ISSN 2201-2982
2023/1

IELTS Research Reports **Online Series**

Speaking and writing features: Distinguishing IELTS proficiency levels and progression over time

Okim Kang, Jesse Egbert and Yongzhi Miao

IELTS™
BRITISH COUNCIL idp Cambridge Assessment English

Speaking and writing features: Distinguishing IELTS proficiency levels and progression over time

This project analysed Korean IELTS test-takers' performance data for writing features and comprehensively examined their association with IELTS proficiency levels and their linguistic progression in relation to their background ability. It is an extension of the previous study, Kang et al. (2021).

Funding

This research was funded by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Grant awarded 2021.

How to cite this report

Kang, O., Egbert, J., & Miao, Y. (2023). Speaking and writing features: Distinguishing IELTS proficiency levels and progression over time. *IELTS Research Reports Online Series, No. 1/23*. British Council, Cambridge Assessment English and IDP: IELTS Australia. Available at https://www.ielts.org/teaching-and-research/research-reports

Introduction

This study by Kang, Egbert and Miao was conducted with support from the IELTS Partners (British Council, IDP: IELTS Australia, and Cambridge University Press and Assessment), as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge University Press & Assessment, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, with over 200 empirical studies receiving grant funding. After undergoing a process of peer review and revision, many of the studies have been published in academic journals, in several IELTS-focused volumes in the *Studies in Language Testing* series (http://www.cambridgeenglish.org/silt), and in the *IELTS Research Reports* series. Since 2012, to facilitate timely access, the research reports have been published on the IELTS website immediately after completing the peer review and revision process.

A proficiency test such as IELTS is a useful tool to classify test-takers according to their level of language ability. As part of this classification process, deciding which features allow raters to distinguish between proficiency levels (and those that influence the development of language proficiency) is an important consideration for language test providers. The current research by Kang et al. has addressed a niche in this area and attempted, using corpus-based findings, to untangle the complexity of second language (L2) proficiency and acquisition.

Conducted in the context of IELTS preparation courses in South Korea, the research looks at the relationship between a sample of Korean test-takers' written and spoken responses and their IELTS test scores. The linguistic features that distinguish IELTS speaking and writing proficiency levels for both the overall and each of the analytic scoring criteria were investigated. Additionally, test-taker background variables were gathered via a series of survey questionnaires. As an extension of earlier research by Kang, Ahn, Yaw and Chung (2021), the researchers of the current study examined the impact of background factors (e.g., the initial level of proficiency and the length of living in English-speaking countries) and gain scores for the select linguistic variables (e.g., essay length and lexical diversity) on IELTS writing score gains.

Findings indicate a complex but expected interplay of various linguistic and individual background factors. Among the 18 speech features being analysed, speech rate was the highest contributing predictor of both the overall and analytic IELTS scores. This was followed by grammatical complexity and choice of neutral tone – each of which had a negative relationship with the test scores. In the writing test, essay length was the most significant predictor of both the overall and all the analytic scores. Lexical sophistication and diversity were the next significant predictors despite being marginal. Gains in essay length between two time periods were also the most significant predictor of increases in the overall and most of the analytic scores. In addition, the initial level of test-taker proficiency was found to predict improvement in the overall and test response score, meaning that the higher test-takers' proficiency was, the less improvement they made over time.

These results corroborate earlier research evidence found in other proficiency tests and learning contexts. Although predicting a language learning pattern is a highly complicated process, the current study aims to move towards this by identifying criterial features that distinguish across the proficiency levels. It is an encouraging result that measures related to fluency in both speaking (i.e., speech rate) and writing (i.e., essay length) were found to be the strongest predictors of proficiency levels, as they can be improved with some success in a set time period (e.g., Kang et al., 2021). It is also of note that the initial level of proficiency appears to determine the extent to which the test scores can improve.

For test-takers and instructors, the findings of this study can guide them in developing an informed curriculum and making a strategic learning plan. Researchers and educators should consider the implications of the study findings for defining and assessing L2 speaking and writing and continue to examine the intricate network of factors that lie behind acquiring the target language across diverse contexts. Taken all together, such accumulated pieces of practical and scientific evidence would have significant implications for IELTS test design, including task types, band score descriptors and more broadly, the validity argument of the test.

Dr Hye-won Lee
Senior Research Manager
Cambridge University Press & Assessment

Speaking and writing features: Distinguishing IELTS proficiency levels and progression over time

Abstract

As an extension of Kang et al. (2021), the current project further analysed 41 L1 Korean IELTS test-takers' performance data for writing features and comprehensively examined their association with IELTS proficiency levels and their linguistic progression in relation to their background. Additionally, by using previously analysed speaking features, the study examined how they could distinguish IELTS speaking proficiency.

Once participants completed the pre-test survey, they took the pre-arranged official IELTS test. Participants' hours of study and target language use information was collected weekly. The post-survey was conducted at the end of the three-month period after the official IELTS post-test. The individual long-run speaking responses from the pre- and post- tests were used for speech analysis. Forty-one (41) participants' writing samples were analysed for writing features (i.e., discourse complexity, lexical sophistication, lexical diversity, grammatical complexity, essay length).

The results showed that speech rate, grammatical complexity, the use of L1 words, intonation level-tone choice, and segmental errors were significant predictors of IELTS speaking proficiency scores with R^2 = 40–60%. As proficiency improved, IELTS test-takers spoke faster but their grammatical structure became less complex. As for writing, essay length and lexical sophistication contributed significantly to the writing proficiency scores (R^2 = 35–46%), and participants' proficiency affected their writing improvement most consistently and significantly.

Findings can inform the development of the IELTS band score descriptors and make contributions to the fields of second language (L2) testing and second language acquisition (SLA) by offering concrete evidence to assist understanding of the relationship between learning outcomes and learner backgrounds.

Authors' biodata

Okim Kang

Okim Kang is Professor of Applied Linguistics and Director of the Applied Linguistics Speech Lab at Northern Arizona University, Flagstaff, Arizona, USA. Her research interests include speech production and perception, L2 pronunciation and intelligibility, L2 oral assessment and testing, automated scoring and speech recognition, World Englishes, and language attitude.

Jesse Egbert

Jesse Egbert is Associate Professor of Applied Linguistics at Northern Arizona University. Jesse specialises in register variation, quantitative methods in linguistics, and corpus linguistic approaches to legal interpretation. He is a founding co-editor of *Register Studies* (with Bethany Gray) and co-editor (with Tony McEnery) of the Routledge series, *Advances in Corpus Linguistics*. His research has been published in a range of peer-reviewed journals.

Yongzhi Miao

Yongzhi Miao is a PhD student in Applied Linguistics at Northern Arizona University, Flagstaff, Arizona, USA. Inspired by his exposure to a variety of English accents in China, England and California, his research interests include speaking and listening, language attitude, Global Englishes, assessment, and research methods. His recent research has appeared in peer-reviewed journals such as *Language Testing* and *Language and Speech*.

Table of contents

List of tables

1 Introduction

1.1 Overview of purpose

Much research has examined the relationship between IELTS (the International English Language Testing System) scores and academic performance for a) the test's predictive validity (e.g., Hill, Storch & Lynch, 2000), b) candidates' attitudes and discourse, task difficulty, and the rating process (e.g., Brown, 2006), and c) the washback of the IELTS writing test on English for academic contexts (Green, 2007). Whilst most of the previous studies were cross-sectional in nature, our recently completed study (Kang et al., 2021) examined the relationships between learner background and IELTS score gains longitudinally. Our study also explored how various speaking features developed over time by analysing IELTS candidates' spoken responses. Findings of the present study have provided IELTS with valuable insights into the language and behaviours of test-takers and examiners of the IELTS test.

However, relatively few studies have addressed the linguistic characteristics of IELTS candidates' production for both speaking and writing skills from a longitudinal perspective. Therefore, as an extension of Kang et al. (2021), the current study seeks to further analyse IELTS test-takers' performance on both speaking and writing features to examine their linguistic progression and their association more comprehensively in relation to IELTS proficiency levels. The proposed project has many practical implications for second language acquisition (SLA) and assessment in general. Understanding how speaking and writing features distinguish proficiency levels can inform the development of the IELTS band score descriptors. Also, knowing how changes in linguistic constructs are linked to learners' proficiency levels, and what individual factors impact those linguistic parameters will have a crucial impact on curriculum planning and development of language learning and assessment.

1.2 Problem statement/rationale

The combined impact of a wide array of linguistic features on IELTS proficiency ratings of candidates' responses needs further investigation as a validation process of second language (L2) oral proficiency tests. The relationships between linguistic characteristics of candidate performances and IELTS band scores can also be the baseline for the development of IELTS scoring descriptors (assessment rubrics) as well as examiner training. In addition, our earlier project (Kang et al., 2021) mainly focused on speaking performances, but writing was the sub-score that actually improved the most significantly over time among Korean test-takers at the test-preparation school in Seoul, South Korea.

The current project focuses on speaking and writing skills because they are known to be the lowest sub-skills among Korean learners of English (British Council database, 2020). The project utilised Kang et al.'s (2021) existing data set and extended its scope. First, the project aimed to examine what linguistic features could distinguish examinees' performances at different proficiency levels in the IELTS tests for the following scoring criteria: for speaking (fluency and coherence, lexical resources, grammatical range and accuracy, and pronunciation); and for writing (coherence and cohesion, lexical resources, and grammatical range and accuracy). Second, the study explored how learner background variables (e.g., hours of study invested, amount of target language use, level of proficiency, and others) affected their writing development in the IELTS tests.

1.3 Theoretical framework: validity argument

Validity is considered crucial in language testing and assessment. Current validity approaches, stemming from the argument-based approach, move toward developing interpretive and validity arguments (e.g., Kane, 2001; Mislevy et al., 2002). This framework starts from the grounds represented by an observation of test-takers' performance on a test. A conclusion of a test-taker's ability is drawn from the observation, based on a chain of reasoning which includes inferences and their backing (usually manifested by empirical research). This validity framework is made up of a set of inferences, moving the argument from the grounds to the conclusion (Chapelle et al., 2008). The proposed project focuses on two inferences—evaluation (i.e., evaluating observed scores that reflect targeted language abilities) and explanation (i.e., warranting that expected scores are attributed to a construct of academic language proficiency). Below, we review a set of linguistic variables highly associated with performance in L2 speaking and writing.

2. Literature review

2.1. Linguistic analysis in speaking

Kang et al. (2021) conducted linguistic analyses to examine test-takers' longitudinal speaking progression, including fluency, lexicon, grammar, and pronunciation. The IELTS speaking section is a composite of scores in four sub-skill areas: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation. It should be mentioned that mean band scores for IELTS Academic for Korean learners of English are 5.7 (reading), 6.0 (listening), 5.6 (speaking), and 5.6 (writing) (IELTS Research, 2020). Notably, this indicates that the two productive skills, speaking and writing, are the lower sub-skills among Korean learners of English.

Fluency was analysed for speech rate, silent pause, and filled pause. Fluency-based measures were found to be contributing significantly to listeners' perceptual judgement of L2 speech (see e.g., Kormos & Denes, 2004). However, this effect does not seem to be linear (Munro & Derwing, 1998; Kang et al., 2022). Very low speech rate may impede communication, as with very high speech rate, and there seems to be a 'sweet spot' for the most effective speech rate under a given circumstance. However, the relationship between fluency and assessment is still not clear, and we seek to provide more empirical evidence in this regard.

Lexical correlates with oral proficiency include vocabulary range and richness (Brown et al., 2005). Vocabulary richness refers to the proportion of low and high frequency vocabulary used in each spoken response, whereas vocabulary range is the ratio of word types (i.e., unique words produced) to word tokens (i.e., all words produced) (Nation, 2013). Iwashita et al. (2008) found that increases in proficiency level were associated with an increase in the number of words produced (tokens) and a wider range of words (type). Accordingly, lexical features analysed in Kang et al. (2021) included type-token ratio, K1 words (i.e., the proportion of the most frequent 1000 words used), K2 words (i.e., the proportion of the most frequent 2000 words used), and Academic Word List words.

As for Grammatical Range and Accuracy, both accuracy and complexity contribute to determinations of language proficiency. Grammatical accuracy, when measured globally (Brown et al., 2005), is suggested as a possible predictor of oral language accuracy (e.g., Foster & Skehan 1996). Global accuracy, measured through errors per communication unit (C-unit), varies significantly among proficiency levels (Iwashita et al., 2008) and speaking tasks and scores (Jamieson & Poonpon, 2013).

The number of verb phrases per C-unit (the verb-phrase ratio) has been identified as the most significant feature that distinguishes proficiency levels among spoken responses (Iwashita et al., 2008). (Although AS-units or C-units are used in the field, the current project employed a more conventional approach.) In addition, grammatical complexity is often examined by counting occurrences of prepositional phrases, passive structures, and adjectives as they revealed a significant effect on task and scores (Jamieson & Poonpon, 2013). Therefore, all of these features, including the number of error-free C-units, C-unit complexity, verb phrase ratio, and dependent clause ratio were re-utilised in the current project as they are also known for salient speaking features that predict L2 speakers' oral proficiency (Brown et al., 2005; Iwashita et al., 2008; Kang & Wang, 2014).

Pronunciation analysis was performed for rhythm, tone choice, pitch range, prominence, lexical stress, and segmental errors. Rhythm (i.e., using a combination of stressed and unstressed syllables), prominence (i.e., emphasising important information whilst de-emphasising unimportant details), and lexical stress are found to be especially important for speech intelligibility and comprehensibility (see Field, 2005; Hahn, 2004; Isaacs & Trofimovich, 2012). For example, if a person stresses every word or every other word in English, this will make it difficult for people to understand if they are not familiar with this rhythmic pattern. Inappropriate word stress is a contributor to communication breakdowns (Jenkins, 2002) and reduced comprehensibility among L2 speakers (Kang, 2010).

In terms of tone choice, many scholars have made theoretical arguments that they will influence how a speaker is perceived and understood (see Brazil, 1997; Hirschberg, 2017). These theoretical arguments have also been corroborated with empirical evidence, suggesting that many L2 speakers overly relied on the use of level tone (i.e., monotone), thus making them sound less genuine, more boring, and less competent (see Kang et al., 2010; Kang et al., 2020; Pickering, 2001; Wennerstorm, 1998). Pitch range, referring to the difference between the speakers' highest pitch value and the lowest, was found to be the best predictor of how accented an L2 speaker sounded in a precursor study (see Kang, 2010). Tone choice, pitch range, and prominence analysed in this study were based on Brazil's (1997) framework for intonation as a communicative tool. Tone choice was determined first by identifying the tone units (similar to thought groups in pronunciation literature) and then locating any prominent syllables within that tone unit. Tone choice refers to the tone (i.e., pitch movement) on the final prominent syllable of a tone unit. Possible tone choices are rising, falling, and level. In Brazil's (1997) model, falling tones are used to present new information, rising tones present known/previously stated information, and level tones are used for procedural language. A greater use of rising tones is associated with higher proficiency as these tones contribute to listener impressions of a shared background with the speaker (Kang et al., 2010). Pitch range refers to the point of F0 minima and maxima on prominent syllables within a given speech sample.

Lastly, segmental errors in vowels and consonants are found to influence comprehensibility (see Caspers, 2010; Munro & Derwing, 1995; 2020), although the type of errors in the information or functional load they carry can be different (see Catford, 1989; Kang & Moran, 2014), with some segmental errors more influential in communication than others. The current study analysed segmental errors (deviations) based on the concept of functional loads. Kang and Moran (2014) analysed 120 test-takers' spoken responses from Cambridge English Language Assessment and demonstrated that highly proficiency (C2) learners still make segmental errors, but their high functional errors dropped drastically as their proficiency increased. Functional loads (FL) rank segmental contrasts according to their importance in English pronunciation (Brown, 1991; Catford, 1987). Those that are most severe, known as "high functional (HF) load", are the segmental errors with phonological contrasts used to distinguish meaning in a large number of words in English.

"Low functional (LF) load" errors are those in which the phonological contrast does not appear in many minimal pairs. High functional load errors tend to have a greater impact on listener comprehension (Kang & Moran, 2014) and may therefore affect proficiency ratings more than low functional load errors.

2.2 Linguistic analysis in writing

In this study, writing analysis solely focused on linguistic components of test-takers' writing scripts but did not refer to the actual Writing rubrics (topics) or prompts. It included lexical sophistication, lexical diversity and density, grammatical complexity, and discourse complexity.

In terms of lexical sophistication, the study measured the sophistication of words used by test-takers in their writing. Sophistication was measured in terms of three broad constructs related to vocabulary use: lexical prevalence, lexical diversity, and academic word list coverage. Measures of lexical prevalence are designed to quantify how frequent or widespread (e.g., range, dispersion) words are in a relevant corpus. Studies have shown that low-level English language learners (ELLs) rely heavily on a small number of frequent and widely dispersed words, but as their writing becomes more advanced, they increasingly use words that are lower in frequency and dispersion (see Kim, Crossley & Kyle, 2018; Kyle & Crossley, 2016). Lexical prevalence, defined as corpus-based measures of word importance, such as frequency and dispersion, was measured at the word level. Thus, for this study, the average lexical prevalence score was computed as the average prevalence for all words in a test-taker's writing sample.

Lexical diversity, or lexical richness, aims to measure how many different words appear in a text. The simplest measure of lexical diversity is type-token ratio. However, type-token ratio is strongly influenced by text length. Thus, alternative measures such as Moving Average Type-Token Ratio (MATTR) have been proposed to measure lexical diversity in a way that is independent of text length (see Covington & McFall, 2010). Based on strong evidence that there is a positive correlation between lexical diversity and ELL writing development (e.g., Malvern et al., 2004; Yu, 2010), we used MATTR to predict development in IELTS writing scores. We also analysed academic word list items such as the Academic Word List (AWL; Coxhead, 2000) and the Academic Vocabulary List (AVL; Gardner & Davies, 2014). We measured the proportion of the words in IELTS writing samples that were contained in the AWL.

Regarding grammatical complexity, in this study, we relied on a widely used alternative, referred to as the Register–Functional Approach to Grammatical Complexity which focuses on individual lexicogrammatical structures that are linked to distinct syntactic forms and functions, and which are motivated by register and functional patterns in actual language use (see Biber et al., 2011; Biber et al., 2020; Biber et al., 2022). In this study we focused on the features of noun phrase complexity (e.g., pre-modifying nouns, attributive adjectives, nominalisations, prepositional phrases as nominal postmodifiers, appositive noun phrases) which have been shown to be strong predictors of L2 writing development (see Ansarifar et al., 2018; Parkinson & Musgrave, 2014; Taguchi et al., 2014).

Last, with reference to discourse complexity, many automated measures have been developed for the purpose of measuring discourse complexity, most of which focus on cohesion and coherence. In this study, we measured the construct of global cohesion, and operationalise this construct in terms of the amount of content word overlap in adjacent sentences in test-taker writing samples. This measure has been shown to be a reliable predictor of L2 writing development (Crossley et al., 2016).

Overall, the current project investigated the impact of all these linguistic features (described above) on IELTS proficiency scores of candidates' responses for both speaking and writing skills. Specifically, it explored how linguistic features could distinguish examinees' performances at different proficiency. In addition, as Kang et al.'s study (2021) did for their speaking performance, the project examined how learner background variables (e.g., hours of study invested, amount of target language use, level of proficiency, and others) affected their writing development in the IELTS tests.

3. Methodology

3.1 Research questions

The project is guided by the following research questions.

1. **What are the overall speaking features that distinguish IELTS speaking proficiency levels for the following scoring criteria: fluency and coherence, lexical resources, grammatical range and accuracy, and pronunciation?**
2. **What are the overall writing features that distinguish IELTS writing proficiency levels for the following scoring criteria: coherence and cohesion, lexical resources, and grammatical range and accuracy?**
3. **How do writing features change over time and how do background variables (i.e., hours of study, amount of L2 use, level of proficiency, and others) correlate with such linguistic progression of IELTS writing?**

3.2 Research design

The proposed study adopted a quantitative/corpus-based approach and a correlational research method to the linguistic analysis of rating criteria. The study analysed the linguistic features of candidate output for each different linguistic criterion in IELTS speaking and writing. Then, it identified those criterial features in candidates' exam scores (and sub-scores) and determined the relationships between those features and learners' proficiency levels. It further examined writing band score changes between pre- and post- tests and their relationship with learner background variables.

3.3 Participants

Participants included 52 Korean students of English who enrolled in a 4-week, 8-week or 12-week IELTS preparation course at a language institute in Seoul, South Korea. Participants ranged in age from 16 to 53 years old (M = 26.75, SD = 8.91). Gender distribution was 61.5% female (n = 32) and 38.5% male (n = 20). The participants were placed into three proficiency levels – beginner (n = 16), intermediate (n = 17), and advanced (n = 19) – based on an in-house placement test with reading and writing sub-components that is regularly used by the language institute. For more details regarding the participants, see Kang et al. (2021).

3.4 Data collection procedure

Data were collected over a one-year period from May 2019 to May 2020. Participants provided informed consent and completed the questionnaires and weekly surveys (e.g., mock exam scores, hours of language study, and amount of target language use), and took the official IELTS tests before and after their IELTS preparation course.

As IELTS scores and sound files (test-takers' spoken responses) were processed by IDP, they were delivered to the research team for transcription and linguistic analysis later on. For the analysis of speech samples, 52 participants were included, but for the analysis of writing samples, only 41 candidates' writing samples were used due to the complexity involved in the data retrieval. They were selected because their data were available; other data were unfortunately unavailable at the time of data retrieval.

IELTS test scores from the pre- and post- tests were the primary outcome measures. Linguistic features for each of the writing and speaking samples were used as independent variables. Our previous project (Kang et al., 2021) also gathered learner data through the background questionnaires, weekly language study/use surveys, and online interviews. The project team received the candidates' samples through a very secure and encrypted process.

In terms of test data privacy, due to the nature of the project which requires access to test-takers' writing test materials, test data privacy and test security was seriously considered. The writing scripts to be analysed were securely stored at the IDP centre in Seoul, South Korea. The test-takers took the IELTS tests as part of an earlier study the research team undertook with a previous IELTS joint research grant (Kang et al., 2021). The IELTS tests were not paid for by the test-takers but were funded from the research grants. The test-takers all signed a comprehensive consent agreement approved by the research team's institution to allow their test materials to be used for research. The previously obtained speaking materials were securely stored in the double-locked, password protected place.

Regarding security, the project team is also aware that test security is of utmost importance, and it is possible that several of the writing rubrics (topics) could still be live (i.e., in use in current IELTS tests). In order to undertake the analysis, therefore, the project did not obtain the writing prompts, but rather used the written scripts from the Writing Task 2 along with the writing sub-scores associated with the task. Writing and speaking features focused solely on linguistic aspects of test-takers' performances and their relationships with the publicly available band score descriptors, but not the prompt-specific questions. This means that the test centre removed the prompt questions from the writing scripts, before providing the research team with the scripts. The speaking prompt questions were removed in the data of the present project. Any prompt-based variations were thus not considered in the current analysis.

3.5 Learner background variables

Based on the Kang et al. (2021) study, the first three primary background variables included hours of study, target language use, and proficiency. Then, information about other types of learner-related variables was collected through a series of surveys (i.e., pre-survey, weekly surveys, and post-survey). These variables were extracted from questionnaire responses and explored as possible contributors to candidates' improvement on the IELTS test. Table 1 below offers the list of eight (8) learner-related variables used in the study.

Table 1: *Learner background variables identified as potential predictors of IELTS gains*

Variables	Operationalisation
Hours of study	Compiled weekly survey (12 weeks) + post-survey responses (composite scores used). Each survey included: • 9 items regarding the hours spent for in-class and outside-of-class study: in-class program, homework, studying alone, studying with others, IELTS practice, & 4 skills practice each (reading/listening/speaking/writing) • 11 options to choose for weekly hours spent: 1=0, 2=less than 1 hr, 3=1-2 hrs, 4=2-4 hrs, 5=4-6 hrs, 6=6-8 hrs, 7=8-10 hrs, 8=10-12 hrs, 9=12-14 hrs, 10=14-16 hrs, 11=more than 16 hrs
Amount of target language use (TLU)	Compiled weekly survey (12 weeks) + post-survey responses (composite scores used). Each survey included: • 11 items regarding English language contact and exposure: communicating with NS friends, with NNSs, with family, with people during online game, watching TV, movies, videos, listening to music, using the internet, social media, & reading in English. • 11 options to choose for weekly hours spent: 1=0, 2=less than 1 hr, 3=1-2 hrs, 4=2-4 hrs, 5=4-6 hrs, 6=6-8 hrs, 7=8-10 hrs, 8=10-12 hrs, 9=12-14 hrs, 10=14-16 hrs, 11=more than 16 hrs
Level of proficiency	IELTS pre-test scores ranging from 4.0 to 7.5. The initial recruitment started with mock exam scores from the language institute: 16 beginners, 17 intermediate, and 19 advanced learners.
Prior English study	Years of studying English since secondary school compiled, including private tutor English courses, as collected through the pre-survey.
Educational level	Level of education, as collected through the pre-survey: 1 = final yr of secondary school, 2 = certificate/diploma, 3 = bachelor's degree, 4 = postgrad certificate/diploma, 5 = master's degree, 6 = PhD.
Prior study abroad experience	Length of living in English-speaking countries (months), as collected through the pre-survey.
Program attendance (i.e., amount of instruction)	Attendance in the program (Averaged 12 weekly surveys + post-survey responses): 1 = less than 1 hr/wk, 2 = 1-2 hrs/wk, 3 = 2-4 hrs/wk, 4 = 4-6 hrs/wk, 5 = 6-8 hrs/wk, 6 = 8-10 hrs/wk, 7 = 10-12 hrs/wk, 8 = 12-14 hrs/wk, 9 = 14-16 hrs/wk, 10 = more than 16 hrs/wk.
Instrumental motivation in studying IELTS	Four questions about different types of instrumental motivation expressed to study IELTS and determined by the presence of its IELTS-related study goals, as collected through pre-test responses: (1) parental suggestion, (2) job-related, (3) further study related, (4) general test-score achievement.

3.6 Data analysis

3.6.1 Phonological analysis

The (pre- and post-) spoken responses were coded for linguistic features in the four IELTS speaking band categories in the precursor study (i.e., fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation) through a combination of automatic computer extraction methods (Kang & Johnson, 2018a, b) and human coding (see Kang, 2010; Kang et al., 2010). A more comprehensive description of the procedures can be found in Kang et al. (2021).

The fluency variables measured were: (a) speech rate, (b) silent pauses, and (c) filled pauses. **Speech rate** was calculated as a composite of *syllables per second* (total number of syllables divided by total speech length), *articulation rate* (total number of syllables divided by time spent talking excluding pauses), and *mean length of run* (average number of syllables produced between pauses of 0.1 seconds or longer). The pause variables were a composite of the *number and duration* of each pause type (i.e., silent and filled). **Number of silent and filled pauses** was calculated as the number of pauses per minute of speech. **Duration of silent and filled pauses** was calculated as the duration of the respective pause type divided by the number of that pause type. These features were automatically extracted from the sound files using Kang's prosody modelling program.

Lexical resource was measured through vocabulary range and richness (Brown et al., 2005). The individual variables were: (a) type-token ratio (TTR), (b) proportion of K1 words, (c) proportion of K2 words, and (d) proportion of AWL words. Type-token ratio was calculated as the total number of word types divided by the total number of word tokens (Nation, 2013). Vocabulary richness was represented by a proportion of K1 (first 1000 most frequent word families), K2 (second 1000 most frequent word families), and AWL (academic word list) tokens used in each spoken response (Coxhead, 2000; Laufer & Nation, 1995).

Grammatical range and accuracy were first identified by coding transcripts for the number of C-units, number of error-free C-units, number of clauses, number of dependent clauses, and number of verb phrases. In this study, a C-unit was operationalised as an independent clause and its modifiers, while a clause was defined as a statement containing both a subject and a predicate (Hughes et al., 1997). Grammatical accuracy was measured globally as the number of error-free C-units divided by the total number of C-units (Brown et al., 2005). **Grammatical complexity** was measured as a composite of: (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).

Pronunciation features were both automatically extracted and manually coded from the sound files, including (a) rhythm, (b) tone choice, (c) pitch range, (d) prominence, (e) lexical stress errors, and (f) segmental errors. **Rhythm** was measured by identifying the first 10 two-syllable words produced in each sound file and determining the length of each syllable. The rhythm ratio was then calculated as the ratio of the length of the stressed syllable to the length of the unstressed syllable. Tone choice was measured as the tone (i.e., rising, falling, or level pitch movement) on the final prominent syllable of each tone unit. **Pitch range** was calculated as the point of F0 minima and maxima appearing on the prominent syllables within the speech sample. **Prominence** was measured as pace and space following Vanderplank's (1993) approach. **Pace** refers to the average number of stressed words per minute of speech; space is the proportion of prominent words to the total word count. **Lexical stress errors** were identified as misplaced syllable stress within words. **Segmental errors** were coded when a speaker's segmental production deviated noticeably from the expected pronunciation. A total of 112 different segmental error types were identified in speakers' language production. After coding these errors, we classified them according to Catford's (1987) functional load levels. Errors with a functional load value of 50 or higher were considered "high" functional load; those with a functional load below 50 were considered "low" functional load (Kang & Moran, 2014).

In sum, the present study investigated the following composite variables targeting the four sub-score criteria in IELTS speaking: speech rate, silent pause, and filled pause for fluency and coherence, TTR, K1 words, K2 words, and AWL words targeting lexical resources, global accuracy and grammatical complexity for grammatical range and accuracy, and rhythm, tone choice (rising, falling, level), pitch range, prominence, lexical stress, and segmentals (HF and LF) targeting pronunciation. A summary and breakdown of these features can be found in Table 2.

Table 2: *Speech features included in the present report*

Rating criteria	Originally measured features	Grouped variables for final analysis
Fluency and coherence	Speech rate Silent pause Filled pause	Speech rate Silent pause Filled pause
Lexical resources	Word type Word token Type token ratio (TTR) K1 (1000) frequent words K2 (2000) frequent words Academic word list (AWL)	TTR K1 words K2 words AWL
Grammatical range and accuracy	Number of C-units Number of error-free C-units Number of clauses Number of verb phrases Number of dependent clauses Global accuracy C-unit complexity Verb phrase ratio Dependent clause ratio	Global accuracy Grammatical complexity
Pronunciation	Rhythm Tone choice (rising, falling, level) Pitch range Prominence (pace and space) Lexical stress Segmentals: high functional (HF) consonant, low functional (LF) consonant, HF vowel, and LF vowel	Rhythm Tone choice (rising, falling, level) Pitch range Prominence Lexical stress Segmentals (HF and LF)

3.6.2 Writing feature analysis

Each of the pre-test and post-test writing samples were analysed for linguistic features that fall within five constructs: grammatical complexity, lexical sophistication, lexical diversity, discourse complexity and essay length. Table 1 contains an overview of these constructs, and the variables and tools we used to measure them.

Discourse complexity was measured with two features that have been shown to be associated with this construct: adjacent paragraph overlap and adjacent sentence overlap (Crossley, Kyle & McNamara, 2016a). These variables were computed using Crossley, Kyle & McNamara's (2016b) TAACO program (i.e., the tool for the automatic analysis of text cohesion).

Essay length was included in this study based on evidence from previous research that shows a positive relationship between the length of a writing sample and the scores assigned to it (e.g., Witte & Faigley, 1981; MacArthur et al., 2019).

To measure lexical sophistication, we used three variables from Kyle, Crossley and Berger's (2018) TAALES program: log content word frequency in BNC writing, normed rate of occurrence of Academic Formulas List items, and age of exposure. Word frequency has been shown to be associated with each of these variables (Kyle & Crossley, 2015; Kim, Crossley & Kyle, 2018).

Kris Kyle's TAALED program was used to measure lexical diversity in two ways: MATTR (moving average type token ratio) and MTLD (measure of lexical textual diversity). These measures of lexical diversity have been shown to be related to writing development (see Kyle, Crossley & Jarvis, 2021).

Finally, to measure grammatical complexity, we focused on features that have been shown to be strongly associated with grammatical complexity and development in English language learners (see Biber, Gray & Poonpon, 2011; Biber, Gray, Staples & Egbert, 2021). We used the Biber tagger to tag each text for the following linguistic features: prepositions, attributive adjectives, predicative adjectives, nominalisations, pre-modifying nouns, short passive and long passive (see Biber, 1988). Biber's TagCount program was then used to calculate normalised rates of occurrence (per 1000 words) for these features. Finally, we calculated z scores for each of these features and added them together to create a single variable for grammatical complexity.

Overall, the present study investigated the following composite variables targeting the four sub-score criteria in IELTS writing (Task 2): discourse complexity for coherence and cohesion, lexical diversity and lexical sophistication for lexical resources, and grammatical complexity for grammatical range and accuracy. Because the criterion task response is sensitive and may relate to the prompt question in the test, we did not investigate it. In addition, we included text length as a holistic measure to investigate the relationship between the number of words produced and the writing quality.

Table 3: *Writing features included in the present report*

Rating criteria	Originally measured features	Grouped variables for final analysis
Task response		
Coherence and cohesion	Overlap between adjacent paragraphs Overlap between adjacent sentences	Discourse complexity
Lexical resources	Log content word frequency in BNC Writing Normed Academic Formulas List Age of exposure MATTR – All words (first 50) MTLD – All words	Lexical sophistication Lexical diversity
Grammatical range and accuracy	Prepositions Attributive adjectives Predicative adjectives Nominalisations Pre-modifying nouns Short passive Long passive	Grammatical complexity
	Word count	Essay length

3.6.3 Statistical analysis

The first research question explored the overall speaking features that could distinguish IELTS speaking proficiency levels for the following scoring criteria: fluency and coherence, lexical resources, grammatical range and accuracy, and pronunciation. To answer this question, Pearson's correlations and five sets of multiple regression models were computed, one for the overall speaking score and one for each of the four IELTS scoring criteria. The dependent variable in each of these models was the speaking (sub-) scores across both pre-test and post-tests. The independent variables include only test administration (pre-test versus post-test) as well as the variables with significant correlations with the dependent variable to improve the degrees of freedom and the power of the models.

The second research question was: *What are the overall writing features that distinguish IELTS writing proficiency levels for the following scoring criteria: coherence and cohesion, lexical resources, and grammatical range and accuracy?* This question was also answered using Pearson's correlations and five multiple regression models, one for the overall writing score and one for each of the four IELTS scoring criteria as dependent variable. In each of these five models, the five linguistic variables were treated as independent variables.

The third research question was: *How do writing features change over time and how do background variables (i.e., hours of study, amount of L2 use, level of proficiency, and others) correlate with such linguistic progression of IELTS writing?* In order to answer the second and the third research questions, we ran five multiple regression models, one for overall writing scores and one for each of the four IELTS scoring criteria. The dependent variable in each of these models was a gain score calculated by subtracting the test score at Time 1 from the test score at Time 2. In each of these models there were eight independent variables related to the background of the test-takers.

1. Hours of study
2. Amount of TLU
3. Proficiency
4. Prior English study
5. Education level
6. Prior study abroad experience
7. Program attendance
8. Instrumental motivation

We also included gain scores for the following five linguistic variables in the study by subtracting the values at Time 1 from the values at Time 2.

1. Discourse complexity
2. Lexical sophistication
3. Lexical diversity
4. Grammatical complexity
5. Essay length

4. Results

4.1 Relationship between speech features and IELTS speaking scores

The first research question explored the overall speaking features that could distinguish IELTS speaking proficiency levels for the following scoring criteria: fluency/coherence, lexical resources, grammatical range and accuracy, and pronunciation. This question was answered by examining the relationship between each speech feature and each speaking (sub-)score for both pre-test and post-test.
This section provides a summary of the correlational and regression-based analysis for each speech feature in relation to each (sub-) score across the two time points. As can be seen from Table 4, the overall speaking scores were positively correlated with a) speech rate, b) K1 words, c) use of falling tone, and d) prominence use of neutral tone, and negatively correlated with a) TTR, b) complexity, c) high-functional and d) low-functional segmental errors.

The Fluency and Coherence sub-score showed significant and positive associations with a) speech rate, b) K1 words, and c) prominence, and negatively correlated with a) TTR, b) complexity, c) use of neutral tone, d) high-functional and e) low-functional segmental errors. The Lexical Resources sub-score was significantly positively correlated with a) speech rate, b) K1 words, c) AWL words, and d) prominence, and negatively correlated with a) TTR, b) complexity, and c) use of neutral tone. The Grammatical Range and Accuracy sub-score was significantly positively correlated with a) speech rate and b) prominence, and negatively correlated with a) TTR, b) complexity, c) use of neutral tone, c) low-functional segmental errors. The Pronunciation sub-score was significantly positively correlated with a) speech rate, b) use of falling tone, and c) prominence, and negatively correlated with a) TTR, b) complexity, c) use of neutral tone, d) word stress errors, e) high-functional and f) low-functional segmental errors. In all relationships, speech rate demonstrated the highest association with the IELTS speaking scores with a moderate strength of $.35 < r < .49$.

Lastly, the speaking features that significantly correlated with the overall speaking features include speech rate, TTR, K1 words, grammatical complexity, neutral and fall tone choice use, prominence, and segmental errors. In particular, as proficiency scores increased, IELTS test-takers increased their speech rate, using proportionally more K1 words. However, as their proficiency went up, their TTR and grammatical complexity reduced, possibly because they were compromising their skills for speeding up their speech.

Table 4: *Correlation between individual speech features with IELTS speaking test scores*

	Fluency & Coherence	Lexical Resources	Grammatical Range & Accuracy	Pronunciation	Overall Speaking
1. Speech rate	0.49**	0.40**	0.41**	0.35**	0.46**
2. Silent pauses	0.03	-0.01	0.06	0.04	0.02
3. Filled pauses	-0.12	-0.14	-0.05	-0.05	-0.11
4. Type-token ratio	-0.30*	-0.20*	-0.26**	-0.30**	-0.29**
5. K1 words	0.22*	0.27*	0.13	0.19	0.23*
6. K2 words	-0.05	-0.12	-0.01	0.00	-0.06
7. Academic Word List words	0.08	0.13	0.08	0.05	0.17
8. Accuracy	0.10	0.01	0.01	0.15	0.03
9. Complexity	-0.36**	-0.43**	-0.34*	-0.24*	-0.35**
10. Rhythm	0.10	0.12	-0.01	0.04	0.06
11. Rise	0.20	0.10	0.16	0.01	0.14
12. Neutral	-0.30**	-0.22*	-0.25*	-0.20*	-0.29*
13. Fall	0.19	0.16	0.17	0.24*	0.22*
14. Pitch	0.10	0.00	0.11	0.06	0.07
15. Prominence	0.45**	0.38**	0.25**	0.30**	0.41**
16. Word stress	-0.09	0.02	-0.03	-0.25*	-0.07
17. High-functional segmentals	-0.22*	-0.18	-0.16	-0.27**	-0.25*
18. Low-functional segmentals	-0.25*	-0.15	-0.20*	-0.21*	-0.22*

*p < 0.05, **p < 0.01, ***p < 0.001

Based on the correlational analysis, five multiple regression models were performed with test administration (pre-test versus post-test) and the significant correlates as the predictors and the (sub-)scores as the outcome variables. We put the test administration as a predictor as a categorical variable to see if the test time itself could contribute to the proficiency scores. For the first model, we measured the degree to which overall speaking scores can be predicted by the speaking features and the test administration (see Table 5). The model was significant: $F(9, 889) = 8.16$, $p < 0.001$ and these predictor variables accounted for a combined 45% (Unadjusted R^2) of the variance in the overall speaking scores (Adjusted $R^2 = 40\%$). Table 5 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. Speech rate was especially an important predictor of the IELTS overall speaking scores followed by grammatical complexity. Pronunciation features (i.e., neutral tone choice, high and low functional load errors) explained about 4–6% of the variance in the overall speaking scores.

Table 5: *ANOVA results and relative importance of linguistic variables and test in predicting overall speaking scores*

	β	F	p	R^2	Cumulative R^2
Speech rate	0.40	33.83	0.03970*	0.10	0.10
Complexity	-0.25	10.81	0.00373**	0.08	0.18
Neutral	-0.23	9.61	0.00686**	0.06	0.24
Prominence	-0.07	0.54	0.72225	0.06	0.30
TTR	-0.15	4.25	0.08839	0.04	0.34
HF	-0.19	7.52	0.02251*	0.04	0.38
LF	-0.18	4.03	0.03616*	0.04	0.42
K1	0.12	1.22	0.17504	0.02	0.44
Test	-0.11	1.59	0.21063	0.01	0.45

*p < 0.05, **p < 0.01, ***p < 0.001

For the second model, we measured the degree to which the sub-score Fluency and Coherence can be predicted by the linguistic variables and test (see Table 6). The model was significant: $F(9, 88) = 8.63$, $p < 0.001$ and these predictor variables accounted for a combined 46% (Unadjusted R^2) of the variance in the overall speaking scores (Adjusted $R^2 = 41\%$). Table 6 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. Overall, the results of this sub-score (Fluency and Coherence) model were similar to the one with the overall scores, but speech rate explained slightly more variance which is somewhat expected as speech rate is a part of fluency features. Although its contribution was not significant, prominence was the next potent predictor of the sub-score of Fluency and Coherence. The use of neutral tone choice remained as a significant predictor in this model.

Table 6: *ANOVA results and relative importance of linguistic variables and test in predicting the sub-score fluency and coherence*

	β	*F*	*p*	R^2	Cumulative R^2
Speech rate	0.38	39.92	0.03982*	0.11	0.11
Prominence	-0.01	0.10	0.95264	0.08	0.19
Complexity	-0.22	9.10	0.00879**	0.07	0.26
Neutral	-0.23	10.76	0.00681**	0.06	0.32
LF	-0.21	6.17	0.01246*	0.05	0.37
TTR	-0.14	4.51	0.08788	0.04	0.41
HF	-0.15	5.71	0.06333	0.03	0.44
K1	0.09	0.82	0.26968	0.02	0.46
Test	-0.07	0.61	0.43840	0.00	0.46

*p < 0.05, **p < 0.01, ***p < 0.001

For the third model, we measured the degree to which the sub-score Lexical Resources could be predicted by the linguistic variables and test (see Table 7). The model was significant: $F(7, 90) = 6.77$, $p < 0.001$ and these predictor variables accounted for a combined 32% (Unadjusted R^2) of the variance in the overall speaking scores (Adjusted R^2 = 29%). Table 7 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. Unlike the Fluency and Coherence model above, the use of K1 words emerged as a significant predictor. Also, Grammatical Complexity turned out to be the strongest predictor of the Lexical Resources sub-scores (p = 0.0006, R^2 = .12). Surprisingly, TTR did not make a significant contribution to the sub-score of this criterion. Speech Rate was also not significantly associated with this sub-score model, although it explained about 6% of the variance. Somewhat unexpectedly, the use of neutral tone choice also emerged as a significant predictor in this Lexical Resources model.

Table 7: *ANOVA results and relative importance of linguistic variables and test in predicting the sub-score lexical resources*

	β	*F*	*p*	R^2	Cumulative R^2
Complexity	-0.32	15.26	0.000610***	0.12	0.12
Speech rate	0.24	22.13	0.236670	0.06	0.18
Prominence	0.01	0.02	0.954511	0.05	0.23
K1	0.20	3.19	0.036426*	0.04	0.27
Neutral	-0.20	5.03	0.024538*	0.04	0.31
TTR	-0.07	1.06	0.461662	0.01	0.32
Test	-0.08	0.69	0.407023	0.00	0.32

*p < 0.05, **p < 0.01, ***p < 0.001

For the fourth model, we measured the degree to which the sub-score Grammatical Range and Accuracy could be predicted by the linguistic variables and test (see Table 8). The model was significant: $F(7, 90) = 6.98$, $p < 0.001$ and these predictor variables accounted for a combined 34% (Unadjusted R^2) of the variance in the overall speaking scores (Adjusted R^2 = 30%). Table 8 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. Not surprisingly, Grammatical Complexity was one of the significant predictors (p = .0147, R^2 = .07). Then, speech rate contributed somewhat more strongly than other variables. Unexpectedly, the use of neutral tone choice and segmental errors still emerged as a significant predictor in this Grammatical Range and Accuracy model. This result shows the inter-connectivity among linguistic features across different speech constructs. The results can also be interpreted as rater bias, in that neutral tone should perhaps not be considered as part of the grammar score.

Table 8: *ANOVA results and relative importance of linguistic variables and test in predicting the sub-score grammatical range and accuracy*

	β	*F*	*p*	R^2	Cumulative R^2
Speech rate	0.45	23.19	0.0252*	0.09	0.09
Complexity	-0.22	8.10	0.0147*	0.07	0.16
Prominence	-0.11	0.95	0.5894	0.05	0.21
TTR	-0.15	3.05	0.1137	0.04	0.25
Neutral	-0.19	5.48	0.0374*	0.04	0.29
LF	-0.22	5.29	0.0151*	0.04	0.33
Test	-0.15	2.78	0.0988	0.01	0.34

*p < 0.05, **p < 0.01, ***p < 0.001

For the last model, we measured the degree to which the sub-score Pronunciation can be predicted by the linguistic variables and test (see Table 9). The model was significant: $F(6, 88) = 5.00$, $p < 0.001$ and these predictor variables accounted for a combined 34% (Unadjusted R^2) of the variance in the overall speaking scores (Adjusted R^2 = 27%). Table 9 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. This model was relatively weaker than the other ones, in that each of the variables showed a somewhat weak strength of associations, and only two variables were statistically significant. Unexpectedly, TTR was a significant predictor of the sub-score of Pronunciation, although its predictive value measured in R^2 was only .06. Low functional errors showed a marginal level of significance (p = .0475). Speech rate was not a significant predictor, but it explained about 6% of the variance in this model, meaning that it made a contribution to the model to some extent.

Table 9: *ANOVA results and relative importance of linguistic variables and test in predicting the sub-score pronunciation*

	β	*F*	*p*	R^2	Cumulative R^2
Speech rate	0.28	15.93	0.1673	0.06	0.06
TTR	-0.21	5.54	0.0246*	0.06	0.12
HF	-0.18	8.05	0.0555	0.05	0.17
LF	-0.18	3.58	0.0475*	0.04	0.21
Word stress	-0.15	3.24	0.1173	0.04	0.25
Prominence	-0.03	0.49	0.8896	0.03	0.28
Complexity	-0.15	2.86	0.1029	0.03	0.31
Neutral	-0.13	3.12	0.1635	0.02	0.33
Test	-0.14	2.22	0.1399	0.01	0.34

*p < 0.05, **p < 0.01, ***p < 0.001

4.2 Relationship between writing features and IELTS writing scores

The second research question is: *What are the overall writing features that distinguish IELTS writing proficiency levels for the following scoring criteria: coherence and cohesion, lexical resources, and grammatical range and accuracy?* As seen in Table 10, there are small positive correlations between the IELTS scoring criteria and the linguistic variables. The linguistic variable that correlated most strongly with the scoring criteria was Essay Length, with correlations that ranged between 0.52 and 0.59. Lexical Sophistication had the next strongest correlations, ranging between 0.26 and 0.37. This was followed by Discourse Complexity which had correlations in the range of 0.22 to 0.33. Next was Grammatical Complexity with correlations between 0.17 and 0.29, followed by Lexical Diversity with correlations ranging from 0.21 to 0.26.

Table 10: *Pearson's r for IELTS proficiency levels and linguistic variables*

	Discourse complexity	Lexical sophistication	Lexical diversity	Grammatical complexity	Essay length
Task Response	0.23	0.30	0.22	0.29	0.53
Coherence and Cohesion	0.22	0.37	0.26	0.18	0.52
Lexical Resources	0.27	0.30	0.21	0.26	0.53
Grammatical Range and Accuracy	0.33	0.26	0.21	0.17	0.59
Overall Writing Score	0.30	0.35	0.25	0.26	0.61

In some cases, these variables correlate in unexpected ways. For example, both Lexical Sophistication and Lexical Diversity correlate more strongly with Coherence and Cohesion than with Lexical Resources. Likewise, Grammatical Complexity correlates more strongly with Task Response and Lexical Resources than with Grammatical Range and Accuracy. As a final example, Discourse Complexity correlates more strongly with Grammatical Range and Accuracy and Lexical Resources than with Coherence and Cohesion.

We also ran five multiple regression models, one for the overall writing score and one for each of the four IELTS scoring criteria as dependent variable. In each of these five models, the independent variables were the five linguistic variables and the test administration (pre-test versus post-test).

For the first model, we measured the degree to which overall writing scores can be predicted by the linguistic variables and test. The model was significant: $F(6, 80) = 11.28$, $p < 0.001$ and these six variables accounted for a combined 46% of the variance in Task Response scores ($R^2 = 46\%$). Table 11 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. Essay Length, Lexical Sophistication, Discourse Complexity, and Lexical Diversity predicted the overall writing scores statistically significantly, but Essay Length was particularly the strongest predictor ($R^2 = 24\%$). That is, with $\beta = 0.40$, as IELTS candidates' proficiency increased, their essay length became longer. The test administration between pre-test and post-test did not make a significant difference in predicting the writing scores.

Table 11: *ANOVA results and relative importance of five linguistic variables and test in predicting overall writing scores*

	β	F	p	R^2	Cumulative R^2
Essay Length	0.40	14.81	0.0002***	0.24	0.24
Lexical Sophistication	0.07	21.21	0.00002***	0.06	0.31
Discourse Complexity	0.02	12.92	0.0006***	0.06	0.37
Lexical Diversity	0.06	16.34	0.0002***	0.05	0.42
Grammatical Complexity	0.02	2.37	0.13	0.03	0.45
Test	0.03	0.04	0.85	0.01	0.46

*p < 0.05, **p < 0.01, ***p < 0.001

For the second model, we measured the degree to which Task Response scores can be predicted by the linguistic variables. The model was significant: $F(6, 80) = 7.55$, $p < 0.001$ and these six variables accounted for a combined 36% of the variance in Task Response scores ($R^2 = 36\%$). Table 12 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. Unlike the overall score prediction above, grammatical complexity significantly predicted the task response scores ($p = 0.05$, $R^2 = 4\%$), although R^2 is still minimal.

Table 12: *ANOVA results and relative importance of five linguistic variables and test in predicting task response scores*

	β	F	p	R^2	Cumulative R^2
Essay Length	0.47	11.47	0.001*	0.19	0.19
Lexical Sophistication	0.06	13.60	0.0004*	0.05	0.24
Grammatical Complexity	0.05	4.09	0.05*	0.04	0.28
Discourse Complexity	-0.01	6.61	0.012*	0.04	0.32
Lexical Diversity	0.05	9.50	0.003*	0.03	0.35
Test	0.04	0.04	0.84	0.01	0.36

*p < 0.05, **p < 0.01, ***p < 0.001

For the third model, we measured the degree to which Coherence and Cohesion scores can be predicted by the linguistic variables. The model was significant: $F(6, 80) = 7.88$, $p < 0.001$ and these six variables accounted for a combined 37% of the variance in Coherence and Cohesion scores ($R^2 = 37\%$). Table 13 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. Essay Length and Lexical Sophistication constantly emerged as strong predictors ($p = 0.003$ and $p < 0.001$). Lexical Diversity was a potent predictor while Grammatical Complexity did not contribute significantly. Test administration (pre-test vs. post-test) did not make a significant difference either ($p = 0.49$).

Table 13: *ANOVA results and relative importance of five linguistic variables and test in predicting coherence and cohesion scores*

	β	F	p	R^2	Cumulative R^2
Essay Length	0.35	9.51	0.003***	0.18	0.18
Lexical Sophistication	0.10	19.59	0.00003***	0.08	0.26
Lexical Diversity	0.06	11.38	0.001***	0.05	0.30
Discourse Complexity	0.01	6.14	0.02*	0.04	0.34
Test	0.10	0.47	0.49	0.02	0.36
Grammatical Complexity	0.00	0.17	0.68	0.01	0.37

*p < 0.05, **p < 0.01, ***p < 0.001

The fourth model accounted for the degree to which Lexical Resources scores can be predicted by the linguistic variables. The model was significant: $F(6, 80) = 7.14$, $p < 0.001$ and these six variables accounted for a combined 35% of the variance in Lexical Resources scores ($R^2 = 35\%$). Table 14 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. All five variables predicted the Lexical Resources scores significantly. Test administration still did not make a significant contribution in this model.

Table 14: *ANOVA results and relative importance of five linguistic variables and test in predicting lexical resources scores*

	β	F	p	R^2	Cumulative R^2
Essay Length	0.37	8.21	0.005***	0.18	0.18
Discourse Complexity	0.03	8.90	0.004***	0.05	0.23
Lexical Sophistication	0.06	13.06	0.0005***	0.05	0.27
Grammatical Complexity	0.04	2.88	0.09	0.04	0.31
Lexical Diversity	0.07	9.78	0.002***	0.03	0.34
Test	0.00	0.00	0.99	0.01	0.35

*p < 0.05, **p < 0.01, ***p < 0.001

The fifth and final model accounted for the degree to which Grammatical Range and Accuracy scores can be predicted by the linguistic variables. The model was significant: $F(6, 80) = 8.65$, $p < 0.001$ and these six variables accounted for a combined 39% of the variance in Grammatical Range and Accuracy scores (R^2 = 39%). Table 15 contains ANOVA results and a breakdown of the contribution of each of the variables to the overall R^2. As seen in Table 9, the sub-category of Grammatical Range and Accuracy scores was not strongly associated with Grammatical Complexity. As a result, Grammatical Complexity was not a significant predictor of the Grammatical Range/Accuracy scores. Again, the pre-/post-test administration did not make a substantial contribution.

Table 15: *ANOVA results and relative importance of five linguistic variables and test in predicting grammatical range and accuracy scores*

	β	*F*	*p*	R^2	Cumulative R^2
Essay Length	0.39	10.77	0.002***	0.22	0.22
Discourse Complexity	0.05	14.56	0.0003***	0.07	0.30
Lexical Diversity	0.08	14.08	0.0003***	0.04	0.34
Lexical Sophistication	0.06	11.87	0.0009***	0.04	0.38
Grammatical Complexity	0.01	0.52	0.47	0.01	0.39
Test	-0.04	0.07	0.79	0.00	0.39

*p < 0.05, **p < 0.01, ***p < 0.001

4.3 Writing improvement and learner background

The third research question is: *How do writing features change over time and how do background variables (i.e., hours of study, amount of L2 use, level of proficiency, and others) correlate with such linguistic progression of IELTS writing?* We ran five multiple regression models, one for overall writing scores and one for each of the four IELTS scoring criteria. The dependent variable in each of these models was a gain score calculated by subtracting the test score at Time 1 from the test score at Time 2. Each of these models included the following eight independent variables related to the background of the test-takers (see Table 1).

- Hours of study
- Amount of Target Language Use (TLU)
- Proficiency
- Prior English study
- Education level
- Prior study abroad experience
- Program attendance
- Instrumental motivation

The other five variables were gain scores for the five linguistic variables.

- Discourse complexity
- Lexical sophistication
- Lexical diversity
- Grammatical complexity
- Essay length

For the first model, we measured the degree to which overall writing gain scores can be predicted by the thirteen independent variables. The overall model was not significant, $F(13, 24) = 2.64$, $p = 0.19$ and these 13 variables accounted for a combined 59% of the variance in overall writing gain scores ($R^2 = 59\%$). A series of ANOVAs revealed that three of the variables were significant predictors: proficiency, prior study abroad experience, and essay length gain score (see Table 16).

Table 16: *Significant predictors of overall writing gain scores*

	β	F	p	R^2
Essay Length gain score	0.43	12.07	0.002	.197
Prior study abroad experience	0.03	8.22	0.008	.153
Proficiency	-0.28	5.26	0.03	.064

For the second model, we measured the degree to which Task Response gain scores can be predicted by the thirteen independent variables. The overall model was not significant, $F(13, 24) = 1.60$, $p = 0.15$ and these 13 variables accounted for a combined 46% of the variance in Task Response gain scores ($R^2 = 0.59$). A series of ANOVAs revealed that two of the variables were significant predictors: proficiency and essay length gain score (see Table 17).

Table 17: *Significant predictors of task response gain scores*

	β	F	p	R^2
Proficiency	-0.56	5.36	0.03	.109
Essay Length gain score	0.54	6.16	0.02	.109

In model 3, we measured the degree to which Coherence and Cohesion gain scores can be predicted by the thirteen independent variables. The overall model was not significant, $F(13, 24) = 1.74$, $p = 0.12$ and these13 variables accounted for a combined 49% of the variance in Coherence and Cohesion gain scores ($R^2 = 0.49$). A series of ANOVAs revealed that two of the variables were significant predictors: essay length gain score and lexical diversity gain score (see Table 18).

Table 18: *Significant predictors of coherence and cohesion gain scores*

	β	F	p	R^2
Essay Length gain score	0.53	10.01	0.004	.221
Lexical diversity gain score	0.10	6.03	0.02	.070

In the fourth model, we measured the degree to which Lexical Resources gain scores can be predicted by the thirteen independent variables. The overall model was not significant, $F(13, 24) = 1.68$, $p = 0.13$ and these 13 variables accounted for a combined 48% of the variance in Lexical Resources gain scores ($R^2 = 48\%$). A series of ANOVAs revealed that the only significant predictor was prior study abroad experience (see Table 19).

Table 19: *Significant predictors of lexical resources gain scores*

	β	F	p	R^2
Prior study abroad experience	0.05	10.39	0.004	.22

In the fifth and final model, we measured the degree to which Grammatical Range and Accuracy gain scores can be predicted by the thirteen independent variables. The overall model was not significant, $F(13, 24) = 1.81$, $p = 0.10$ and these 13 variables accounted for a combined 50% of the variance in which Grammatical Range and Accuracy gain scores ($R^2 = 50\%$). A series of ANOVAs revealed that three of the variables were significant predictors: essay length gain score, prior study abroad experience and lexical diversity gain score (see Table 20).

Table 20: *Significant predictors of grammatical range and accuracy gain scores*

	β	*F*	*p*	R^2
Essay Length gain score	0.35	5.67	0.03	.145
Prior study abroad experience	0.03	5.38	0.03	.121
Lexical diversity gain score	0.09	5.14	0.03	.034

5. Discussion

5.1 Relationship between speech features and IELTS speaking scores

This section seeks to interpret the findings that emerged from the present report in light of the previous literature. The first research question examined the relationship between speech features and IELTS overall scores and sub-scores for speaking across different time points. Overall, we found a relatively stable set of variables that predicted the speaking scores and sub-scores, explaining a significant proportion of the variances in each model: a) speech rate, c) K1 words, d) complexity, e) use of neutral tone, and f) high-functional and low-functional segmentals. In short, the findings tentatively suggested that a variety of speech features were related to IELTS speaking scores and sub-scores.

It is interesting that even though each sub-category has its unique criterion description (i.e., fluency, lexical resources, grammatical range & accuracy, and pronunciation), we found that some of the linguistic features intertwined among themselves. For example, grammatical complexity was the most significant predictor for the Lexical Resources sub-scores, or speech rate contributed most substantially to the Grammatical Range and Accuracy sub-scores. In addition, the use of neutral tone choice made a constant contribution to many sub-scores such as Fluency, Lexical Resources, and Grammatical Range and Accuracy, although they were not directly related to pronunciation. This phenomenon confirms that speaking is the most difficult language skill to assess reliably (Luoma, 2004). As Douglas (1997) stated, language test developers might hope that, "if raters focus attention only on pronunciation, grammar, fluency and comprehensibility, for example, the many other features of the discourse will not influence them. This is mounting evidence that this is a vain hope" (p.22). Indeed, this finding verifies this argument that each of the speech constructs does not work independently, but instead they all are inter-connected. An alternative explanation is that the raters have difficulty distinguishing between features, in which case future rater training could attend to this aspect.

Speech rate was consistently associated with the overall scores and most of the sub-scores. The importance of speech rate in language assessment has been well documented (e.g., Kang, 2010; Kormos & Denes, 2004). Faster speech tended to receive higher ratings in the present study as well as in many precursor studies. That is, as IELTS candidates' rate of speech went up, the scores they received were higher.

A similar finding was observed in Kang and Wang (2014) and Kang and Yan (2018)'s linguistic analysis results with 120 speech samples in the Cambridge English Language Assessment (CELA). Although the relationship was entirely straightforward as there was no clear distinction between adjacent levels of proficiency, speech rate was clearly a potent variable in predicting or distinguishing proficiency levels. However, this finding should be interpreted with caution, as the relationship between speech rate and listener ratings may not be necessarily linear, but perhaps it happens in a U-shaped fashion (see Munro & Derwing, 1998; Kang et al., 2022).

Complexity was negatively correlated with speaking scores. To interpret this finding, it is possible that complex, opaque language did not reflect the conversation situation prescribed by the IELTS speaking test. It is also possible that given complexity, test-takers might have made more mistakes which resulted in lower scores. As their fluency improved, candidates' use of sentence complexity might have been compromised. This pattern was also found in the spoken responses in the CELA (see Kang & Yan, 2018).

It is interesting to see that TTR was significantly but negatively associated with some of the sub-scores in the current dataset as seen in Table 3. TTR is a measure of lexical diversity (i.e., word type to word token ratio). The fact that TTR is negatively associated with speaking scores meant that the less diverse the vocabulary, the higher the scores. This finding is related to another finding where the use of K1 words were very strongly associated with scores. That is, the more test-takers used words within the first 1000 words (i.e., high-frequency words), the higher scores they tended to receive. Comparatively, using advanced words did not contribute to the proficiency scores. This finding appears to be surprising, as one would think the use of more advanced words could lead to a higher score. Perhaps, this can be explained by the fact that fluency (speech rate) was such an important contributor to the proficiency ratings and other complex grammatical structure, or that vocabulary uses did not affect raters' rating scores.

In addition, such findings can be interpreted in the following two ways. First, IELTS speaking is characterised by day-to-day conversations, meaning that the prompt questions are designed as a way for test-takers to talk about their regular daily activities. O'Keeffe et al. (2007) found that most of the words in conversations were very common English words. Thus, it is possible that the use of high-frequency words resembled natural English conversations and thus received higher ratings. Furthermore, this finding can be interpreted in light of the current proficiency level of the test-takers in the current sample. The sample in the study were around intermediate level. Thus, it is possible that they could use more high-frequency words at ease. Using less frequent words may be associated with errors which could lead to lower scores. These interpretations, however, should be tested with empirical data.

It is not surprising that the use of tone choice was predictive of test-takers' performance. The present report found that overuse of neutral / level intonation was oftentimes associated with lower scores. This echoed previous finding where L2 English speakers tended to rely on the use of neutral / level intonation in their speech, and this could lead to them being perceived as more boring, less engaged in the communication, or less sociable, and less effective in communication (Kang, 2010; Pickering, 2001; Wennerstorm, 1998).

The last prominent finding was that segmental features, regardless of high-functional or low-functional segmentals, could feed into test-takers' IELTS speaking scores. This finding is in line with previous studies which found that segmental errors (e.g., vowel / consonant substitutions, additions, or deletions) were associated with decreased comprehensibility or intelligibility (e.g., Caspers, 2010; Kang & Moran, 2014; Munro & Derwing, 1995; 2020). The contribution of segmental errors to speaking proficiency has been rather consistent all throughout sub-skills in the IELTS speaking scores.

5.2 Relationship between writing features and IELTS writing scores

In this section, we interpret the findings for research questions 2 and 3. For the second research question, we attempted to identify writing features that distinguished IELTS writing proficiency levels for the following scoring criteria: coherence and cohesion, lexical resources, and grammatical range and accuracy. We found small positive correlations between the IELTS scoring criteria and the set of linguistic variables included in this study.

The results of five regression models revealed the linguistic features that most strongly predicted IELTS writing proficiency levels. In all five models, essay length was the strongest predictor, with R^2 values ranging from 18% to 25%. This finding is consistent with findings from previous research on writing quality (e.g., MacArthur et al., 2019). In addition to essay length, other significant predictors of overall writing scores were lexical sophistication, discourse complexity, and lexical diversity. This finding corroborated previous research findings that have shown a positive correlation between lexical diversity and ELL writing development (e.g., Malvern et al., 2004; Yu, 2010). In this study, we operationalised the discourse complexity as the amount of content word overlap in adjacent sentences in test-taker writing samples. As this feature predicted L2 writing development (Crossley et al., 2016) strongly, it also demonstrated a significant association with the IELTS writing scores.

Moreover, Task Response scores were also significantly predicted by lexical sophistication and grammatical complexity. Note that in this study, grammatical complexity was analysed through a Register–Functional Approach which involves individual lexico-grammatical structures (Biber et al., 2020). It particularly focused on the features of noun phrase complexity. As many studies (e.g., Parkinson & Musgrave, 2014; Taguchi et al., 2014) showed strong correlations between these features and L2 writing development, it is not surprising to see that they predicted IELTS writing scores strongly as well.

Coherence and Cohesion scores were significantly predicted by lexical sophistication, lexical diversity, and discourse complexity, after essay length was accounted for. According to the IELTS writing task band descriptors, the criterion of Coherence and Cohesion is described by logical sequences of information and ideas along with the good use of cohesive devices and the clear management of paragraphing. Therefore, it is not unexpected to see such lexical and discourse features predicting the scores. In addition to essay length, Lexical Resources was significantly predicted by discourse complexity, lexical sophistication, and lexical diversity. Given that Lexical Resources include band descriptors about a wide range of vocabulary with natural and sophisticated control of lexical features, these lexical features as significant predictors are somewhat predictable. These findings also support previous findings (e.g., Kim et al., 2018; Kyle & Crossley, 2016).

Finally, after essay length, significant predictors of Grammatical Range and Accuracy were discourse complexity, lexical diversity, and lexical sophistication. It was surprising to find that grammatical complexity was the only linguistic variable not to predict Grammatical Range and Accuracy scores. Overall, it seems that linguistic features can predict IELTS writing scores and sub-scores with moderate levels of variance accounted for.

5.3 Impact of learner background factors on learners' writing improvement

For the third research question, we aimed to determine how writing features change over time, and how these changes are explained by test-taker background variables and the linguistic variables. Essay length gain scores was a significant predictor in four of the five models. This confirms our findings for research question 2 regarding the importance of this variable. Essay length not only predicts test scores, but also predicts gains over time. It has long been known that this variable correlates with writing quality (e.g., Witte & Faigley, 1981). There is an ongoing debate about "whether text length is a construct-relevant aspect of writing competence or a source of judgment bias" (Fleckenstein et al., 2020). Scholars who take the latter position have argued that essay length is a variable that should be controlled for rather than analysed in research on writing quality (Jo, 2021). We have included essay length in this analysis in an effort to measure the degree to which it accounts for variance in writing scores that is unique from other constructs such as discourse complexity. The strong effect of essay length revealed in this study suggest the need for further research into this variable and its relationship to the purpose, constructs, and design of the IELTS writing exam.

Prior study abroad experience emerged as a significant predictor in three of the five models. This underscores the importance of extended learner immersion in ESL contexts. In fact, study abroad experience or language contact has demonstrated positive effects on learners' lexical development (see Collentine, 2004; Milton & Meara, 1995). Study abroad and ESL experiences can provide immersion in the target language which could facilitate the improvement of performance skills in general.

For the overall writing score gains, learners' proficiency, and study abroad experience contributed significantly. In particular, proficiency was negatively associated with writing gains, which means that the higher candidates' proficiency was, the less improvement they made. This finding is in line with Kang et al.'s (2021) findings that there was a restricted improvement in IELTS test-takers' speaking skills, especially among high proficiency learners. Proficiency also revealed a negative relationship with the Task Response sub-score gains. Surprisingly, the hours of study or the amount of target language use did not predict writing score gains.

Finally, Lexical diversity gain scores was also a significant predictor in two of the models, revealing that test-takers who increased the lexical diversity of their writing were more likely to experience gains in their test scores. This finding is in line with previous research that has shown that lexical complexity is a powerful predictor of L2 writing quality (Lee et al., 2021).

Overall, the results of the writing component of this study demonstrate that learners' writing development can be explained by a combination of linguistic variables, as well as variables related to learners' background characteristics. As expected based on previous research, text length was a strong predictor of writing scores. Additionally, other linguistic variables associated with lexical sophistication, discourse complexity, and lexical diversity accounted for variance in writing scores. The learner background characteristic that most strongly predicted learner writing development was prior study abroad experience, which highlights the importance of language exposure in ESL contexts.

6. Conclusion

Through this project, we have attempted to expand our understanding of language learning and progress by trying to answer such questions as: (1) what are the overall speaking features that distinguish IELTS speaking proficiency levels? (2) what are the overall writing features that distinguish IELTS writing proficiency levels? and (3) how do writing features change over time and how do background variables correlate with such linguistic progression of IELTS writing? However, predicting a language learning pattern is indeed not a simple process, as it can involve various unforeseen factors affected by individuals' personal, social, and environmental situations. Undoubtedly, the variables examined in this study are limited in scope and length. As seen in Kang et al.'s (2021) report, even though we used IELTS score gains as an indicator of improving language ability in that study, it may not necessarily mean evidence of a real gain in language proficiency. Despite these limitations, some implications can be drawn from the findings of this follow-up study, particularly regarding the proficiency level-related features for both writing and speaking skills.

According to Kang et al.'s (2021) findings, an intensive 12-week course of study in an EFL context did not bring substantial changes in IELTS writing band scores particularly if learners already had a high level of proficiency. Especially, advanced learners of English might need to be informed of the fact that score gains might be a bit slower at the upper levels than at the lower levels. Low-proficiency learners, however, may demonstrate measurable improvement in their overall writing score. This improvement can also provide the lower-proficiency learners with genuine motivation which can affect their general attitude towards study. One of the most important patterns of linguistic change was also how strongly proficiency was linked to various linguistic features in writing performances. As a result, language programs and institutions could consider offering diagnostic tests to assess students' initial proficiency levels before they start the program for this type of performance-based skills such as writing and speaking skills and offer level specific learning objectives and outcomes accordingly.

Also, speech rate was a strong predictor for the overall speaking scores as well as some of the sub-scores. This perhaps suggests that features related to speech rate can be improved somewhat more quickly than other sub-skills. Essay length itself was an important factor in predicting writing performances. Some of these features seemed to be more salient than others when it came to affecting learners' proficiency ratings. Overall, educators and test practitioners should keep in mind that language learning does not follow a linear and uniform relationship. As we have iterated a few times already in this report, it is a complex and unpredictable process. No single feature independently predicted proficiency, but it is a combination of all linguistic properties intertwined. Overall, the field would benefit from taking multi-dimensional approaches to better understand language learners and their progress, their needs and backgrounds, and their expectations as well as their learning behaviours.

REFERENCES

Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes*, 31, 58–71.

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

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.

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

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, 46, 100869.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2022). *The Register–Functional Approach to Grammatical Complexity: Theoretical foundation, descriptive research findings, application*. New York: Routledge.

Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge: Cambridge University Press.

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. *IELTS Research Report No. 6*. IELTS Australia, Canberra; British Council, London.

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*. Princeton, NJ: Educational Testing Service.

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

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

Caspers, J. (2010). The influence of erroneous stress position and segmental errors on intelligibility, comprehensibility and foreign accent in Dutch as a second language. *Linguistics in the Netherlands*, 27(1), 17–29. https://doi.org/10.1075/avt.27.03cas

Chapelle, C., Enright, M., Jamieson, J. (Eds.). (2008). *Building a validity argument for TOEFL*. New York: Routledge/Taylor and Francis Group.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.

Collentine, J. (2004). The effects of learning contexts on morpho-syntactic and lexical development. *Studies in Second Language Acquisition*, 26(2), 227–248.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237.

Csomay, E., & Prades, A. (2018). Academic vocabulary in ESL student papers: A corpus-based study. *Journal of English for Academic Purposes*, 33, 100–118.

Douglas, D., & Smith, J. (1997). Theoretical underpinnings of the Test of Spoken English revision project. *TOEFL Monograph Series*. Princeton, NJ: Educational Testing Services. Retrieved from https://www.ets.org/Media/Research/pdf/RM-97-02.pdf

Ejzenberg, R. (2000). The juggling act of oral proficiency: A psycho-Sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on Fluency* (pp. 287–313). Ann Arbor: The University of Michigan Press.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39(3), 399–423. https://doi.org/10.2307/3588487

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. *Frontiers in Psychology,* 11, 562462.

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, 28(3), 410–439.

Gardner, D., & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics*, 35(3), 305–327.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. London: Cambridge University Press.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly,* 38(2), 201–223. https://doi.org/10.2307/3588378

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *IELTS Research Reports Volume 2*, 52–63. IELTS Australia, Canberra; British Council, London.

Hirschberg, J. (2017). Pragmatics and prosody. In Y. Huang (Ed.). T*he Oxford Handbook of Pragmatics* (pp. 1–19). Oxford University Press.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3), 475–505. https://doi.org/10.1017/s0272263112000150

Iwashita, N., Brown, A., McNamara, T., O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How difficult? *Applied Linguistics*, 29, 24–49.

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

Jo, C. W. (2021). Short vs. extended adolescent academic writing: A cross-genre analysis of writing skills in written definitions and persuasive essays. *Journal of English for Academic Purposes*, 53, 101014.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38(2), 301–315.

Kang, O., Ahn, H., Yaw, K., & Chung, S. (2021). Investigation of relationship among learner background, linguistic progression, and score gain on IELTS. *IELTS Research Report Series, No. 1/21*. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. https://www.ielts.org/-/media/research-reports/ielts-rr_2021-1_kang-et-al.ashx

Kang, O., & Hirschi, K., Hansen, J., & Looney, S. (2022). *Characterization and normalization of second language speech intelligibility through lexical stress, speech rate, rhythm, and pauses.* Acoustic Society of America. TN.

Kang, O., & Johnson, D. (2018a). Contribution of suprasegmental to English speaking proficiency: Human rater and automated scoring system. *Language Assessment Quarterly,* 15(2), 150–168.

Kang, O., & Johnson, D. (2018b). Patent Serial No. 9,947,322; NAU Case 2013-015, *Systems and methods for automated evaluation of human speech*.

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly*, 48(1), 176–187.

Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. J*ournal of Language and Social Psychology*, 28, 441–456. https://doi.org/10.1177/0261927X09341950

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal*, 94(4), 554–566.

Kang, O., & Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing and Assessment*, 1(1), 24–39, https://doi.org/10.23977/langta.2018.11003

Kang, O., & Wang, L. (2014). Impact of different task types on candidates' speaking performances. *Research Notes*, 57, 40–49.

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, 102(1), 120–141.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2), 154–170.

Lee, C., Ge, H., & Chung, E. (2021). What linguistic features distinguish and predict L2 writing quality? A study of examination scripts written by adolescent Chinese learners of English in Hong Kong. *System*, 97, 102461.

Luoma, S. (2004). *Assessing Speaking*. Cambridge University Press: UK, Cambridge.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496.

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, 32(6), 1553–1574.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.

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

Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19(4), 477–496.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. https://doi.org/10.1111/0023-8333.49.s1.8

Munro, M. J. & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, 48(2), 159–182.

Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, 6(3), 283–309. https://doi.org/10.1075/jslp.20038.mun

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.

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

Stoddard, S. (1991). *Text and texture: Patterns of cohesion*. New Jersey: Ablex.

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(2), 420–430.

Wennerstrom, A. (1998). Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in Second Language Acquisition*, 20(1), 1–25. http://www.jstor.org/stable/44486381

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College composition and communication*, 32(2), 189–204.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2), 236–259.