltrq.2025.51.09

Linguistic Progression in IELTS Speaking and Learner Background Factors

Okim Kang*
Northern Arizona University, Arizona, USA
Kate Yaw
University of South Florida, USA
Hyunkee Ahn
Seoul National University, South Korea

Correspondence
Email: Okim.Kang@nau.edu

Abstract
This study examined the degree of change in test takers' speaking performances over a 3-month period. Furthermore, it investigated the impact of learner background variables on learners' linguistic progresses. Fifty-two Korean learners of English, who were enrolled in IELTS preparation classes, took part in the study. Their proficiency was initially established using scores from an in-house placement instrument. Upon completing a preliminary questionnaire, participants sat for an officially-administered IELTS pre-test. Their language learning data were collected each week by survey. A final questionnaire was then completed after 12 weeks of study, immediately following the official IELTS post-test. For speech analysis (i.e., lexico-grammatical and pronunciation features), pre- and post-test individual long-run speaking responses were coded to examine participants' linguistic gains over time. Findings indicated that fluency features improved most significantly over time, although the relationships between speech construct changes and learner background variables were more complex. Implications of these findings are useful for curriculum planning and for developing of language assessment and testing, as well as validity evidence for the IELTS speaking test.

ARTICLE HISTORY
Received: 01 July 2024
Revised: 10 October 2025
Accepted: 23 November 2025

KEYWORDS
Linguistic Progression, Linguistic Analysis, Learner Background, IELTS, Pronunciation

How to cite this article (APA 7th Edition):
Kang, O., Yaw, K., & Ahn, H. (2025). Linguistic progression in IELTS speaking and learner background factors. *Language Teaching Research Quarterly*, *51*, 353–375. https://doi.org/10.32038/ltrq.2025.51.09

©The Author(s), 2025. Published by EUROKD. This article is licensed under a Creative Commons Attribution 4.0 International License, which grants permission to use, share, adapt, distribute and reproduce in any medium or format provided that proper credit is given to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if any changes were made. www.EUROKD.com

[1]Introduction

It is a common belief that second language (L2) learners can acquire their target language skills at various paces. Various linguistic analysis research on learner language production has also provided evidence of such L2 speech development in a study abroad (SA) context (e.g., Brecht et al., 1993; Freed et al., 2004). Among learners, oral fluency gains tend to be most consistent and observable (Segalowitz & Freed, 2004). Then, linguistic development does not occur in isolation; instead, it interacts with other linguistic features as seen in Gass (1999) where lexical weaknesses accounted for inaccurate grammatical structures. However, it is still unknown how linguistic progression takes place in test-takers' oral performances, particularly in a high-stakes assessment context.

In addition, learners' background factors can affect their linguistic progression. For example, a learner's proficiency level can affect his/her language development and learning outcomes (Benigno et al., 2017; Elder & O'Loughlin, 2003). Furthermore, the amount of target language use (TLU) or language contact can play a big role in learners' language development, such as with regards to oral fluency (Freed et al., 2004). Also, time has been considered as the single best predictor of outcomes in L2 learning development (Lightbown & Spada, 2020). In fact, the impact of these variables on learning gains are particularly important to consider in an English as a Foreign Language (EFL) context, given that student access to the TLU is limited. In addition, research examining the relationships between these learner backgrounds and their linguistic progress from a longitudinal perspective has been extremely rare.

In the assessment context, longitudinal research on learners' linguistic progression and the background factors that affect this development also contributes to the wide body of evidence needed to construct a validity argument for the use of speaking test scores. Indeed, Chapelle et al. (2008) provide a valuable structure for a high-stakes language testing context, with a series of inferences beginning with domain description and building to utilization of test scores. These inferences can be categorized into four broad types: test-development related (i.e., domain definition), consistency related (i.e., evaluation, generalization), construct related (i.e., explanation, extrapolation), and use related (i.e., utilization, consequence implication) (Chapelle, 2021; Chapelle & Lee, 2021). Research involving linguistic analysis of speech produced on a high-stakes assessment and the background factors that influence this over time may offer important insights for the explanation and generalization inferences, as these relate to how scores reflect variation in test-takers proficiency levels and are comparable across parallel versions of tasks, test forms, and raters, respectively (Chapelle, 2011).

[1] This paper is part of a special issue (2025, 50-51) entitled: In honour of Carol A. Chapelle's contributions to language assessment and learning (edited by Christine Coombe, Tony Clark, and Hassan Mohebbi).

Therefore, the current study investigated learners' language development across a semester-long period (12 weeks) within an EFL context. Additionally, it explored the effects of learner background characteristics (i.e., proficiency level, hours of study invested, target language use amount) on the linguistic gains measured by speaking performances in the IELTS test. Understanding the links between linguistic construct changes and learner background characteristics can inform both the planning of curriculum and the development of tools for L2 assessment and learning while also contributing research-based evidence for consistency- and construct-related inferences in a validity argument for using the IELTS speaking test to measure adults' spoken academic English language skills.

Review of the Literature

Linguistic Evidence of Language Development in Speaking Assessment

As measures of linguistic development, English proficiency exams, especially high-stakes tests, contain tasks to elicit evidence of both productive and receptive language use. The IELTS is one example, with an academic version designed to assess the language proficiency of non-native speakers of English who plan to continue their education in a tertiary institution. The current study focused on the productive language task, particularly IELTS Academic speaking performances of Korean learners of English. Within this learner population, speaking has been demonstrated as one of the lowest sub skills (IELTS Research, 2019). IELTS speaking scores are a composite from four sub-score areas: Pronunciation, Lexical Resource, Fluency and Coherence, and Grammatical Range and Accuracy. The current study examined speaking features related to each of these four criteria respectively.

As for fluency and coherence, confirmed oral performance rating predictors (Ginther et al., 2010; Trofimovich & Baker, 2006) include both pause structures (Brown & Yule, 1983) and speech rate (Kormos & Dénes, 2004). In fact, up to half (50%) of the variance seen in oral performance ratings can be explained by these suprasegmental features (Kang et al., 2010), with non-target-like patterns of pausing and speech rate accounting for a respective 46-48% and 12-14% of listeners' negative impressions of speakers (Rossiter, 2009). Additionally, suprasegmental features demonstrate high correlations with the global discourse structure present in oral performances, given that listeners depend upon prosodic features as markers of key discourse boundaries (e.g., Pickering, 2001, 2004; Swerts, 1998) and signals of the speaker's intent to hold the floor (Wennerstrom & Siegel, 2003).

Lexical correlates with oral proficiency include vocabulary richness and range (Brown et al., 2005; Yu, 2010). As described in Nation (2013), vocabulary richness, or lexical sophistication, signifies the proportion of high and low frequency vocabulary items produced within a speaker's response, while vocabulary range, or lexical diversity, is the ratio of word types (i.e., the total number of unique words a speaker produces) to word

tokens (i.e., the total number of words a speaker produces). Lexical diversity consistently emerges as a significant predictor of oral proficiency (Iwashita et al., 2008; Lu, 2012) and comprehensibility (Saito et al., 2016), with increases in proficiency level associated with increased type-token ratios (TTR; Iwashita et al., 2008). The role of lexical sophistication, however, is less clear. In oral proficiency exams (e.g., TOEFL iBT and ACTFL), lexical sophistication plays a stronger role given than raters are explicitly trained to attend to this aspect of lexical resources (see Crossley et al., 2011; Crossley & McNamara, 2013). In contexts without such explicit rater training, though, lexical sophistication has not demonstrated the same link (Lu, 2012; Saito et al, 2016).

On the grammar front, language proficiency is impacted by both complexity and accuracy. Global measures of grammatical accuracy (Brown et al., 2005) may predict accuracy of oral language production, following empirical research in second language acquisition and language testing (e.g., Foster & Skehan, 1996). Quantified as number of errors for each C-unit, global accuracy shows significant variation among proficiency levels (Iwashita et al., 2008), as well as across speaking scores and tasks (Jamieson & Poonpon, 2013). Specifically, the most significant feature to distinguish levels of proficiency across spoken responses is the verb-phrase ratio, or number of verb phrases per C-unit (Iwashita et al., 2008). Additional measures of grammatical complexity include counts of passive structures, adjectives, and prepositional phrases, demonstrating significant effects on scores and tasks (Jamieson & Poonpon, 2013).

As for pronunciation, various features have demonstrated an association with ratings of oral proficiency (see Kang et al., 2010; Kormos & Dénes, 2004). Examples of such speech features are lexical stress, rhythm, segmental errors, tone choice, pitch range, and prominence. Inappropriate word stress contributes to both breakdowns in communication (Jenkins, 2002) and reductions in comprehensibility among non-native speakers of English (Kang, 2010). As a stress-timed language, English additionally relies on stress for appropriate rhythm patterns. For inner circle varieties of English, native speakers commonly produce rhythms with a stressed to unstressed syllable length ratio greater than 1, indicating that unstressed syllables have a consistently shorter duration than stressed ones (Kang et al., 2020). In addition, learner speech at higher levels of proficiency is associated with more frequent use of rising tones, which bolster listeners' impressions of sharing background knowledge with their interlocutor (Kang et al., 2010, Taguchi et al., 2022). Speech with narrow pitch ranges can increase listener challenges with determining prosodic units (see Pickering, 2004; Wennerstrom, 1994). Moreover, lower-proficiency speech may be characterized by excessive use of prominence, which causes listener difficulty in the appropriate allocation of their attentional resources (Wennerstrom, 2000). Finally, segmental errors with a high functional load tend to affect listener comprehension more notably than low functional load errors (Kang & Moran, 2014) suggesting that errors with a high functional load may affect proficiency ratings to a higher degree.

Language Development and Learner Background Factors in the Assessment Context

Given that learner background factors have been broadly implicated in both L2 development (e.g., Benigno et al., 2017; Dörnyei, 2005; Lightbown & Spada, 2020) and ensuing language test performance (e.g., Elder & O'Loughlin, 2003), further attention to how learners' individual differences (e.g., proficiency level) and behaviors (e.g., target language use, hours devoted to L2 study) affect language development in the assessment context is warranted. To include both behavioral and individual difference factors, we use the broader term "learner background factors" in acknowledgement of the dynamic, L2-learner-specific system that these factors comprise (Dörnyei, 2005). Three principal background factors – hours of study, proficiency level, and target language use (TLU) – were deliberately chosen for this study based on the needs of test-takers and testing agencies (e.g., IELTS).

One factor vital to development of language in learners is time. Indeed, Lightbown and Spada (2020) assert that time "may be the single best predictor of outcomes in L2 learning" (p. 422). Existing research focuses on hours of study as a predictor of L2 proficiency change (e.g., Benigno et al., 2017), thereby providing indirect evidence of linguistic development over time. Time required to reach proficiency milestones, such as those measured by the Common European Framework of Reference (CEFR), depends on both 1) the linguistic distance between the learner's L1 and L2, and 2) the starting proficiency of the learner.

Further evidence from IELTS research shows that learners at differing levels of proficiency acquire language skills at different paces, with the greatest proficiency gains among those with the lowest starting proficiency. This pattern holds in the ESL context across a range of pre/post-exam intervals, from a 12-week semester (Elder & O'Loughlin, 2003; Humphreys et al., 2012) to an entire program of study (from six months to 2.5 years; O'Loughlin & Arkoudis, 2009). This is perhaps not surprising, as higher proficiency levels are characterized by more complex linguistic features (Gray et al., 2019), which take longer to acquire. However, the relationship between learners' hours of study and L2 linguistic feature development (as opposed to overall proficiency change) remains relatively unexplored.

In addition to the aforementioned interplay between proficiency and hours of study, L2 linguistic feature development also occurs as a function of learner proficiency. A longitudinal study of pronunciation development by Korstomitina and Kang (2021) revealed that higher proficiency learners showed more improvements in prominence and fluency measures over a 15-week study period than their lower-proficiency peers, though proficiency was not a strong predictor of segmental deviations. Looking at syntactic complexity development over a three-semester period, Vercellotti (2019) found that productive (e.g., clause length, subordination) and structural (e.g., syntactic variety, weighted structural complexity) measures all increased across proficiency levels,

suggesting that higher proficiency learners possess a greater capacity for allocating their cognitive resources for more complex language production. Notably, these speech studies were conducted in ESL contexts; more research is needed to determine if similar development patterns occur among EFL learners.

Beyond the classroom, the degree of contact that learners have with their target language can also be consequential for language development (Freed et al., 2004). Immersion in the target language, such as through study abroad (SA) or second language learning (e.g., ESL), provides contact opportunities which may lead to development of learners' reading (Dewey, 2004), listening (Cubillos et al., 2008), lexical skills (Milton & Meara, 1995), spoken rhythm (i.e., stress timing in English; Trofimovich & Baker, 2006), and, perhaps most salient, oral fluency (Freed, Segalowitz, & Dewey, 2004; Kang et al., 2021). Learners in contexts other than ESL or SA may try to imitate this immersion by using language courses, technology-mediated interactions in the target language (e.g., chatting, social media, online gaming), and contact with speakers of the target language. In the assessment context, however, it is unclear what impact these EFL learner efforts can have.

The Current Study

The current study examined test-takers' linguistic progression over time particularly in IELTS Speaking and the relationships among learner background factors and various linguistic constructs in speaking. This is part of a larger project funded by the IELTS Joint Research Program, and a complete report is available to readers in Kang et al. (2021). The IELTS Speaking test is structured as a one-on-one interaction between an examiner and the candidate. Candidates have a chance to demonstrate various speaking skills across the three-part exam, which includes an introduction and short interview (Part 1), a monologic extended response (Part 2), and an interactive discussion (Part 3). As a development in the revised speaking exam (Taylor, 2001), Part 2 offers the opportunity for candidates to display sustained language production and take initiative in the interaction. Therefore, this study analyzed linguistic features of the responses produced by candidates on Part 2 of the Academic Speaking exam. There were two research questions that guided this study:

RQ$_1$: How do linguistic features of speaking in IELTS develop over the time of 3 months?

RQ$_2$: In what ways do learner background factors (i.e., hours of study, hours of target language use, and level of proficiency) impact test-takers' linguistic development as demonstrated through their IELTS speaking performance?

Methodology

Participants

The present study involved 52 Korean learners of English as a foreign language. All

participants were registered for IELTS preparation courses at a language institute located in Seoul, South Korea. Their age ranged from 16 to 53 years old (M = 26.75, SD = 8.91) with 32 females and 20 males. By using in-house IELTS-based placement test scores, participants' proficiency was determined to be beginner (IELTS band 1.0-4.0, n = 16), intermediate (IELTS band 4.0-6.0, n = 17), or advanced (IELTS band 6.0 and higher, n = 19). The placement process also considered prior IELTS scores in cases where participants reported previous IELTS experience. The preparation courses included content and practice in all four skill areas of the IELTS (listening, reading, writing, speaking).

Research Instruments

In addition to IELTS test speaking performances, we employed two study-specific measures of learner data: pre- and post-study background questionnaires and weekly surveys of language study and use. In the survey responses, learners were asked to evaluate their process of English learning and report on their individual background factors.

The IELTS test

Current versions of the official IELTS test were administered twice to participants within the context of a regularly-scheduled test administration session. Test 1 was done before starting the 12-week IELTS preparation course; Test 2 occurred after completing the course. In all cases, exams were free of charge to the participants. Once the exams were scored, we received participants' speaking band sub-scores, plus recordings of their speaking performances on the test. Measurements of participants' proficiency were the pre- and post-test band scores, with pre-test scores also used to indicate their initial proficiency level.

Background questionnaires

Participants responded to pre- and post- background questionnaires on Qualtrics twice during the study – at the start and the conclusion of the 12-week period. These questionnaires, adapted from Elder and O'Loughlin (2003), were devised to obtain data on the variables under investigation which included the hours of study and target language use as well as participants' demographics.

Hours of study was reported via nine survey items asking learners to share how many hours that they devoted to their studies inside and outside of class each week. Survey items focused on time that participants spent attending class, studying with others, studying alone, doing homework, and practicing for the IELTS (covering all four language skill areas on the IELTS: listening, reading, writing, speaking). In each item, there were 11 possible answer options in a range from 0 hours to more than 16 hours. For analysis, we generated composite scores based on responses to these nine items; scores were calculated for each of the 12 study weeks plus from responses on the post-questionnaire.

The *amount of target language use* (TLU)*,* which was adapted from Freed, Dewey, et al. (2004), was measured through 11 items focused on the weekly number of hours that learners had contact with or exposure to English outside of their language studies. These items asked learners to report on English use when communicating with L1 English users , L2 English users, family members, and online gaming peers. Items also gauged English language exposure through television, music, online videos (e.g., YouTube), movies, general internet use, social media, and reading for pleasure. Like the measure for hours of study, items on the TLU survey also provided learners with 11 answer choices spanning from 0 hours to more than 16 hours. Similar to hours of study, composite scores for data analysis were calculated based on responses to these 11 items for the 12 weekly surveys and the post-questionnaire.

Weekly language use/study survey

Throughout the duration of the study (12 weeks), learners were tasked with responding to a weekly survey reporting their hours spent studying and using English. Table 1 offers an overview of how the primary learner background variables were operationalized for this study.

Table 1

Primary Background Factors Impacting IELTS Speaking Band Score Gains

Variables	Operationalization
Hours of study	Composite of responses from 12 weekly surveys + 1 post-survey.. Each survey measured in-class and out-of-class hours of study
TLU	Composite of responses from 12 weekly surveys + 1 post-survey. Each survey measured exposure to and contact with English language: communicating in English (i.e., with native and non-native speaking friends, family, and people while online gaming), listening to music, viewing media (i.e., television, films, and videos), using social media, using the internet, & reading in English.
Level of proficiency	IELTS scores from the pre-test, spanning from 4.0 to 7.5.

Data Collection

Data collection occurred over a 12 month period. Once participants gave informed consent and responded to the pre-questionnaire, they sat for the officially-administered IELTS test. Next, they took an IELTS preparation course while responding to weekly surveys reporting their hours of language study, amount of TLU, and mock exam scores. After finishing their IELTS course, learners completed the post-questionnaire and then sat for a second official IELTS test. IELTS scores and audio files from the speaking exam were processed by IDP, then posted to research team members in the US to be transcribed and analyzed linguistically.

Speech Analysis and Coding

The first sixty seconds of the Part 2 (monologic extended run) audio responses were selected to code for linguistic features in the four IELTS speaking band categories (i.e., pronunciation, lexical resource, fluency and coherence, and grammatical range and

accuracy). The speech samples were trimmed and converted to digital .wav files using Audacity (Version 2.4.1); they were then transcribed according to the convention from Biber et al. (2024). Researchers verified the transcripts with the original audio files to ensure accuracy. To measure suprasegmental features, Kang and Johnson's (2018) prosodic modeling program was used to extract speech rate, filled pauses, silent pauses, pitch range, tone choice, and prominence. The reliability of this program was tested with a correlation of .95 between the computer and trained linguists, by using 20% of the data with .95 for fluency (Johnson & Kang, 2017; Kang et al., 2018) and .75 for prosody (Kang et al., 2021).

For vocabulary, the vocabulary profiling tool LexTutor (Version 4; Cobb, 2020) was used to calculate lexical features (i.e., type-token ratio, K1 words, K2 words, and AWL words). For grammar, rhythm, and segmentals, two trained human raters coded features with PRAAT (a computer-assisted speech analysis program; Boersma & Weenink, 2007). Inter-coder reliability values for the three sets of features that were manually coded (grammar = .99, rhythm = .98, segmental = .93) were all considered acceptable.

Linguistic analysis and features

The linguistic variables used for analysis include fluency and coherence, lexical resources, grammatical range and accuracy, and pronunciation. Please see Kang et al.'s (2021) IELTS report for more details of these analyses. Within each speaking band category, variables were clustered to reflect the final linguistic dimension for individual rating criteria along with the original linguistic features measured prior to and following the process of variable reduction. This effort at category-specific clustering was done intentionally to help readers understand the results of the speech analysis.

Fluency and coherence measures were selected on the basis of extensive research findings on L2 suprasegmental use (e.g., Kang et al., 2010; Kormos & Dénes, 2004). Given the high correlation among some of the fluency variables (r>.76), the fluency variables were reduced down to the following three features: (a) speech rate, (b) silent pauses, and (c) filled pauses. *Speech rate* was measured using a composite of *syllables per second* (total syllable count divided by total seconds of speech), *articulation rate* (total syllable count divided by the duration of speech omitting pauses), and *mean length of run* (mean number of syllables that speakers articulated between pauses ≥ 0.1 seconds). Pause variables were generated by combining the *number* and *duration* of filled and silent pauses. Due to the strong correlation between the two silent pause variables (r >.66) and the two filled pause variables (r >.62) respectively, these measures were also clustered for each of the pause dimensions. For the *number of silent and filled pauses,* a count of total number of pauses for one minute of speech was calculated. *Duration of silent and filled pauses* was determined by calculating the average length of each pause type (i.e., dividing the pause duration by the total number of the respective type of pause). Kang and Johnson's (2018) prosody modelling program was used to extract these features

automatically.

Lexical resource was operationalized according to Brown et al. (2005) as vocabulary richness and range. Within this, lexical measures included: (a) type-token ratio (TTR), (b) proportion of K1 words, (c) proportion of K2 words, and (d) proportion of AWL words. TTR was measured according to Nation (2013), by taking the total word types count and dividing this by the total word tokens count. For vocabulary richness, the measures were a proportion of tokens produced within each audio response from the following three categories: K1 (first 1000 most frequent word families), K2 (second 1000 most frequent word families), and AWL (academic word list) (Coxhead, 2000; Laufer & Nation, 1995). Correlational analysis indicated relative independence among the individual variables; weak correlation coefficients ($r < .285$) provided evidence for the retention of all the lexical resource variables.

To measure *grammatical range and accuracy,* transcripts were initially coded to identify how many C-units, error-free C-units, clauses, dependent clauses, and verb phrases were produced. For the purposes of the current study, a C-unit was operationalized as an independent clause plus its modifiers, while the operational definition for a clause was a statement that contained both a subject and a predicate (Hughes et al., 1997). A global calculation of *grammatical accuracy* measured the proportion of error-free C-units to total C-units (Brown et al., 2005). For *grammatical complexity,* a composite variable was calculated that included: (a) C-unit complexity (number of C-units divided by number of clauses), (b) verb phrase ratio (number of C-units divided by number of verb phrases), and (c) dependent clause ratio (number of dependent clauses divided by total number of clauses). There was only a weak correlation ($r = .152$) between C-unit complexity and global accuracy, so global accuracy was retained as its own independent variable. However, there were strong, statistically significant correlations between C-unit complexity and both verb phrase ratio ($r=.94$) and dependent clause ratio ($r=.87$), leading us to merge these three variables into one composite variable of grammatical complexity.

Though numerous *pronunciation* features were coded using the audio files, the features determined to be most pertinent to the tasks on the IELTS speaking test plus justified through existing literature (see Kang et al., 2010; Kormos & Dénes, 2004) included: (a) rhythm, (b) tone choice, (c) pitch range, (d) prominence, (e) lexical stress errors, and (f) segmental errors. Given the relatively small correlations ($r < .38$) among these categories and their representations of unique phonological properties, each was retained as an independent variable. When necessary within categories, however, some pronunciation features were consolidated into one variable. For instance, space and pace were highly correlated ($r = .78$) and therefore combined into one prominence variable. With segmental features, consonants and vowels were clustered by functional load, yielding two segmental variables: high and low functional load. In contrast, the treatment of tone choices was somewhat different. Although they demonstrated rather medium-strong

collinearity (especially between rising and neutral, .18 < r < .62), the three tone choices were all preserved as independent variables because of their autonomous discourse nature (Kang et al., 2010) and for the aim of enhancing the interpretation for each sound phenomenon.

Rhythm was measured by locating the first 10 two-syllable words that speakers produced in each response excerpt, measuring each syllable's length, and identifying the stressed syllable. To calculate the rhythm ratio, then, we divided the stressed syllable length by the unstressed syllable length.

To measure *tone choice,* the tone (i.e., rising, level, or falling pitch movement) on the last prominent syllable within every tone unit was identified. *Pitch range* was calculated for the 60-second speech sample as the point of F0 minima and maxima for the prominent syllables. *Prominence* was measured as *pace* and *space* in line with Vanderplank's (1993) and Kang's (2010) approach. *Pace* is the mean number of stressed words per 60 seconds of speech; *space* refers to the proportion of prominent words to the total word count. *Errors in lexical stress* were coded when speakers misplaced the syllable stress in a multisyllabic word. *Segmental errors* were identified when the segmental aspects of a speaker's oral production diverged markedly from their anticipated pronunciation. Within the language produced for analysis, a total of 112 distinct segmental error patterns were noted. Errors were coded and then classified as "high" or "low" based on Catford's (1987) determination of functional load. "High" functional load errors were those with a functional load value greater than or equal to 50 (out of 100), while "low" functional load errors were those below 50 (Kang & Moran, 2014).

Statistical Analysis

Both frequencies and descriptive statistics were first used to explain the linguistic patterns that emerged in this study. Once the linguistic construct dimensions were identified, correlational analyses were performed on each variable in order to investigate each linguistic dimension's overall saliency and any systematic changes in linguistic features produced by test-takers over the three-month study period. To address the first research question focusing on the linguistic parameter changes, linguistic gains were measured by calculating the differences between Test 1 and Test 2 speaking performances. A series of paired t-tests were performed for any significance of changes. Bonferroni adjustments were made for multiple comparisons (i.e., adjusted p=0.05/18= 0.003). As for the second research question, we conducted four multiple regression analyses using the learner background variables as predictors in the models and treating each of the linguistic variables as dependent variables, including Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation.

Results

Linguistic Construct Changes over Time

In order to better contextualize the linguistic changes, we briefly examined the actual IELTS' speaking score gains as well as its sub-scoring criteria, although it is not a research questions in the current study. The results showed no notable changes in the overall IELTS speaking test score across the 3 month period (M=.125, SD=.58, *p*=.19, using a 9-band scoring system), but the difference in the sub-scoring criterion of fluency and coherence (M=.24, SD=.74) was statistically significant with a small-medium effect size (*p* =.013, *d* =.28). The fluency and coherence gain scores were larger than those of the other sub-score categories: lexical resource (M=.12, SD=.85, *p*=.33), grammatical range and accuracy (M=-.02, SD=.70, *p*=.84) , and pronunciation (M=-.02, SD=.82, *p*=.91). Accordingly, when it comes to the linguistic changes and patterns, significant developments in fluency features could be expected, but possibly no substantial changes might follow from the grammatical, lexical, and pronunciation features.

Table 2 presents a summary of descriptive statistics for the significant linguistic variable changes. [See Kang et al.'s (2021) IELTS report for comprehensive lists for these linguistic changes.] The two columns on the right display the statistical significance level (i.e., *p*-value) for the linguistic variable change when modeled through paired t-tests. Most noticeably, all of the features related to fluency (speech rate, filled pauses, and silent pauses) showed statistically significant changes during the 12 week study program. The negative t-value with medium effect size (t=-3.07, d = .50) for speech rate indicates participants produced statistically significantly more rapid speech during the post-test performance when compared to the pre-test (12 weeks prior).

Table 2

Descriptive Statistics and Linguistic Construct Changes

Rating criterion	Linguistic Variable	Mean	SD	95% CI		t	*p*
				Lower	Upper		
Fluency/ coherence	Speech rate	-1.07	2.51	-1.775	-.372	-3.07	**.003**
	Silent pause	2.96	7.71	.811	5.104	2.767	**.008**
	Filled pause	6.84	5.611	5.28	8.404	8.794	**.000**
Lexical resource	TTR	.033	.067	.0142	.051	3.525	**.001**
	K1 words	-12.02	23.96	-18.69	-5.34	-3.62	**.001**
Pronunciation	Rhythm	-.238	.55	-.391	-.085	-3.125	**.003**
	Tone_level	.047	.177	-.002	.096	1.907	**.062**
	Prominence	.244	.505	-.384	.103	3.481	**.001**

On the other hand, the number and length of silent (near significant) and filled pauses reduced significantly from their Time 1 performance with average changes of 2.96 (d=.43) and 6.94 (d=7.68) respectively. Within these pause findings, the very large effect size for the change in filled pauses was particularly promising, as it indicated that participants produced fewer hesitation markers at Time 2 compared to Time 1. The positive development in the fluency features reported here is indeed reinforced by our

previous finding that the sub-rating criterion of Fluency and Coherence showed a statistically significant change from Time 1 to Time 2 with a small effect size (p=.013, *d*=.28).

Two Lexical Resource features demonstrated statistically significant changes with medium effect sizes: type token ratio (TTR) (p=.001, d=.65) and K1 words (p=.001, d=.47). In other words, participants showed the ability to use a greater variety of word types and the 1000 most frequent English word families after 12 weeks of studying. Yet Grammatical Range and Accuracy displayed no statistically significant changes for the linguistic features measured.

In terms of pronunciation changes, rhythm and prominence choice features were most noteworthy as variables with statistically significant improvements following three months of study. That is, there was a statistically significant change with medium effect size (t=-3.125, p=.003, d=.63) in the rhythm patterns that participants produced, as determined by dividing the stressed syllable length by the unstressed syllable length . In their performance at Time 2, test-takers produced stressed syllables that were notably longer than their unstressed syllables. It is also worth mentioning that the proportion of prominent words to total word count showed a statistically significant decrease with medium effect size (t=3.481, p=.001, d=.65). This finding indicates that participants produced significantly fewer prominent syllables during their spoken test performance after 12 weeks of study.

Impact of Three Background Factors on Linguistic Progression of IELTS Speaking

Table 3 displays descriptive statistics of the three learner background variables measured in the study (i.e., hours of study, TLU, and proficiency). Over the period of 12 weeks, participants' mean hours of study is 284.38 and their mean TLU is 273 hours.

Table 3

Descriptive Statistics of Three Background Variables

Variables	N	Minimum	Maximum	Mean	SD
Hours of study	52	120	720	284.38	170.79
Target language use	52	107	675	272.80	109.17
Level of proficiency	52	4.0	7.5	5.70	.88

Table 4 summarizes the results of analyses regressing each outcome measure (i.e., linguistic features that displayed statistically significant, or near significant, relationships with the predictor at $p < .05$) on the three learner background factors. Most striking in these findings is the potent association between proficiency and various linguistic features, most notably all of the changes in fluency features and some of the prosody feature change. In other words, with increases in proficiency, participants' speech rate became faster (t=2.151, p =.037), while both silent (t=-2.153, p=.036) and filled pauses (t=-2.389, p = .021) grew shorter. In addition to hours of study and TLU, these learner

background factors accounted for roughly 9-15% of the variance in the linguistic change variables included in the statistical model.

Table 4

Summary of Multiple Regression of Background Factors on Linguistic Features

Linguistic Features	Predictors	Coefficient (*B*)	t	Sig.	Adjusted R^2
Speech rate					.12
	Proficiency	.292	2.151	.037	
Silent pause					.092
	Proficiency	-.297	-2.153	.036	
Filled pause					.151
	Proficiency	-.318	-2.389	.021	
	Target language use	-.313	-2.228	.031	
AWL					.033
	Hours of study	.302	1.964	.055	
Grammatical complexity					.086
	Proficiency	-.376	-2.718	.009	
Rhythm					.064
	Proficiency	.326	2.332	.024	
Rising tone					.058
	Proficiency	.250	1.783	.081	
Falling tone					.087
	Target language use	.338	2.325	.024	
Level tone					.086
	Proficiency	-.314	-2.275	.027	
Prominence					.089
	Target language use	-.281	-1.935	.059	
Pitch range					.124
	Proficiency	-.395	-2.921	.005	
Lexical stress					.119
	Proficiency	-.372	-2.740	.009	
Segmental_HF					.043
	Target language use	-.262	-1.999	.051	

On the pronunciation side, proficiency strongly predicted changes to rhythm, rising and level tone, pitch range, and lexical stress. As reported in Table 5, the regression coefficients and t-test values showed that proficiency has a positive relationship with rhythm (t=2.332, p=.024) and rising tone choice (t=1.783, p=.081) changes, whereas it demonstrated a negative link to level tone choice (t=-2.275, p=.027), pitch range (t= -2.921, p=.005), and lexical stress error changes (t=-2.740, p=.009). Only rising tone choice failed to meet the critical alpha level (=.05) for statistical significance. These findings imply that with increases in proficiency, stressed syllables on average grew longer and rising tone was used more frequently. At the same time, participants used fewer level tone choices and produced less frequent errors in lexical stress with improvements in proficiency. Additionally, the pitch range contracted more when proficiency increased. These predictors, plus the other learner background factors (hours of study and TLU) explained roughly 6% to 12% of the variance in pronunciation features.

Proficiency also demonstrated a statistically significant link to grammatical complexity (t=-2.718, p=.009), which was calculated as a composite variable comprising c-unit complexity, dependent clause ratio, and verb phrase ratio. This negative association showed that when proficiency went up, the changes to the number of indicators of grammatical complexity and range were smaller. That is, although learners at lower proficiency levels attempted to generate more complex sentences and produced more changes during the three months of study, students at higher proficiency levels did not display much difference after the 12 weeks of learning, perhaps attributable to a higher baseline ability to produce more grammatically complex structures.

TLU showed strong relationships with filled pause, falling tone, and high functional segmental error changes. Together with hours of study and proficiencylevel, TLU accounted for up to 15% of variance in changes to filled pause patterns. The negative regression coefficient and t-value (t=-2.228, p =.031) indicate that greater TLU (e.g., by watching films, using social media, or communicating with friends) led participants to produce fewer and shorter hesitation markers. TLU also impact the choice to use falling tone. Participants' amount of TLU had a positive connection to this change in intonation patterns (t=2.325, p=.024), meaning that participants with more daily English use in their lives outside of class also produced falling tones more frequently. Lastly, TLU displayed a somewhat significant but negative association with segmental errors that carried a high functional load (t=-1.999, p=.051). While the statistical significance level is slightly higher than the critical alpha value of .05, it was encouraging to see the possibility of greater hours of TLU leading to positive developments in segmental production.

Finally, hours of study were unfortunately not necessarily associated with linguistic feature changes aside from AWL (academic word list), which had a relatively weak significance level (t=1.964, p=.055). Nonetheless, this result suggests a potential connection between hours of study and participants' production of AWL items, which in turn could impact the lexical resource evaluation category.

Discussion

The project investigated learners' IELTS linguistic construct changes over a 12-week period in an EFL context and further examined to what extent learner background characteristics (e.g., hours of study, TLU, and proficiency level) affected the linguistic development demonstrated in their IELTS speaking performance. Note that the present study involved a very targeted population in an EFL context, namely adult learners taking preparation classes for the IELTS at a South Korean private language school. All findings must be interpreted in a context-specific manner. Additionally, while it is beyond the scope of this paper to provide a full validity argument for the use of the IELTS speaking test to measure academic oral English skills – and indeed, such an argument would require more than one single study (Chapelle & Lee, 2021) – we offer suggestions for

ways in which these findings may contribute to a broader validity argument in line with the structure developed by Chapelle et al. (2008, 2010).

Changes of Linguistic Constructs in IELTS Speaking

Across the 12-week learning period, many of the linguistic features did not improve significantly. At least, all fluency-related features showed statistically significant improvement. Official IELTS' sub-score reports also confirmed this pattern, indicating a statistically significant increase in fluency and coherence sub-scores from Time 1 to Time 2 following 12 weeks of study. While some pronunciation features are considered more controversial in terms of their likelihood of improvement over time, there have been more consistent findings that fluency can, minimally, show some progression over time (Derwing et al., 2006; Derwing et al., 2008) and in contexts of study abroad (Segalowitz & Freed, 2004). The current study found significant improvements (with large effect sizes) in the fluency features of speech rate, silent pauses, and filled pauses. In particular, the filled pauses had a very large effect size (d=7.68), indicating that learners produced dramatically fewer markers of hesitation at Time 2 (following 12 weeks of study) than in their oral production at Time 1.

Although some of the lexical features measured in this study (type token ratio and K1 [most frequent 1000 words] usage) displayed positive developments, grammatical accuracy and complexity did not show similar changes. These results are perhaps not surprising, as a substantive body of vocabulary acquisition research has shown that gains in vocabulary occur over time (e.g., Milton & Meara, 1995) yet grammatical complexity and accuracy features have not demonstrated significant improvements within a somewhat short time period (Coleman, 1997 Freed, 1998). Moreover, the lack of statistically significant changes in the overall speaking band score and criterion scores for lexical resources and grammatical range and accuracy confirmed the finding of no statistically significant differences in grammatical features during the 12-week study period. It is possible that three months was simply not a sufficient amount of time to generate any grammatical changes.

The finding that students participating in the current project did not show any notable speaking skill changes over the three-month period also led to limited pronunciation feature gains. Indeed, only prominence and rhythm features demonstrated improvements after three months of study. Generally speaking, pronunciation gains are known to be limited in nature, often occurring only in specific contexts or with certain features (Derwing et al., 2008). However, participants' progress in their production of prominence (i.e., sentence stress) is especially note-worthy. In their Time 2 spoken responses, learners had significantly fewer prominent syllables compared to their Time 1 productions. In Kostromitina and Kang (2021), which analyzed 75 ESL students' spoken production while enrolled in an intensive English program, the only variable that demonstrated a statistically significant improvement across a semester time period was

prominence. Additionally, speakers with lower proficiency levels are more likely to produce words at relatively equal pitch levels no matter the role of individual words in the discourse structure (Kang et al., 2010; Pickering, 2001). Thus, the present result illustrates the learnability of certain types of stress features within a 2-3 month period.

Learners' rhythmic patterns, which were measured by dividing the duration of the stressed syllable by the duration of the unstressed syllable, showed statistically significant changes with a medium effect size. For Korean learners of English, with a syllable-timed L1 in which all syllables tend to be equally long, the improvement in stress-timed language pattern production that was found in this study can be viewed as a big improvement. Indeed, Levis (2005) notes that not all pronunciation features are considered learnable, but the current study provides evidence that learners may have acquired some pronunciation features without being explicitly taught.

In terms of their contribution to a validity argument for IELTS speaking score use, these findings are quite promising, specifically for the generalization inference (Chapelle et al., 2008). As one of the consistency-related inferences, generalization focuses on the claim that test scores remain consistent across raters, test forms, and testing occasions (Chapelle & Lee, 2021). The longitudinal design of the current study, with test-takers taking the IELTS before and after 12 weeks of instruction, provided evidence that as certain linguistic features (i.e., fluency) changed, so too did the test scores. For areas with no statistically significant score change (i.e., pronunciation, grammar), there were also not many notable changes in the linguistic features produced by test-takers. This suggests that the IELTS speaking scores are accurately reflecting changes in actual language production, which is to be hoped for with a high-stakes proficiency test.

Relationship between Background Variables and the Linguistic Progression

The relationship between proficiency and various linguistic features was one of the most striking patterns found in this part of the study. Proficiency demonstrated a potent association with all of the changes in fluency features, in addition to some of the prosody feature changes. There were statistically significant changes between Time 1 and Time 2 test performances for all of the fluency features measured in the study. On top of this, students' proficiency levels showed a strong link to these changes; i.e., with increases in proficiency, learners also spoke more quickly and had shorter filled and silent pauses. Moreover, proficiency strongly predicted rhythm, level and rising tone choices, pitch range, and lexical stress changes. When proficiency improved, stressed syllables became longer on average, and the use of rising tone occurred more frequently.

Additionally, higher-proficiency learners tended to make less frequent level tone choices and produce fewer lexical stress errors in comparison to their lower-proficiency counterparts. These findings align with those from the existing literature (Kang & Moran, 2014; Kang et al., 2020), which found that advanced-level learners made fewer errors

with word stress, and level tones had a negative association with proficiency. Furthermore, the changes in pitch range grew more compressed as proficiency improved. Pitch range can be a useful proficiency marker, with beginning-level students frequently producing quite narrow pitch ranges in comparison to advanced-proficiency learners (Kang, 2010). Students' proficiency level, as measured by their Time 1 IELTS test scores, could offer more extensive predictions of pronunciation development patterns. Learners' proficiency similarly predicted the changes in Grammatical Complexity. The negative association shows that with increases in proficiency, there were decreased changes in the amount of grammatical complexity and range markers , likely due to advanced learners entering the 12-week program of study with already-high abilities for creating complex sentences.

The distinct linguistic features produced by speakers at different proficiency levels provides evidence for the explanation inference of a validity argument for IELTS score use. Explanation is a construct-related inference, meaning that the inference is about how the test has assessed the target speaking construct (Chapelle & Lee, 2021). In this case, current findings demonstrate that different oral performance scores on the IELTS speaking test correspond to measurable differences in fluency, pronunciation, and grammatical complexity features of the speech produced. In short, the test scores reflect actual proficiency differences as evidenced by linguistic features in test-takers' oral responses.

Beyond proficiency, TLU was strongly related to developments in filled pauses, falling tone choice, and high functional load-based segmental errors. As students increased their TLU through their communications with friends, movie viewing, book reading, and usage of social media, they used shorter and fewer markers of hesitation. This finding supports the speculation that frequent TLU can make learners more comfortable using English and lead to fluency improvements, as literature on study abroad has demonstrated (e.g., Freed et al., 2004; Segalowitz & Freed, 2004). Greater TLU also was helpful in terms of learners' improvements to their intonation pattern by increasing their frequency of falling tone use, which is a pattern typically found in speech produced by native speakers of English (Kang, 2010; Pickering, 2001). Finally, TLU demonstrated a significant but negative association with the frequency of segmental errors most notably with regards to high functional load errors. Indeed, this progression is quite encouraging given that longitudinal pronunciation improvements, especially those in the domain of consonants and vowels, have been found to happen slowly or not at all (Kang et al., 2020).

Interestingly, hours of study did not show connections to changes in any of the linguistic features aside from AWL (academic word list) word use. While this association was rather weak, it suggests that hours of study has a direct relationship to academic word use, as well as learning. This result additionally implies that language learning is a complex process rather than following a linear, straightforward path (Larsen-Freeman,

1997, 2012). The learning journey that students follow can be unique and unpredictable, meaning that sometimes, continued practice may not result in performance gains because of some restructuring processes (McLaughlin, 1990). In order to understand these learning phenomena better, researchers can develop more refined and specified methods to better elicit learners' varying behaviors and patterns.

Conclusion

Predicting linguistic progression may not be a straightforward process, as individual learners' environmental, personal, and social factors can impact learning outcomes. In fact, the current study limitedly investigated only three background factors and the duration of this study can perhaps be too short to see any substantive gains. Also, the audio files analyzed (i.e., one-minute each) could have been longer. Finally, the assessment performances in an EFL context could have narrowed the types of performances. Future research can expand its timeline and scope to further examine learners' linguistic progression in detail. In spite of these limitations, however, the results of this study point to some practical implications.

First, test-takers can be informed that in regard to their speaking skills, it is realistic to expect fluency to be able to improve more quickly than other sub-skills. For features like segmental (i.e., vowel and consonant) errors and grammatical complexity, it is generally rather challenging to see changes in a short period of time. When test-takers are making decisions about whether to invest their time in test preparation or language learning programs, it could be helpful for them to be aware of which aspects of spoken language are likely to improve more easily than others. Then, proficiency plays a significant role in these linguistic changes.

Next, TLU may serve a valuable role in improving learners' fluency and promoting further developments in pronunciation. Although gains in test scores may perhaps require instruction that is more structured and explicit, some other speaking sub-skills (e.g., rhythm and intonation) have the potential to improve through frequent TLU and practice. Therefore, test practitioners and educators would be well served to remember that learning a language is a complex and unpredictable process – it does not follow a uniform, linear trajectory. It is important that we take a multi-dimensional approach for better understanding of our learners, including their progress, backgrounds, and needs, as well as their expectations and learning behaviors.

Generally, understanding the connections between linguistic construct development and hours that learners spend on test preparation, along with the individual factors that impact those linguistic features, can have notable effects on both curriculum planning and the development of language learning and testing. In sharing the results of this project, we hope to offer concrete evidence for those seeking to understand more about

longitudinal language learning outcomes and their connection to learner background factors.

ORCID
https://orcid.org/0000-0002-7721-5283
https://orcid.org/0000-0002-4382-9762
https://orcid.org/0009-0006-7676-4420

Publisher's Note
The claims, arguments, and counter-arguments made in this article are exclusively those of the contributing authors. Hence, they do not necessarily represent the viewpoints of the authors' affiliated institutions, or EUROKD as the publisher, the editors and the reviewers of the article.

Acknowledgements
Not applicable.

Funding
This work was funded with a 2018 grant from the IELTS Partners (British Council, Cambridge Assessment English, and IDP: IELTS Australia).

CRediT Authorship Contribution Statement
Okim Kang: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project administration, Funding acquisition
Kate Yaw: Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing
Hyunkee Ahn: Conceptualization, Resources, Data Curation, Supervision

Generative AI Use Disclosure Statement
We did not use any AI tool in this project.

Ethics Declarations
World Medical Association (WMA) Declaration of Helsinki–Ethical Principles for Medical Research Involving Human Participants
Not applicable.

Competing Interests
We have no competing interest.

Data Availability
Contact the corresponding author.

References

Benigno, V., de Jung, J., & Van Moere, A. (2017). *How long does it take to learn a language? Insights from research on language learning* (Global Scale of English Research Series). Pearson. https://www.pearson.com/english/about/gse/research.html

Biber, D., Conrad, S. and Cortes, V. (2004) If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Boersma, P., & Weenink, D. (2007). Praat, http://www.praat.org (Version 4.5.25).

Brecht, R., Davidson, D., & Ginsberg, R. (1993). *Predictors of foreign language gain during study abroad.* National Foreign Language Center. https://files.eric.ed.gov/fulltext/ED360828.pdf

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks*. (TOEFL Monograph Series MS-29). Educational Testing Service.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge University Press.

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systematic description of English phonology. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 87-100). TESOL.

Chapelle, C. A. (2011). Validation in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 717-730). Routledge.

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE Publications. https://doi.org/10.4135/9781071878811

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A., & Lee, H. (2021). Validation of spoken language assessments for adult L2 learners. In T. Haug, W. Mann, & U. Knoch (Eds.), *The handbook of language assessment across modalities* (pp. 273-284). Oxford University Press. https://doi.org/10.1093/oso/9780190885052.003.0023

Cobb, T. (2020, May). *Web VP classic v.4* [computer program]. https://www.lextutor.ca/vp/eng/

Coleman, J. A. (1997). Residence abroad within language study. *Language Teaching*, *30*(1), 1–20. https://doi.org/10.1017/S0261444800012659

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. https://doi.org/10.2307/3587951

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology, 17*(2), 171-192. http://dx.doi.org/10125/44329

Crossley, S A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45*(1), 182-193. https://doi.org/10.5054/tq.2010.244019

Cubillos, J. H., Chieffo, L., & Fan, C. (2008). The impact of short-term study abroad programs on L2 listening comprehension skills. *Foreign Language Annals, 41*(1), 157-185. https://doi.org/10.1111/j.1944-9720.2008.tb03284.x

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34*(2), 183-193. https://doi.org/10.1016/j.system.2006.01.005

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*(3), 359-380. https://doi.org/10.1093/applin/amm041

Dewey, D.P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 303-327. https://doi.org/10.1017/S0272263104262076

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410613349

Elder, C., & O'Loughlin, K. (2003). *Investigating the relationship between intensive English language study and band score gains on IELTS* (IELTS Research Reports, 4). IDP: IELTS Australia. https://www.ielts.org/-/media/research-reports/ielts_rr_volume04_report6.ashx

Foster, P., & Skehan, P. (1996). The Influence of Planning and Task Type on Second Language Performance. *Studies in Second Language Acquisition*, *18*(3), 299–323. https://doi.org/10.1017/S0272263100015047

Freed, B. (1998). An overview of issues and research in language learning in a study-abroad setting. *Frontiers: The Interdisciplinary Journal of Study Abroad, 4*, 21-60.

Freed, B. F., Dewey, D. P., Segalowitz, N. S., & Halter, R. H. (2004). The language contact profile. *Studies in Second Language Acquisition, 26*(2), 349-356. https://doi.org/10.1017/S027226310426209X

Freed, B., Segalowitz, N., & Dewey, D. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*, 275-301. https://doi.org/10.1017/S0272263104262064

Gass, S. (1999). Discussion: Incidental vocabulary learning. *Studies in Second Language Acquisition, 21*(2), 319–333. https://doi.org/10.1017/S0272263199002090

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379-399. https://doi.org/10.1177/0265532210364407

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-RR-90, ETS Research Report No. RR-19-45). ETS. https://doi.org/10.1002/ets2.12280

Hughes, D. L., McGillivray, L., & Schmidek, M. (1997). *Guide to narrative language: Procedures for assessment*. Thinking Publications.

Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A., Michael, R., & Walkinshaw, I. (2012). *Tracking international students' English proficiency over the first semester of undergraduate study* (IELTS Research Report Series, 1). IDP: IELTS Australia. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2012-1.ashx

IELTS Research (2019). Test taker performance 2019. Retrieved November 17, 2020, from https://www.ielts.org/en-us/research/test-taker-performance.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Jamieson, J., & Poonpon, K. (2013). *Developing analytic scoring guides for TOEFL iBT's Speaking Measure* (TOEFL Monograph Series RR-13-13). ETS.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23*(1), 83–103. https://doi.org/10.1093/applin/23.1.83

Johnson, D. O., & Kang, O. (2017). Comparison of algorithms to divide noisy phone sequences into syllables for automatic unconstrained English speaking proficiency scoring. *Artificial Intelligence Review, 52*(3), 1781–1804. https://doi.org/10.1007/s10462-017-9594-y

Kang, O. (2010). Salient prosodic features on judgments of second language accent. In *Proceedings of Speech Prosody 2010 (paper 016)*. International Speech Communications Association. https://doi.org/10.21437/SpeechProsody.2010-32

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal, 94*(4),554-566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O., & Moran, M. (2014). Pronunciation features in non-native speakers' oral performances. *TESOL Quarterly, 48,* 173-184. https://doi.org/10.1002/tesq.152

Kang, O., & Johnson, D. (2018). Contribution of suprasegmental to English speaking proficiency: Human rater and automated scoring system. *Language Assessment Quarterly*, *15*(2), 150-168. https://doi.org/10.1080/15434303.2018.1451531

Kang, O., Thomson, R., & Moran, M. (2020). Which Features of Accent affect Understanding? Exploring the Intelligibility Threshold of Diverse Accent Varieties. *Applied Linguistics, 41*(4), 453-480. https://doi.org/10.1093/applin/amy053

Kang, O., Ahn, H., Yaw, K., & Chung, S. (2021). Investigation of relationship among learner background, linguistic progression, and score gain on IELTS. *The IELTS Research Report Series. https://www.ielts.org/-/media/research-reports/ielts-rr_2021-1_kang-et-al.ashx*

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Kostromitina, M., & Kang, O. (2021). The effects of ESL immersion and proficiency on learners' pronunciation development. *Frontiers in Communication, 6*, Article 636122. https://doi.org/10.3389/fcomm.2021.636122

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics, 18*(2), 141-165. https://doi.org/10.1093/applin/18.2.141

Larsen-Freeman, D. (2012). From unity to diversity*:* Twenty-five years of language teaching methodology. *English Teaching Forum, 50*(2), 28–38. https://files.eric.ed.gov/fulltext/EJ982846.pdf

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369-377. https://doi.org/10.2307/3588485

Lightbown, P. M., & Spada, N. (2020). Teaching and learning L2 in the classroom: It's about time. *Language Teaching, 53*(4), 422-432. https://doi.org/10.1017/S0261444819000454

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2),, 190-208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

McLaughlin, B. (1990). Restructuring. *Applied Linguistics, 11*(2), 113-128. https://doi.org/10.1093/applin/11.2.113

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL. Institut voor Togepaste Linguistik*, (107-08), 17-34.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

O'Loughlin, K., & Arkoudis, S. (2009). *Investigating IELTS exit score gains in higher education* (IELTS Research Reports, 10). IELTS Australia. https://www.ielts.org/-/media/research-reports/ielts_rr_volume10_report3.ashx

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly 35*(2), 233-255. https://doi.org/10.2307/3587647

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes, 23*(1), 19-43. https://doi.org/10.1016/S0889-4906(03)00020-6

Rossiter, M. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review, 65*(3), 395-412. https://doi.org/10.3138/cmlr.65.3.395

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech. *Studies in Second Language Acquisition, 38*, 677-701. https://doi.org/10.1017/S0272263115000297

Segalowitz, N., & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173-199. https://doi.org/10.1017/S0272263104262027

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485-496. https://doi.org/10.1016/S0378-2166(98)00014-9

Taguchi, N., Hirschi, K., & Kang, O. (2022). Longitudinal L1 development in the prosodic marking of pragmatic meaning: Prosodic changes in L2 speech acts and individual factors. *Studies in Second Language Acquisition*, *44*(3), 843-858. https://doi.org/10.1017/S0272263121000486

Taylor, L. (2001). Revising the IELTS Speaking test: Developments in test format and task design. *Research Notes 5*, 3-5.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1-30. https://doi.org/10.1017/S0272263106060013

Vanderplank, R. (1993). Pacing and spacing as predictors of difficulty in speaking and understanding English. *English Language Teaching Journal, 47*(2), 117–125. https://doi.org/10.1093/elt/47.2.117

Vercellotti, M. L. (2019). Finding variation: assessing the development of syntactic complexity in ESL speech. *International Journal of Applied Linguistics, 29*, 233-247. https://doi.org/10.1111/ijal.12225

Wennerstrom, A. (1994). *Intonational meaning in English discourse: A study of nonnative speakers.* Applied Linguistics, *15*(4), 399–420. https://doi.org/10.1093/applin/15.4.399

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102-127). University of Michigan Press.

Wennerstrom, A., & Siegel, A. F. (2003). Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes, 36*(2), 77-107. https://doi.org/10.1207/S15326950DP3602_1

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236-259. https://doi.org/10.1093/applin/amp024