Studies in Second Language Acquisition (2026), 1–28
doi:10.1017/S0272263126102046

RESEARCH ARTICLE

Exploring linguistic features in young foreign language learners' speech: *The interplay of learner background, educational context, and time in 6 months of development*

Mohammadreza Dalman[1], Okim Kang[2] and Maria Nelly Gutierrez Arvizu[3]

[1]English, Arizona State University, Tempe, AZ, USA; [2]English, Northern Arizona University, Flagstaff, AZ, USA and [3]Foreign Languages, Universidad de Sonora, Hermosillo, Sonora, Mexico
Corresponding author: Mohammadreza Dalman; Email: reza.dalman@asu.edu

(Received 29 July 2025 UTC; Revised 31 August 2026 UTC; Accepted 09 September 2026 UTC)

Abstract

The widespread early teaching of English in primary schools has prompted extensive research on young learners' proficiency and assessment. However, studies exploring young learners' development in spoken language features—especially in relation to background and contextual factors in EFL settings—are still limited. This longitudinal study examined the development of young foreign language learners (ages 10–12) in various linguistic dimensions of speaking (i.e., fluency, pronunciation, vocabulary, grammar, and content) over 6 months and investigated how these changes related to their educational context and background variables. A total of 102 young foreign language learners from three different educational contexts in Mexico took the speaking section of the TOEFL Primary® test at both the start and end of the 6 months. They also completed background questionnaires at the beginning, midpoint, and end of the study. Results indicated that learners in the advantaged educational contexts outperformed those with limited resources on most linguistic features of speaking, and that learners with lower initial proficiency made greater gains in all spoken linguistic features. These findings offer important insights for school administrators and curriculum developers.

Keywords: learner differences; linguistic features; oral proficiency; young learners

The large-scale introduction of English in primary schools is one of the most influential educational policy developments (Johnstone, 2009). This shift toward early English instruction has created a high demand for age-appropriate proficiency assessments to meet the needs of stakeholders, including schools and parents (Huang, Bailey, Sass, & Shawn Chang, 2021). The global increase in English language education and assessment for young learners has fueled growing research interest in these topics over the

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

past 25 years (see, e.g., Cho & Blood, 2020; De Wilde, Brysbaert, & Eyckmans, 2021; Lázaro-Ibarrola, 2024, 2025; Lázaro-Ibarrola & Roothooft, 2025).

Despite the surge in interest in teaching and assessing young learners, previous research has predominantly investigated young learners' overall proficiency gains over time across various language skills (e.g., Lázaro-Ibarrola, 2025; Lázaro-Ibarrola & Roothooft, 2025). While findings on overall proficiency gains provide evidence of learners' general improvement, they offer limited, if any, insights into the nature of that development. In fact, gains in overall proficiency entail improvement in specific linguistic features, which, in turn, exerts a positive influence on overall proficiency (see Saito & Plonsky, 2019). Fine-grained analyses of linguistic features of speaking reveal the underlying mechanisms of language growth (Norris & Ortega, 2009), especially for young learners whose linguistic development exhibits uneven progress across different dimensions of language (Muñoz, 2014). However, despite the importance of examining such fine-grained features, research on the developmental trajectories of young foreign language (FL) learners' linguistic features of speaking and their relationships with learner background characteristics from a longitudinal perspective is scarce. These features have been shown to differentiate young learners across different proficiency levels (Gu & Hsieh, 2019; Hsieh & Wang, 2019) and are amenable to the influence of individual differences (Muñoz, 2014). Yet, no research has systematically examined how growth in linguistic features of young learners' speaking unfolds in relation to background variables. Furthermore, the role of educational context in shaping young learners' development of spoken language features remains largely unexplored.

To address the existing gaps, the current investigation adopts a longitudinal approach to exploring the relationships between young FL learners' (aged 10 to 12 years) background and contextual factors and their progression in linguistic features of speaking over a semester in Mexico, an English as a Foreign Language (EFL) context that has received less attention in the literature compared with other South American countries (Ruiz de Zarobe & Banegas, 2024).

Literature review

Linguistic features as correlates of oral proficiency

The construct of second language (L2) oral proficiency is multi-componential, composed of a set of principal dimensions, including fluency, vocabulary, grammar, and pronunciation (De Jong et al., 2013). These dimensions distinguish learners across proficiency levels, including young learners who are the focus of the current study (Gu & Hsieh, 2019). This section reviews the literature on the contribution of linguistic features to L2 learners' oral proficiency, focusing primarily on 10–12-year-old learners.

Fluency

Oral fluency is both a predictor and a product of L2 proficiency (Suzuki & Kormos, 2025; Tavakoli, Nakatsuhara, & Hunter, 2020). In L2 fluency research, the temporal characteristics of speech are referred to as utterance fluency (Segalowitz, 2010), which encompasses several measurable aspects, including breakdown fluency, speed fluency, and repair fluency (Tavakoli & Skehan, 2005). *Breakdown fluency* refers to pausing behavior and can be measured by counting the number and length of filled and silent pauses. *Speed fluency* is concerned with the speed with which speech is delivered and can be measured by calculating measures of speech rate, including syllables per second, articulation rate, and mean length of run. *Repair fluency* is a speaker's false starts,

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

corrections, and repetitions during oral communication (De Jong et al., 2013; Suzuki, Kormos, & Uchihara, 2021). Empirical research on young learners' fluency development over time is relatively rare. Djigunović (2016) investigated the oral proficiency development of young Croatian learners (aged 11 at the start and 14 at the end of the study) over 4 years and found that fluency manifested a consistent growth over time and that the young learners' fluency ratings did not vary across different speaking tasks. Hsieh and Wang (2019) found that high-proficiency TOEFL Junior test takers (mean age 13.6) spoke significantly faster than their lower proficiency peers, and their speech featured shorter pause duration, relatively fewer filled and unfilled pauses, and longer stretches of chunks. These findings were further confirmed by Gu and Hsieh (2019), who reported that higher proficiency young learners aged 9–12 produced fewer pauses and longer stretches of speech than their lower proficiency peers.

Vocabulary

Vocabulary knowledge, including lexical range (the ratio of word types to word tokens) and lexical richness (the proportion of high- and low-frequency vocabulary), is associated with L2 oral proficiency (e.g., Brown, Iwashita, & McNamara, 2005; Iwashita, Brown, McNamara, & O'Hagan, 2008). Investigating young learners' TOEFL Junior spoken responses across four proficiency levels, Hsieh and Wang (2019) found that more proficient learners produced a wider range of words as well as words that were less familiar, more abstract, and harder to associate with an image. Comparable findings were echoed in Gu and Hsieh (2019), who found that more proficient young learners tended to use a wider range of words in their speech compared with their less proficient peers. Orosz (2009) examined young learners' (aged 9–12) vocabulary size in the first 4 years of English study at a Hungarian primary school. The results revealed that during the first 3 years of English learning, young learners' vocabulary size grew consistently every year, but the progress started to diminish in the fourth year. The findings also showed that the acquisition of more frequent words preceded the acquisition of less frequent vocabulary. This pattern was further supported by Roessingh and Elgie (2009), in which K-2 English language learners acquired high-frequency English words very quickly and developed basic literacy skills. However, their speech lacked low-frequency words usually used by native speakers. In fact, young learners relied heavily on the first 250 high-frequency words to communicate their meaning. Overall, the empirical findings support that high-proficiency young learners know more words and that high-frequency words are acquired earlier than less frequent vocabulary.

Grammar

The grammar measures of syntactic accuracy and complexity are key components of oral proficiency (Norris & Ortega, 2009). Syntactic accuracy is assessed at two levels: *global*, in which all errors are identified, and *specific*, in which only a particular set of errors is considered, such as prepositions, article use, and plural markers (Iwashita et al., 2008). Syntactic complexity is conceptualized as the characteristics of an utterance at the clausal level, such as the number of clauses per T-unit, the ratio of dependent clauses, and the verb phrase ratio (Iwashita et al., 2008). Hsieh and Wang (2019) and Gu and Hsieh (2019) found that the spoken responses of high-proficiency TOEFL Junior test takers featured fewer grammatical errors and more clauses per speech unit than the spoken responses of lower proficiency test takers. Investigating the global accuracy of young EFL learners' speech over 4 years, Djigunović (2016) observed a curvilinear

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

pattern in which young learners progressed consistently in grades 5 and 6. The progress dipped in grade 7 and then picked up again in grade 8.

Pronunciation

Among various pronunciation features, lexical stress (i.e., stressing the appropriate syllable in a given word) has been identified as the most robust predictor of non-native speakers' oral proficiency rating, surpassing other features such as fluency, segmentals, and tone choice (Kang, 2012). Lexical stress errors are also strongly associated with the intelligibility of accented L2 speech (Kang et al., 2020). Despite its importance in L2 speech, to the best of our knowledge, no studies have systematically examined lexical stress development using speech data from young EFL learners, and research on adult learners has yielded mixed results. For example, while Kang, Ahn, Yaw, and Chung (2021) found that the lexical stress accuracy of EFL learners did not improve significantly after participating in a 3-month IELTS preparation course, Kang and Kermad (2020) reported that adult learners significantly reduced their lexical stress errors after 3 months of immersion in an English as a Second Language (ESL) context. In this regard, the findings of the current study add to the limited existing body of research on lexical stress acquisition among young FL learners.

Content

Few empirical attempts have examined the content aspect of speech. In the domain of young language test takers, Hsieh and Wang (2019) examined the content quantity (i.e., the number of speech units and the number of clauses) and quality (i.e., the number of key points) of TOEFL Junior test takers and reported that both measures of content were influenced by proficiency level, in that more proficient test takers produced significantly more numbers of key points in the form of speech units and clauses compared to the less proficient ones. The current study extends the limited research base on the development of content quantity production among young FL learners.

Learner background variables and language proficiency development

Language proficiency development is shaped by various learner-related factors. Out-of-school exposure through frequent interactions with native speakers (e.g., Du, 2013), playing video games (e.g., Sylvén & Sundqvist, 2012), reading different texts (e.g., Montag & MacDonald, 2015), or exposure to various media (e.g., Peters, 2018) enhances the acquisition of different aspects of L2 proficiency, including fluency, vocabulary, and syntactic complexity and accuracy (Muñoz, 2014). Additionally, the length of instruction, measured as the number of years of education inside and outside of school, has been shown to predict L2 learners' gains in vocabulary (e.g., Peters, 2018; Unsworth, Persson, Prins, & De Bot, 2014), syntactic complexity (e.g., Muñoz, 2014), lexical stress accuracy (e.g., Saito & Hanzawa, 2016), and fluency (e.g., French, Gagné, & Collins, 2020). L2 learning experience and assessment practices have also been shown to impact the quality and quantity of L2 learning as well as assessment outcomes. Positive L2 learning experience is associated with increased motivation, greater metacognitive awareness, and a broader range of learning strategies among young learners, all of which contribute to academic achievement (Kim & García, 2014; Wolf, 2024). Classroom assessment practices, including frequency, mode, and task type used by language teachers to evaluate students' progress, play a critical role in shaping instruction, influencing learner behavior,

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

and ultimately affecting L2 learning outcomes (Latif & Wasim, 2024; Yu & Xu, 2021). In fact, frequent and multimodal assessment practices yield ample opportunities for L2 production and targeted feedback (see, e.g., Gass et al., 2020). Given the established link between assessment practices and opportunities to learn, incorporating a measure of students' assessment experiences will deepen our knowledge about how variations in L2 learning outcomes can be attributed to assessment-related practices. Prior L2 knowledge is likewise reported to influence language proficiency growth. De Wilde et al. (2021) found that the initial level of proficiency was a robust predictor of young learners' (aged 10–12) receptive vocabulary and speaking proficiency gains, explaining 58% and 72% of the variance. Unsworth et al. (2014) reported that prior knowledge was the best predictor of young EFL learners' (mean age 10.37) receptive vocabulary and grammar proficiency growth. Jackel, Schurig, Florian, and Ritter (2017) also indicated that proficiency accounted for 35% of the variance in young FL learners' scores in receptive skills.

Educational context and language proficiency development

L2 learners' proficiency gains are likely to vary as a function of the educational context in which they study (Lázaro-Ibarrola, 2024, 2025; Lázaro-Ibarrola & Roothooft, 2025; Lorenzo, Granados, & Rico, 2021). Such contexts can be broadly categorized into English as a Medium of Instruction (EMI), where English serves primarily as the language to teach academic content; Content Language Integrated Learning (CLIL), which places equal emphasis on language and academic content; and English as a Foreign Language (EFL), in which language is the main subject of instruction. In a recent meta-analysis, Lee, Lee, and Lo (2023) found a moderate overall advantage of EMI-CLIL programs over mainstream education settings. Their findings further revealed that EMI-CLIL instruction was more effective in improving receptive skills than productive skills. Complementing this evidence, numerous primary studies have shown that EMI-CLIL is associated with improvements in overall speaking (e.g., Lázaro-Ibarrola, 2012, 2025; Merino & Lasagabaster, 2018) and reading proficiency (Lázaro-Ibarrola & Roothooft, 2025) as well as in specific linguistic features of speaking, including vocabulary (e.g., Azpilicueta-Martínez, 2025), grammar complexity, and subordination/dependent clauses (e.g., Azpilicueta-Martínez, 2024), and fluency (e.g., Pérez Cañado & Lancaster, 2017). In contrast, pronunciation has emerged as a feature that appears less sensitive to instructional settings (see, e.g., Pérez Cañado & Lancaster, 2017; Rallo Fabra & Jacob, 2015). These findings, however, should be interpreted with caution, as EMI-CLIL is typically characterized by a higher level of exposure to the FL, making it difficult to discern whether improvements can be ascribed to the EMI-CLIL factor or just to the additional exposure. Against this backdrop, the present study explored variations in gains in linguistic features of speaking among Mexican young learners across three educational contexts—Bilingual, Private, and Public schools—where the first two represent the EMI-CLIL models and the latter follows an EFL model.

The current study

The current study is part of a larger project examining the relationship between young FL learners' background variables, educational context, and proficiency gains on the TOEFL Primary test over time. Our previous findings showed that speaking was the only skill that improved significantly over 6 months, reflecting a CEFR transition from A1 at T1 to A2 at T2. These gains in speaking proficiency were influenced by learners'

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

educational context and initial proficiency, which further indicated that progress was not uniform across all learners. Building on that, the present study specifically explores the development of linguistic features of speaking, including fluency, grammar, vocabulary, pronunciation, and content, and their relationship with learner background (i.e., hours of study, target language use, English learning experience, assessment practices, and initial proficiency) and contextual factors.

Research on young learners' (aged 10–12) development of spoken linguistic features remains scarce, particularly from a longitudinal perspective. This age range is theoretically significant because previous research (e.g., Hartshorne, Tenenbaum, & Pinker, 2018) suggests that L2 learning ability begins to decline after age 12, making this age range theoretically informative. Moreover, young learners at this age may possess more cognitive advantages, such as superior working memory compared to much younger learners, which can facilitate L2 learning/acquisition (see Swanson, 2015). Finally, the 6-month timeframe adopted in this study aligns with prior recommendations suggesting that a 6-month interval between test administrations is appropriate for young learners to demonstrate gains in their scores (Educational Testing Service [ETS], 2023; see also Cho & Blood, 2020; Gu, Lockwood, & Powers, 2015).

In the current study, participating schools assessed young learners' English proficiency using school-designed proficiency tests. Although those in-house assessments share some similarities with the TOEFL Primary test, none of the students had prior exposure to the TOEFL Primary. We selected the TOEFL Primary test because it is a standardized instrument specifically developed for young learners (as young as 8 years old) whose proficiency ranges from A1 to B2 (ETS, 2023), making it appropriate for capturing developmental trajectories in our sample. Analyses focus on speaking tasks 5 and 6 of the TOEFL Primary test, picture description and narration, which elicit comparably longer, more detailed responses, providing a richer sample for examining linguistic development. The selection of linguistic features is informed by prior research demonstrating that they distinguish young learners across proficiency levels (Gu & Hsieh, 2019; Hsieh & Wang, 2019), as well as by the authors' pilot studies, which showed their potential to capture developmental change. Background variables were chosen based on previous research with young learners in FL contexts (e.g., De Wilde et al., 2020, 2021; Muñoz, 2014).

The current study was conducted at three primary schools in Sonora, Mexico: (a) a Public school (low-exposure EFL), (b) a Private school (EMI/low-CLIL), and (c) a Bilingual school (immersion/high-CLIL). The English instruction program in the Public school consists of 2.5 hr per week, spread across three 50-min sessions, with the curriculum relying primarily on state-funded textbooks. The Public school does not charge tuition and generally serves students from lower socioeconomic status families. The Private school offers 7.5 hr of English instruction per week, divided into five sessions of English, where the language is taught for communicative purposes, and 12 sessions of Science class in English (content-based instruction). The Private school typically serves students from middle to high socioeconomic status families. In the Bilingual/bicultural school, English immersion serves as the learning context for students. Core content classes are taught in English for a total of 30 hr of exposure to English every week. Due to high tuition and enrollment fees, this school mainly serves students from higher socioeconomic status families.

Although teaching oral skills is often neglected in Public schools in Mexico, Private and Bilingual schools prioritize oral skills through a variety of strategies designed to provide students with practice and build their confidence in using English. Common classroom activities include role-playing, singing, drilling (e.g., repeating after the

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

teacher or a model, modifying words in a sentence), describing and predicting, interviewing and reporting, telling or retelling stories and anecdotes, and rephrasing ideas using different vocabulary, among others. Teachers view it as important to provide constructive feedback while supporting students and offering them encouragement and praise, with the goal of boosting their self-confidence. Additionally, the use of the internet, videos, or other online tools is becoming an integral part of instruction. Overall, students are motivated to use and practice English actively during class time (School coordinators, personal communication, July 9, 2025).

Research questions

The present study addresses the following research questions:

1. How do young FL learners' linguistic changes in the picture description and narration tasks of the TOEFL Primary speaking test differ across educational contexts (i.e., Bilingual [immersion/high-CLIL], Private [EMI/low-CLIL], and Public school [low-exposure EFL])?
2. To what extent do learner background variables (e.g., hours of study, target language use, English learning experience, assessment practice, initial proficiency) predict young FL learners' linguistic changes in speaking in the picture description and narration tasks of the TOEFL Primary speaking test across educational contexts (i.e., Bilingual [immersion/high-CLIL], Private [EMI/low-CLIL], and Public school [low-exposure EFL]) and why?

Hypotheses

H1: Based on previous research on CLIL-oriented instruction, Bilingual and Private school learners are expected to exhibit comparatively greater gains across several linguistic dimensions of speaking than Public school learners.

H2: Drawing on previous literature, it is expected that learner background variables significantly predict young learners' development in linguistic features of speaking over time.

Method

The study adopted a longitudinal research design, with data collection occurring at three stages: at the onset (T1), at midpoint, and at the end (T2) of the study. At T1, after completing informed consent, participants responded to a set of online background questionnaires that collected information about (a) participants' demographics, (b) the number of hours they spent studying English, (c) the amount of target language use they had, and (d) participants' perceptions of their English learning experience. After completing the online background questionnaires, participants took the official computer-based TOEFL Primary speaking test provided by ETS. At midpoint (3 months after T1), the participants completed the online hours of study and target language use questionnaires for a second time. At T2 (6 months after T1), participants completed another set of online background questionnaires that included (a) hours of study, (b) target language use, and (c) assessment practices at school. After completing the online questionnaires, the participants took *the same* computer-based TOEFL Primary speaking test for a second time. Two rounds of oral interviews were also conducted with learners and

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

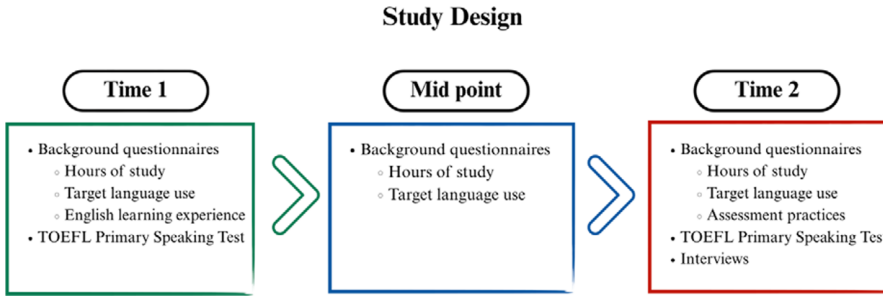


Figure 1. Overall design of the study.

teachers at the end of the study, and the recorded interviews were transcribed for further analysis. Figure 1 provides a summary of the study design.

Participants

The study included two groups of participants: learners and teachers. The learner participants were 110 young FL learners from nine sixth-grade classes in Sonora, Mexico. The learners were 55 females and 55 males, ranging in age from 10 to 12 years, with an average age of 10.98 (SD = 0.40). Due to incomplete speaking test data at either T1 or T2 from eight learners (7.3%), the final dataset comprised 102 learners. The learners were particularly recruited from three pedagogically distinct elementary schools: 36 were from three classes at a Bilingual school, 33 were from four classes at a Private school, and the remaining 33 were from two classes at a Public school. All learners were native speakers of Spanish, and 72.73% self-reported speaking only Spanish at home, while 22.72% reported speaking a combination of Spanish and English. The learners' English speaking proficiency fell within pre-A1 to B2 as indicated by their performance on the speaking section of the TOEFL Primary test (see Table 1 for a breakdown of the learners' demographics across the three learning contexts). Six teachers were also recruited from the respective schools (Bilingual school, n = 3; Private school, n = 2; Public school, n = 1) to participate in oral interviews at the end of the study. The teachers were five females and one male, ranging in age from 28 to 45 years, with an average age of 37 years (SD = 6.75). The teachers were all Spanish L1 speakers and self-reported an average of 14.83 years of teaching experience (SD = 6.08).

Variables and instruments

The study variables included (1) young learners' scores on a range of linguistic features extracted from responses to the TOEFL Primary picture description and narration tasks:

Table 1. Learners' demographic characteristics across the three learning contexts

Educational context	*N*	Hours of English instruction per week	Speaking proficiency
Bilingual	36	30	B1–B2
Private	33	7.5	A2–B1
Public	33	2.5	Below A1–A1

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

(a) fluency (speech rate, silent pauses), (b) pronunciation (lexical stress), (c) vocabulary (lexical resource), and (d) grammar (syntactic complexity, accuracy, and dependent clause ratio); (2) learning context (Bilingual, Private, or Public schools); and (3) background variables, including (a) initial proficiency (T1 scores on the linguistic features), (b) hours of English study, (c) target language use, (d) perceptions of English learning, and (e) assessment practices.

The TOEFL Primary speaking test

The *same* official computer-based TOEFL Primary speaking test was administered twice, at the beginning (T1) and at the end of the 6 months (T2), to assess young learners' improvement in the linguistic features of speaking. The speaking section of the test included eight questions and took approximately 20 min to complete. The questions assessed young learners' ability to perform a variety of tasks, ranging from seeing and describing parts of a picture to telling a story based on watching an animation. The speaking score range was 0–27, which corresponds to A1–B2 levels (ETS, 2023). For this study, learners' responses to questions 5 and 6 were selected for speech analysis, as these prompts elicited comparatively longer utterances. Question 5 was a picture description task in which the learners saw a picture, and then a character asked them to speak for 20 s, describing parts of the picture. Question 6 was a narration task in which learners watched a short animation and then were asked to speak for 30 s, telling a story based on what they had seen.

Background questionnaires

Online background questionnaires were administered to all learners at T1, midpoint, and T2 via Qualtrics (see Supplementary Appendix A). The questionnaires solicited information about a range of background variables that could potentially predict young learners' score gains in the linguistic features of speaking. These variables included young learners' demographics, overall hours of English study, amount of target language use, perceptions of English learning experience, and assessment-related practices. To ensure that young learners could comprehend the survey questions, the questions were presented in English and Spanish.

Hours of study. Hours of study was measured through nine items, which were administered to the learners at T1, midpoint, and T2. The items were adapted from Kang et al. (2021) and asked about the total number of hours young learners spent studying in class and outside of class last week. The items targeted attending English class; doing homework; going to language school; studying alone and with others outside of class; and practicing language skills, including reading, listening, speaking, and writing. For each item, participants had 10 answer options to choose from, ranging from 0 hr to more than 16 hr. Composite scores from these nine items across T1, midpoint, and T2 were used for analysis.

Target language use. The amount of target language use was measured through 11 items in which young learners reported on the hours of weekly English contact and exposure they had outside of the study context. The items were adapted from Freed, Dewey, Segalowitz, and Halter (2004) and have been used by Kang et al. (2021). These items focused specifically on using English when communicating with native speakers (NSs), non-native speakers (NNSs), family members, and when playing online games. They also measured exposure to English through watching TV, movies, and YouTube videos, as well as listening to music, using social media, and reading for pleasure. The items included 10 answer options that ranged from 0 hr to more than 16 hr and were

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

administered to the learners at the T1, midpoint, and T2 phases of the study. Composite scores from the 11 items across T1, midpoint, and T2 were used for analysis.

Learners' perceptions of English learning experience. Learners' perceptions were measured through five items adapted from Kim and García (2014) and Hunt (2011). On a Likert scale ranging from 1 (not true at all) to 10 (very true), the items asked about the extent to which young learners enjoyed their English classes as well as the activities and homework they completed in their English classes, the extent to which they had clear English learning objectives, and to what extent they were satisfied with their improvements in their English classes. The measure was administered only at T1, and composite scores from these five items were used for analysis.

Assessment practices. Assessment practices were measured through four items adapted from Gu and So (2017). The items asked learners about the frequency, type, and mode of English tests (individual vs. group projects) they completed in class and at school throughout the semester. The measure was administered only at T2, and composite scores from these four items were used for analysis.

Oral interviews

To gain additional insights into learners' progress over the 6-month study period and to complement the quantitative findings, two sets of semi-structured oral interviews (see Supplementary Appendix B) were conducted at the end of the study with six teachers (Bilingual school, $n = 3$; Private school, $n = 2$; and Public school, $n = 1$) and 15 learners (Bilingual school, $n = 6$; Private school, $n = 5$; and Public school, $n = 4$). The interview participants were selected randomly from each school type to ensure representation of the target population. Informed consent was obtained from all interview participants and the young learners' guardians where applicable. Each interview session lasted for approximately 30 min. The learner interview questions explored factors that could have influenced the language development during the study period, whereas the teachers' interview questions solicited information about classroom practices and contextual factors that could help explain variation in the learner outcomes. Questions were presented in English; however, the participants had the option of responding in English or Spanish. Responses to the oral interview questions were transcribed by the third author, whose L1 is Spanish.

Data analysis

Data analysis involved transcribing the speech files and coding them for a range of linguistic features through manual coding and automatic feature extraction. The data were then submitted to a set of bivariate and multivariate statistical analyses to answer the research questions.

Linguistic analysis

Learners' responses to the picture description and narration tasks were transcribed orthographically using a consistent transcription convention (Biber, Conrad, & Cortes, 2004), with all repetitions, false starts, and disfluency features such as "um" and "uh" removed to ensure clarity and consistency in the linguistic analysis. At T1, responses ranged from 5 to 20 s ($M = 11.74$, $SD = 4.08$) for the picture description task and from 5 to 30 s ($M = 17.40$, $SD = 6.31$) for the narration task; at T2, they ranged from 3 to 20 s ($M = 12.45$, $SD = 7.74$) for picture description and from 5 to 30 s ($M = 17.74$, $SD = 6.68$)

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

for narration. Responses from speaking tasks 5 and 6 were combined for all subsequent analyses. To establish intercoder reliability for transcriptions, the second author retranscribed 10% of the total data. Intercoder reliability between the first and second authors' transcriptions of the same 10% subset of the data was assessed using a two-way consistency, average-measures intraclass correlation coefficient (ICC), which indicated excellent reliability (ICC = .99). Upon completing transcriptions, the speech files were analyzed for the following linguistic features: (a) fluency, (b) pronunciation, (c) vocabulary, (d) grammar, and (e) content quantity.

Fluency measures included (a) speech rate, (b) number of silent pauses, and (c) mean length of silent pauses. Speech rate was calculated as the total number of syllables divided by the total speech length, including pauses (Kang et al., 2021). Number of silent pauses was computed as the total number of silent pauses longer than 0.30 s, and the mean length of silent pauses was computed as the total length of silent pauses divided by the total number of silent pauses. The fluency features were extracted automatically using a Praat script by de Jong and Wempe (2009).

For pause-related measures, learners who attended the speaking test but produced no analyzable speech were assigned a value of zero. This coding decision was made to distinguish cases of non-production attributable to limited speaking ability from missing data resulting from absence from the speaking test. Accordingly, zero values were treated as valid observations reflecting an inability to produce speech rather than missing responses. Learners who were absent from the assessment were treated as missing data and were not assigned pause values.

Lexical stress errors were identified as the total number of deviations divided by the total number of polysyllabic words and were coded manually and instrumentally using Praat (Boersma & Weenink, 2016). Lexical stress deviations were coded by the first author, and to establish the reliability of data coding, 10% of the data was coded by the second author. The intercoder reliability was acceptable, ICC = .87.

The vocabulary measure—lexical resource—was a composite of a set of individual features, including (a) the proportion of K1 words (first 1,000 most frequent word families), (b) the proportion of K2 words (second 1,000 most frequent word families), (c) tokens (the total number of words produced in the speech sample), (d) types (the range of unique words produced in the speech sample), and (e) type-token ratio (the total number of word types divided by the total number of word tokens). Features of the lexical resource were generated using the *Vocabprofile* function in *Lextutor* (Cobb, 2024), based on written transcripts (see also Brown et al., 2005; Gu & Hsieh, 2019). The feature of Coxhead's academic word list (AWL) was excluded, as the speaking tasks did not prompt learners to use academic words.

Grammar measures encompassed (a) syntactic complexity, (b) syntactic accuracy, and (c) dependent clause ratio. To analyze the speech samples for grammar measures, the transcriptions were first coded for (a) the number of C-units (an independent clause with its modifiers), (b) the number of clauses (a statement containing both a subject and a predicate), and (c) the number of verb phrases. Syntactic complexity was measured as a composite of (a) C-unit complexity (the number of C-units divided by the number of clauses) and (b) verb phrase ratio (the number of C-units divided by the number of verb phrases) (Gu & Hsieh, 2019; Kang et al., 2021). Syntactic accuracy was measured globally as the number of error-free C-units divided by the total number of C-units (Gu & Hsieh, 2019). The dependent clause ratio was measured as the number of dependent clauses divided by the total number of clauses. Grammar measures were coded manually by the first author, and to establish the reliability of data coding, 10% of the data was coded by the

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

second author. Intercoder reliability was acceptable for syntactic complexity, ICC = .88, and syntactic accuracy, ICC = .86, and excellent for dependent clause ratio, ICC = .96.

To measure content quantity, in line with previous research (Brown et al., 2005), the number of C-units in the speech samples was used instead of the number of words, as the number and variety of words would not reliably represent the number of ideas presented.

Statistical analysis

All quantitative data were analyzed in R 4.5.0. Variables for which lower scores indicated higher performance, such as lexical stress, number of pauses, mean length of pauses, C-unit complexity, and verb phrase ratio, were directionally transformed so that higher values consistently represented better performance.[1] Pearson correlations were then computed based on raw scores to assess collinearity among linguistic variables and group highly correlated, conceptually related features into composite speech constructs. Variables intended to represent the same construct showed very strong intercorrelations (vocabulary measures $r > .75$–$.96$; syntactic complexity measures $r > .80$–$.99$). Composite variables were created by first standardizing the component variables as z-scores and then averaging the standardized values within each domain construct. The formulae used to derive the composite scores are as follows:

- Lexical resource = (types [z] + tokens [z] + type/token ratio [z] + K1 proportion [z] + K2 proportion [z]) / 5.
- Syntactic complexity = (reverse-coded C-unit complexity [z] + reverse-coded verb phrase ratio [z]) / 2.

Given that grammar accuracy and dependent clause ratio were weakly related to other grammar measures ($r < .35$, $r < .20$, respectively), they both served as separate dependent variables. Thus, the final analysis included nine dependent variables grouped into five categories: fluency (speech rate, number of pauses, mean length of pauses), pronunciation (lexical stress), vocabulary (lexical resource), grammar (syntactic complexity, syntactic accuracy, and dependent clause ratio), and content quantity.

The first research question investigated whether young learners' linguistic changes in speaking differed as a function of learning contexts. Given that the three groups were not equal at T1, a set of nine ANCOVA tests was used, with T1 scores treated as covariates, learning context as an independent variable, and T2 scores on each one of the linguistic features as the dependent variable. Before conducting each ANCOVA, the homogeneity of regression slopes was checked to ensure that the relationship between the dependent variables and the covariate was the same in each category of the independent variable. The Tukey HSD test was used for post hoc analyses for each statistically significant model. To interpret the effect sizes (Cohen's d values) for post hoc comparisons, we followed Plonsky and Oswald's (2014) guidelines with lower thresholds of $d = .40$ as a small effect, $d = .70$ as a medium effect, and $d = 1.00$ as a large effect. Descriptive statistics were reported for all linguistic variables at both T1 and T2,

[1]As recommended by Tabachnick and Fidell (2019), variables for which lower scores indicated stronger performance were directionally transformed by multiplying scores by –1, so that higher values consistently represented better performance. Where composite variables were created, transformed variables were standardized prior to aggregation.

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

including both individual and composite variables. To answer the second research question regarding the effects of learner backgrounds on linguistic changes in speaking, four sets of standard multiple linear regressions were computed. Descriptive statistics included means and standard deviations for background variables. Regression assumptions were checked following the recommendations by Field, Miles, and Field (2012), and none of the assumptions were violated.

To analyze the interview data, a reflexive thematic coding scheme was developed inductively following Braun and Clarke's (2006) six-phase framework. The process involved (a) familiarizing oneself with interview data through repeated readings, (b) creating initial codes systematically across the dataset, and (c) organizing the codes into broader categories to form preliminary themes. The themes were then reviewed for coherence and alignment with the data, refined and named to capture their distinctiveness, and finally illustrated with representative quotes. Given that the interview questions were straightforward and the answers they elicited lent themselves to suitable categorization, no interrater reliability check was conducted.

Results

This section is divided into two subsections. The first subsection presents quantitative findings for each research question, followed by the second subsection, which reports the qualitative findings organized by research questions.

Quantitative findings

RQ1: How do young FL learners' linguistic changes in the picture description and narration tasks of the TOEFL Primary speaking test differ across educational contexts (i.e., Bilingual [immersion/high-CLIL], Private [EMI/low-CLIL], and Public school [low-exposure EFL])?

Tables 2 and 3 summarize the unadjusted descriptive statistics for the linguistic features of speaking across the three learning contexts at T1 and T2. Overall, Bilingual and Private school learners scored higher at T1 across most linguistic features, while Public school learners generally exhibited lower initial mean scores and greater variability. Changes in raw mean scores from T1 to T2 varied across learning contexts and linguistic features and did not consistently align with the ANCOVA results reported below, in which T2 scores were adjusted after partialling out differences in T1 performance. Figure 2 illustrates adjusted T2 scores for all linguistic features across the three learning contexts.

The results of ANCOVA showed that, for speech rate, T1 scores were significantly related to the learners' T2 scores, $F(1, 98) = 4.18$, $p = .04$, and after partialling out the effect of T1 scores, a significant difference was observed in learners' T2 speech rate performance across the learning contexts, $F(2, 98) = 17.71$, $p < .0001$, $R^2 = .71$. Post hoc analyses revealed significant differences with medium-to-large effect sizes between Bilingual and Private ($p = .01$, $d = .68$), Bilingual and Public ($p < .0001$, $d = 2.76$), and Private and Public school learners ($p < .0001$, $d = 2.08$). These findings indicate that, after controlling for T1 performance, Bilingual and Private school learners produced faster speech at T2 than Public school learners, and that Bilingual school learners also outperformed Private school learners.

For number of silent pauses, T1 scores were significantly related to learners' T2 performance, $F(1, 98) = 6.55$, $p = .01$. A significant effect of learning context on learners'

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

Table 2. Unadjusted descriptive statistics for individual linguistic features of speaking across the three learning contexts

		Bilingual				Private				Public			
		T1		T2		T1		T2		T1		T2	
Category	Variables	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Fluency													
	Speech rate	2.83	.43	2.69	.40	2.24	.72	2.20	.68	.11	.48	.25	.85
	Pauses	−10.48	3.75	−7.73	3.18	−9.35	4.40	−8.62	3.95	.93	3.93	.34	1.75
	Mean length of pauses	−1.31	4.44	−.60	.15	1.42	4.74	.68	.25	.04	.16	.03	.14
Pronunciation													
	Lexical stress	−.04	.07	−.001	.007	.000	.000	−.004	.02	.000	.000	.000	.000
Vocabulary													
	K1 words	.84	.04	.83	.13	.78	.21	.81	.22	.04	.18	.09	.28
	K2 words	.08	.03	.09	.03	.09	.04	.08	.05	.005	.02	.008	.03
	Types	36.86	8.54	33.79	9.33	22.37	11.34	25.24	10.55	.86	3.60	1.81	6.74
	Tokens	69.72	18.89	64.97	18.12	42.88	24.35	45.60	21.94	1.55	6.52	3.03	12.20
	Type/token ratio	.54	.06	.55	.06	.55	.18	.54	.16	.03	.13	.08	.24
Grammar													
	C-unit complexity	−.92	.10	−.92	.10	−.92	.24	−.90	.24	−.05	.23	−.07	.26
	Verb phrase ratio	−.91	.10	−.91	.10	−.91	.24	−.89	.24	−.05	.23	−.07	.26
	Global accuracy	.66	.23	.75	.18	.41	.30	.48	.33	.02	.12	.04	.18
	Dependent clause ratio	.10	.17	.08	.10	.02	.06	.04	.07	.000	.000	.004	.02
Content													
	Content quantity	10.69	2.45	10.16	2.59	7.54	3.53	7.76	3.30	.19	.82	.48	.20

Table 3. Unadjusted descriptive statistics for selected composite variables across the three learning contexts

	Bilingual				Private				Public			
	T1		T2		T1		T2		T1		T2	
Variables	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Syntactic complexity	–.64	.21	–.67	.23	–.65	.54	–.61	.53	1.24	.51	1.20	.58
Lexical resource	.75	.20	.71	.22	.42	.48	.42	.48	–1.14	.36	–1.05	.54

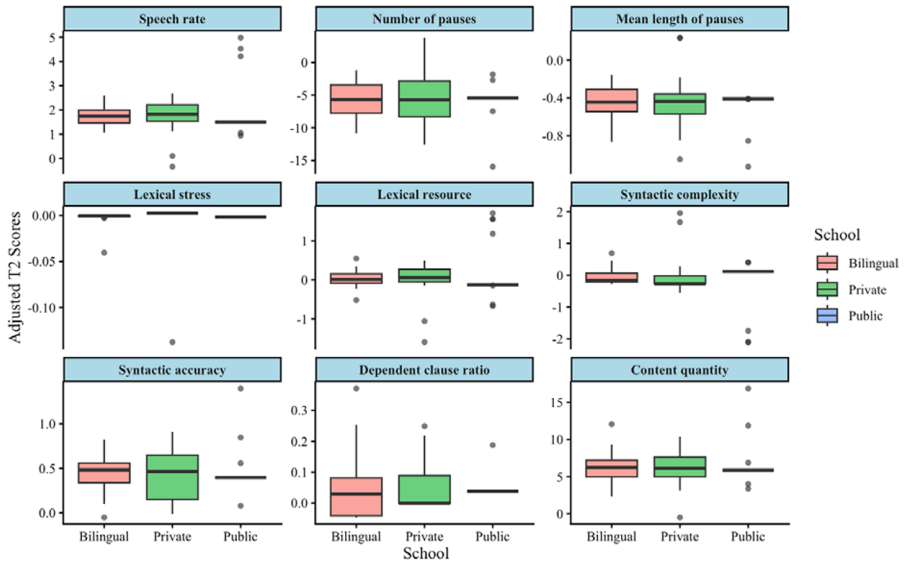


Figure 2. Summary of results for the linguistic features score changes before and after 6 months across the three educational contexts.

adjusted T2 scores was also observed, $F(2, 98) = 23.58$, $p < .0001$, $R^2 = .60$. Post hoc analyses revealed significant differences with large effect sizes between Bilingual and Public ($p < .0001$, $d = 1.82$) as well as Private and Public school learners ($p < .0001$, $d = 2.19$), whereas no significant difference was observed between Bilingual and Private school learners ($p = .29$, $d = .36$). A comparable pattern was observed for mean length of pauses. After controlling for T1 scores, a significant effect of learning context on learners' adjusted T2 scores was observed, $F(2, 98) = 112.55$, $p < .0001$, $R^2 = .69$. Post hoc comparisons revealed significant differences with large effect sizes between Bilingual and Public ($p < .0001$, $d = 3.01$) as well as Private and Public school learners ($p < .0001$, $d = 3.40$), whereas no significant difference was observed between Bilingual and Private school learners ($p = .25$, $d = .39$). These results indicate that, after controlling for T1 scores, the Public school learners obtained lower pause frequency and pause duration scores at T2 than the Bilingual and Private school learners. However, this finding should be interpreted with caution. For pause-related measures, learners who attended the speaking assessment but produced no analyzable speech were assigned

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

pause values of zero. As a result, the lower pause scores observed for the Public school group may partly reflect limited or absent speech production rather than greater fluency. Consequently, lower pause frequencies and shorter pause durations should not be interpreted as evidence of superior fluency, but rather as an outcome that may be influenced by restricted speech production among some learners in this group.

For lexical stress, T1 scores were not associated with learners' scores at T2, $F(1, 98) = .06$, $p = .79$, and no significant effect of context was observed in the learners' T2 scores after partialling out the effect of T1, $F(2, 98) = .73$, $p = .48$, $R^2 = .01$.

For lexical resource, T1 scores significantly predicted learners' adjusted T2 scores, $F(1, 98) = 9.34$, $p = .002$. After controlling for T1 performance, a significant effect of learning context on the learners' T2 lexical resource scores was observed, $F(2, 98) = 11.03$, $p < .0001$, $R^2 = .79$. Post hoc analyses revealed significant differences with large effect sizes between Bilingual and Public ($p < .0001$, $d = 2.64$) and Private and Public school learners ($p < .0001$, $d = 2.23$), whereas no significant difference was observed between Bilingual and Private school learners ($p = .25$, $d = .41$). These results suggest that after adjusting for T1 scores, Bilingual and Private school learners produced a wider range of lexical items at T2 than Public school learners.

For syntactic complexity, T1 scores were not significantly related to T2 performance, $F(1, 98) = 1.40$, $p = .24$. A significant effect of learning context on adjusted T2 scores was observed, $F(2, 98) = 24.38$, $p < .0001$, $R^2 = .76$. Post hoc analyses revealed significant differences with large effects between Bilingual and Public school ($p < .0001$, $d = -3.33$) and Private and Public school ($p < .0001$, $d = -3.21$), whereas no significant difference was observed between Bilingual and Private school learners ($p = .88$, $d = -.11$). **These results suggest that Public school learners exhibited higher syntactic complexity at T2 than both Bilingual and Private school learners.** However, these findings should be interpreted with caution since **this pattern may reflect shorter utterances yielding higher ratio-based complexity scores rather than qualitatively more advanced syntactic production.**

For syntactic accuracy, T1 scores were significant predictors of learners' adjusted T2 performance, $F(1, 98) = 19.37$, $p < .0001$. After partialling out differences at T1, a significant effect of learning context on learners' syntactic accuracy scores at T2 was observed, $F(2, 98) = 13.92$, $p < .0001$, $R^2 = .65$. Post hoc analyses indicated significant differences with medium-to-large effect sizes between Bilingual and Private ($p = .02$, $d = .73$), Bilingual and Public ($p < .0001$, $d = 1.92$), and Private and Public school learners ($p < .0001$, $d = 1.19$). These findings show that after controlling for T1 performance, Bilingual school learners produced more accurate speech at T2 than both Private and Public school learners.

Regarding dependent clause ratio, T1 scores were not significantly related to learners' adjusted T2 scores, $F(1, 98) = 2.39$, $p = .12$. However, a significant main effect of learning context on learners' T2 performance was observed after controlling for T1 scores, $F(2, 98) = 10.01$, $p < .0001$, $R^2 = .14$. Post hoc analysis indicated significant differences with medium-to-large effect sizes between Bilingual and Private ($p = .024$, $d = .63$) and Bilingual and Public ($p < .0001$, $d = 1.16$), whereas the difference between Private and Public school learners did not reach statistical significance ($p = .08$, $d = .52$). These findings suggest that after adjustment, Bilingual school learners produced a higher proportion of dependent clauses at T2 than both Private and Public school learners.

For content quantity, T1 scores were significant predictors of learners' adjusted T2 performance, $F(1, 98) = 49.93$, $p < .0001$. After controlling for T1 scores, a significant effect of learning context on learners' T2 content quantity was observed, $F(2, 98) = 4.97$, $p = .009$, $R^2 = .79$. Post hoc analyses revealed significant differences with large effect

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

sizes between Bilingual and Public school (p = .01, d = 1.38) and Private and Public school learners (p = .006, d = 1.18), whereas no significant difference was observed between Bilingual and Private school learners (p = .75, d = .20). These findings indicate that, after adjustment, both Bilingual and Private school learners produced more content, operationalized as the number of C-units, at T2 than Public school learners.

RQ2: To what extent do learner background variables predict young FL learners' linguistic changes in speaking in the picture description and narration tasks of the TOEFL Primary speaking test across educational contexts? and why?

Overall, learners reported spending an average of 1–2 hr studying English (M = 2.38/10; SD = 1.11) and 4–6 hr (M = 2.86/10; SD = 1.44) using English each week. The mean of young learners' self-reported perceptions of English learning experience was 7.57/10 (SD = 1.89), which shows that, overall, the learners had a favorable and engaging educational experience. Finally, the learners reported having completed a mean of 4.48 (SD = 2.90) individual and group assignments per month during the semester. Descriptive statistics for young learners' self-reported background variables are summarized in Table 4.

The regression results are all summarized in Table 5. Turning first to fluency features, three multiple regression models were fit to learners' score gains on speech rate, number of pauses, and mean length of pauses. For speech rate, the model explained 36% of the variance in score gains for the Bilingual school (F(5, 30) = 5.02, p = .002), 25% for the Private school (F(5, 27) = 3.14, p = .02), and 19% for the Public school (F(5, 23) = 2.34, p = .07). For number of pauses, the model accounted for 31% of the variation in score gains for the Bilingual school (F(5, 30) = 4.24, p = .005), 41% for the Private school (F(5, 27) = 5.40, p = .001), and 83% for the Public school (F(5, 23) = 29.00, p < .001). For mean length of pauses, the model explained 99% of the variance in score gains for the Bilingual school (F(5, 30) = 6506.30, p < .001), 99% for the Private school (F(5, 27) = 2347.49, p < .001), and 57% for the Public school (F(5, 23) = 8.54, p < .001). Across all three learning contexts, initial proficiency (performance at T1) emerged as a negative significant predictor of gains in speech rate, number of pauses, and mean length of pauses, indicating that learners with lower initial proficiency exhibited greater improvement over time (delivered faster speech with fewer and shorter pauses) over 6 months. Assessment practices were positively and negatively associated with gains in mean length of pauses of Bilingual and Private school learners, respectively, although this association was only marginally significant for the Bilingual school.

In the second set of analyses, three regression models were fit to the learners' score gains on grammar measures, including syntactic complexity, accuracy, and dependent

Table 4. Summary of the self-reported background variables across the three learning contexts

	Bilingual school				Private school				Public school			
Variables	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Hours of study	3.11	1.15	1.67	6.07	2.10	.60	1.26	3.85	1.98	1.08	1.04	6.74
Target language use	3.68	1.71	1.52	8.67	2.65	1.07	1.21	5.88	2.32	1.12	1.11	5.79
English learning experience	7.34	1.74	2.00	10.00	8.36	1.72	2.00	10.00	7.13	2.00	1.60	9.60
Assessment practices	5.97	2.51	1.00	12.00	2.42	1.33	1.00	6.25	4.88	3.25	1.00	13.00

Note: Hours of study is a composite of nine items, each rated from 1 (0–1 hr) to 10 (more than 16 hr). *Target language* use is a composite of 11 items rated on the same scale. *English learning experience* is a composite of five items rated from 1 (not true at all) to 10 (very true). *Assessment practices* is a composite of four items in which students reported the number of assignments, quizzes, individual projects, and group projects they completed during the 6 months.

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

Table 5. Summary of multiple linear regression results

	Bilingual				Private				Public			
Variables	β	SE	p	R^2	β	SE	p	R^2	β	SE	p	R^2
Speech rate				.36				.25				.19
Hours of study	.20	.27	.47		.16	.24	.42		-.08	.34	.81	
Target language use	−.25	.27	.36		−.005	.25	.95		.10	.28	.73	
English learning perception	.08	.14	.57		.04	.16	.78		−.20	.20	.32	
Assessment practices	.25	.14	.08		.03	.16	.86		.04	.29	.88	
Proficiency	−.55	.14	<.001		−.61	.15	<.001		−.52	.17	.005	
Number of pauses				.31				.41				.83
Hours of study	−.29	.28	.31		.09	.22	.70		.07	.16	.64	
Target language use	.24	.28	.39		.03	.23	.87		−.10	.13	.41	
English learning perception	−.006	.14	.96		−.16	.14	.27		.13	.09	.17	
Assessment practices	.11	.14	.45		.01	.14	.91		−.08	.13	.56	
Proficiency	−.64	.14	<.001		−.65	.14	<.001		−.87	.07	<.001	
Mean length of pauses				.99				.99				.57
Hours of study	.002	.01	.85		.01	.01	.37		.14	.25	.57	
Target language use	.005	.01	.67		−.01	.01	.46		−.18	.20	.37	
English learning perception	.007	.006	.22		−.001	.01	.89		.16	.15	.28	
Assessment practices	.01	.006	.06		−.02	.01	.03		−.09	.21	.67	
Proficiency	−1.00	.006	<.001		−1.00	.009	<.001		−.75	.12	<.001	
Syntactic complexity				.53				.24				.39
Hours of study	.18	.23	.44		−.001	.24	.99		.07	.30	.82	
Target language use	.09	.23	.69		−.10	.25	.68		−.07	.24	.79	
English learning perception	−.16	.12	.18		−.02	.16	.90		.14	.17	.43	
Assessment practices	−.05	.12	.70		−.19	.16	.26		−.01	.26	.96	
Proficiency	−.70	.12	<.001		−.61	.16	.001		−.67	.14	<.001	
Dependent clause ratio				.78				.51				
Hours of study	.13	.16	.41		−.12	.20	.54		–	–	–	–
Target language use	.05	.16	.75		.20	.20	.32		–	–	–	
English learning perception	−.11	.08	.41		.01	.13	.94		–	–	–	
Assessment practices	−.02	.08	.77		−.05	.13	.68		–	–	–	
Proficiency	−.84	.08	<.001		−.78	.13	<.001		–	–	–	
Syntactic accuracy				.54				.36				.25
Hours of study	−.11	.23	.63		.05	.22	.82		.05	.34	.88	
Target language use	.29	.24	.24		.47	.25	.07		−.07	.27	.78	
English learning perception	.06	.11	.52		.16	.15	.30		−.16	.19	.41	

(*Continued*)

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

Table 5. (*Continued*)

	Bilingual				Private				Public			
Variables	β	SE	p	R^2	β	SE	p	R^2	β	SE	p	R^2
Assessment practices	.08	.12	.52		.34	.15	.04		.01	.28	.96	
Proficiency	−.82	.12	<.001		−.47	.18	.01		−.57	.16	.002	
Lexical resource				.07				.22				.27
Hours of study	−.09	.32	.78		.03	.26	.91		−.05	.31	.87	
Target language use	.05	.33	.88		.27	.26	.30		−.03	.26	.90	
English learning perception	.08	.16	.64		.22	.17	.19		−.11	.19	.56	
Assessment practices	.13	.17	.45		.14	.16	.38		.15	.27	.57	
Proficiency	−.39	.18	.04		−.49	.16	.006		−.58	.16	.001	
Lexical stress				.99				–				
Hours of study	−.009	.03	.80		–	–	–		–	–	–	–
Target language use	.01	.03	.71		–	–	–		–	–	–	
English learning perception	−.01	.02	.54		–	–	–		–	–	–	
Assessment practices	.003	.02	.89		–	–	–		–	–	–	
Proficiency	−.10	.02	.000		–	–	–		–	–	–	
Content quantity				.29				.20				.12
Hours of study	.30	.28	.30		−.04	.26	.87		−.19	.36	.59	
Target language use	−.65	.30	.04		.03	.28	.92		.25	.29	.40	
English learning perception	.07	.14	.64		.35	.17	.04		−.25	.21	.25	
Assessment practices	.33	.15	.03		−.04	.16	.81		.14	.31	.65	
Proficiency	−.18	.17	.27		−.40	.18	.03		−.41	.17	.03	

Note: Regression models were not estimated for lexical stress gain in the Private school group (SD = .02), or for lexical stress (SD = .000) and dependent clause ratio gains (SD = .02) in the Public school group, due to insufficient outcome variability for regression analysis. R^2 values represent adjusted R^2.

clause ratio. For syntactic complexity, the background variables accounted for 53% of the variance among Bilingual school learners ($F(5, 30) = 8.79$, $p < .001$), 24% among Private school learners ($F(5, 27) = 3.02$, $p = .03$), and 47% among Public school learners ($F(5, 23) = 4.54$, $p = .005$). For syntactic accuracy, the model explained 54% of the variance for Bilingual ($F(5, 30) = 9.34$, $p < .001$), 36% for Private ($F(5, 27) = 4.59$, $p = .004$), and 25% for Public school learners ($F(5, 23) = 2.84$, $p = .04$). For dependent clause ratio, the model explained 78% of the variance in score gains for Bilingual school learners ($F(5, 30) = 26.02$, $p < .001$) and 51% for Private school learners ($F(5, 27) = 7.82$, $p < .001$). Across all grammar measures and learning contexts, initial proficiency emerged as a consistent negative predictor of gains, indicating that learners with lower initial proficiency could achieve greater improvement over time. Additionally, assessment practices were positively related to gains in syntactic accuracy among Private school learners.

The next set of analyses examined the impacts of background variables on the vocabulary measure of lexical resource. The background variables explained 7% of the variance in gains for Bilingual school learners ($F(5, 30) = 1.58$, $p = .20$), 22% for Private school learners ($F(5, 27) = 2.77$, $p = .04$), and 27% for Public school learners ($F(5, 24) =$

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

3.20, $p = .02$). Proficiency was the only significant predictor across all three learning contexts, with negative coefficients suggesting that learners with lower initial proficiency demonstrated greater gains in lexical resource, reflected in the use of a wider range of lexical items.

In the last set of analyses, two regression models were fit to the learners' gains in lexical stress and content quantity. For lexical stress, the model accounted for 99% of the variance in gains among Bilingual school learners ($F(5, 30) = 685.72$, $p < .001$), with initial proficiency emerging as the sole significant negative predictor. For content quantity, the background variables explained 29% of the variation in gains for Bilingual school learners ($F(5, 30) = 3.89$, $p = .008$), 20% for Private school learners ($F(5, 27) = 2.65$, $p = .04$), and 12% for Public school learners ($F(5, 23) = 1.74$, $p = .16$). Assessment practices and target language use were positively associated with gains in content quantity among Bilingual school learners. For Private school learners, gains were negatively associated with initial proficiency and positively associated with perceptions of English learning. In the Public school context, initial proficiency was the only significant predictor, suggesting that lower proficiency learners achieved greater gains.

Qualitative findings

The interview data partially support the quantitative findings by confirming that the three learning contexts offered varying degrees of English exposure, instructional support, and opportunities for meaningful language use. Across teacher and student interviews, Bilingual and Private school participants described more extensive and ongoing contact with English both inside and outside the classroom, while Public school participants generally focused on classroom learning, limited resources, motivation, and teacher-created activities/materials. These contextual differences help explain the ANCOVA results that indicated stronger performance for Bilingual and Private school learners across several linguistic domains, especially lexical resource, syntactic accuracy, speech rate, and content quantity.

In the Bilingual school, teachers consistently reported incorporating English throughout the curriculum and using it regularly during the school day. One teacher postulated that "Everything has to be in English. So I think it has to be getting better," suggesting ongoing exposure to the target language. Another teacher stated that students had improved because "the materials help them a lot to learn a lot of new words, vocabulary mostly," while "the immediate correction of some grammar structure has helped them internalize" language forms. One more teacher emphasized reading comprehension, explaining, "I like the reading comprehension… to make them comprehend what they're reading," and noted that students had improved in fluency and confidence when reading aloud. Students' responses complemented this pattern of extensive English use beyond school. One Bilingual school learner mentioned, "I talk English in my house and with my parents… I read in English, I write in English, and I watch movies in English. Everything's English." Another student similarly reported, "I mostly do English. I mostly talk in English. I mostly read in English." One learner described reading for pleasure and vocabulary development: "I read books in English… many words that I don't understand, I look them in the dictionary and… I learn many new words." Others linked their progress to speaking practice, with one student noting, "Talking makes me feel comfortable… I feel that I am succeeding each time more." Another said that listening helped because "just hearing people talk and listening to new words helps me understand."

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

In the Private school context, teachers highlighted explicit instruction in pronunciation, grammar, vocabulary, and speaking. One teacher explained, "I usually emphasize more on pronunciation, vocabulary, grammar… because what we want is for them to produce the language, to speak." Another teacher stated that she aimed to develop "the whole skills… reading, grammar, listening" so that students could improve overall comprehension. At the same time, Private school teachers acknowledged limitations in instructional time. One teacher hinted, "They have an hour a day or one hour and a half and sometimes that's not enough. Maybe if they had more hours a day they would be able to practice more." Students in the Private school context frequently described self-directed English learning through media, technology, and extracurricular activities. One student reported that she had learned "mostly by looking on the internet," while another stated that she surrounds herself with English through "watching movies in English, listening to music in English… socials… I make sure it's always English." She also added, "The more you're around English… the more you learn." Another learner reported attending additional English classes outside school: "I go in the afternoon classes… four days a week." Some students also described reading and writing in English at home, with one learner stating, "I like to write… which also helps a lot."

In the Public school setting, interviews portrayed a learning environment highly dependent on classroom instruction and teacher initiatives. One student explicitly linked school access to opportunity, explaining, "Since I don't have so many resources… it's very good that they teach us English for free." This comment suggests that access to English learning opportunities outside school may have been more limited for some learners. The Public school teacher repeatedly underscored motivation as integral to success, explaining that "motivation… is very crucial or important" because some students face challenges before arriving at school. She also described the importance of documenting progress, stating that "their notebook is their portfolio of evidence… the evidence of what you have learned during day by day the whole year." In addition, she noted adapting materials to learners' needs: "I don't really like the books… most of the times I just make changes because I think they are very difficult for them." She also highlighted technology use, explaining, "We have used a lot of technology here in our classroom so we have improved a lot." Taken as a whole, these comments help explain why Public school learners generally scored lower on speech rate, lexical resource, syntactic accuracy, and content quantity. More limited access to English beyond school and heavier reliance on classroom instruction may have reduced opportunities for vocabulary growth and extended oral production.

The interview data also help interpret the regression findings, in which proficiency emerged as the most consistent predictor across contexts and outcomes. Both teachers and students repeatedly linked progress to frequent use, accumulated exposure, and confidence. Learners who reported speaking English at home, exposure to English media, reading independently, or attending extra classes often described stronger progress. By contrast, estimates of "hours of study" were highly variable (e.g., "one hour," "all day," "two hours," or counting screen time), suggesting that quantity of study alone may be less informative than the quality and regularity of English engagement.

Overall, the qualitative findings converge with the ANCOVA and regression results, demonstrating that learning context influenced learners' opportunities to use English, the types of instruction they received, and their access to complementary educational resources. Bilingual school learners benefited from immersion-like exposure across subjects and home use. Private school learners benefited from explicit instruction

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

combined with extracurricular English contact. Public school learners, on the other hand, benefited from committed teachers and classroom motivation efforts but appeared to face more limited external resources and fewer opportunities for sustained English use. These contextual differences are consistent with the quantitative findings in which Bilingual and Private school learners generally outperformed Public school learners on multiple linguistic domains.

Discussion

This longitudinal study explored young FL learners' progression in various linguistic features of speaking (i.e., fluency, pronunciation, vocabulary, grammar, and content quantity) and how their progress was related to the learners' educational context and background variables. The following discussion summarizes and interprets the results.

To address the first research question, a series of ANCOVA models was computed, with educational context treated as an independent variable, adjusted T2 scores on the linguistic features as a dependent variable, and scores at T1 as a covariate. The findings indicated that, for most linguistic features, including speech rate, lexical resource, dependent clause ratio, syntactic accuracy, and content quantity, Bilingual and Private school learners outperformed Public school learners. Bilingual school learners also surpassed Private school learners in speech rate and content quantity at T2. These findings coincide with those of previous research (e.g., Azpilicueta-Martínez, 2024, 2025; Lee et al., 2023; Pérez Cañado & Lancaster, 2017) demonstrating the advantage of EMI-CLIL programs, which typically involve greater instructional intensity and increased exposure to the target language, for the development of linguistic features of speaking in young learners. These findings also provide validity evidence for the TOEFL Primary speaking test, which is built on the assumption that speaking test performance varies according to learners' language learning experience and exposure (see, e.g., Hsieh, 2024). In a similar vein, Hsieh and Wang (2019) and Gu and Hsieh (2019) demonstrated that proficiency level significantly differentiated young learners' performance across multiple linguistic features of speaking, including fluency, lexical resource, grammatical complexity, and grammatical accuracy. Likewise, findings of the current investigation suggest that educational contexts associated with greater English exposure and opportunities for meaningful language use are likely to contribute to differences across the spoken linguistic dimensions. In the context of this study, Bilingual school learners received 30 hr of in-class exposure to English per week, which was 22.5 hr more than Private school learners and 27.5 hr more than Public school learners. In addition, Bilingual school learners were engaged in various school-organized extracurricular activities, such as programs to experiment with creating projects in the makerspace, yoga, and other extracurricular activities. On the other hand, Public school learners generally focused on classroom learning and had limited access to English-mediated resources and opportunities for interaction. Supporting these quantitative findings, one Bilingual school teacher noted during the interview session that "Everything has to be in English." Similarly, a Bilingual school student mentioned that "I talk English in my house and with my parents… I read in English, I write in English, and I watch movies in English. Everything's English." Private school students also reported frequent engagement with English outside of school through digital media and online resources. For example, one student indicated that they had learned "mostly by looking on the internet," whereas another explained that they surrounded themselves with English through "watching movies in English, listening to music in English… socials…" and

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

emphasized that "the more you're around English… the more you learn." Consistent with the interview findings, the analysis of the self-reported target language use questionnaire further revealed that the Bilingual school learners had significantly more out-of-school exposure than the Private ($p = .005$, $d = .77$) and Public school learners ($p < .01$, $d = .98$).

As for the pronunciation feature of lexical stress, the findings revealed no significant difference in learners' adjusted T2 scores across the three learning contexts. This finding may be explained by the fact that young learners acquire lexical stress patterns through repeated exposure to familiar lexical items rather than through internalizing abstract phonological rules (e.g., Redford & Oh, 2016). As a result, stress patterns in familiar words become consolidated in long-term memory, resulting in relatively stable performance across different contexts. Given the high frequency and familiarity of the words featured in the assessment tasks, performance on this prosodic feature was expected to remain relatively stable over time, with minimal variation across the three learning contexts. This finding also partially aligns with those of Pérez Cañado and Lancaster (2017) and Rallo Fabra and Jacob (2015), which indicated that pronunciation tends to be less sensitive to instructional context differences (CLIL vs. non-CLIL). However, it should be noted that these studies did not specifically investigate lexical stress. Thus, the present finding remains largely exploratory and should be explored in future research. Finally, contrary to our expectations and previous research findings (Azpilicueta-Martínez, 2024; Pérez Cañado & Lancaster, 2017), the Public school learners outperformed the Bilingual and Private school learners on pause-related and syntactic complexity measures. These findings, however, should be interpreted with caution, as **only a small number of Public school learners produced analyzable speech at T2, and their comparatively shorter utterances may have potentially inflated ratio-based complexity indices and affected pause-related estimates. Therefore, these findings are likely to reflect statistical properties associated with shorter production rather than qualitatively more advanced performance. Future research with larger and more balanced samples is needed to determine if these findings are replicable.**

The results of multiple regression analyses revealed that initial proficiency, operationalized as young learners' scores at T1, consistently and significantly predicted growth across all linguistic features and learning contexts over time. This finding implies that young learners' initial proficiency level at the beginning of the course is likely to influence the extent to which they progress over 6 months. This finding comports with previous research by De Wilde et al. (2021), Jackel et al. (2017), and Unsworth et al. (2014), who reported proficiency as one of the most potent predictors of L2 learners' gains in different linguistic features of speaking. Since score gains are more likely to occur at lower proficiency levels, it may be pedagogically important to assess young learners' proficiency at the beginning of the semester so that each learner can be guided and informed about their progress throughout the semester. Furthermore, low-proficiency learners should be encouraged to benefit from contextualized learning, especially in a context where English is readily available and young learners are motivated to learn the language (De Wilde et al., 2021).

In addition to initial proficiency, assessment practices and target language use emerged as positive significant predictors of young learners' improvement over time. Specifically, assessment practices were associated with improvement in mean length of pauses and syntactic accuracy among Private school learners, as well as content quantity production among Bilingual school learners. Target language use was also

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

positively related to Bilingual school learners' content quantity production. As for assessment practices, the findings underscore the significance of frequency, mode, and types of assessment practices employed by language teachers in shaping learners' L2 development (Latif & Wasim, 2024; Yu & Xu, 2021), particularly among young L2 learners aged 10–12. However, given that previous research has rarely investigated the effects of assessment practices on the development of young learners' linguistic dimensions of speaking, the findings of this study should be interpreted as exploratory and warrant further exploration.

Likewise, the positive relationship between target language use and Bilingual school learners' content quantity production is unsurprising, given that the Bilingual school represented a high-CLIL/immersion context that involved ample opportunities for target language use and meaningful interaction using the L2. Frequent exposure to and active use of the L2 over time may have reinforced young learners' ability to retrieve linguistic resources more efficiently to elaborate their ideas and produce longer spoken responses. That the remaining variables did not emerge as significant was somewhat unexpected and warrants further research. One potential reason would be that this study collected background information through young learners' self-reports, which, while appropriate for this context, may have reflected young learners' subjective interpretation rather than objective measures. This is an avenue for future research to adopt more objective measures for assessing young learners' backgrounds.

Conclusion and implications

This longitudinal study investigated the extent to which young learners' linguistic features of speaking improved over 6 months and how the learners' progress was related to their educational context and background factors. Several limitations should be acknowledged. First, the study narrowed the study population to young learners aged 10–12 in one EFL context, which may limit the generalizability of the findings. Second, it collected learners' background information solely through self-report questionnaires and may therefore be subject to reporting bias. Third, it investigated score gains within a rather short timeframe. Fourth, it used only two time points and relatively small samples within each school context. Fifth, it combined the linguistic features across the picture description and narration tasks that may have masked potential task-specific effects on learners' spoken responses. Sixth, the study did not explore other factors such as motivation, aptitude, and learning style that may potentially influence score gains over time. Seventh, the same TOEFL Primary speaking test was administered at T1 and T2. Thus, score gains might have been partly influenced by test familiarity or practice effects. Finally, school context was associated not only with different instructional models and the amount of exposure to the English language they provided, but also with the socioeconomic level of students they served. Although information on the socioeconomic profile of schools suggests that, generally, students from the Private and Bilingual schools came from advantaged backgrounds, information on individual students' socioeconomic backgrounds was not available. Therefore, it is hard to disentangle how much of the observed differences could be attributed to instructional factors, students' socioeconomic background, school resources, or an interaction of all. Thus, the findings should be interpreted with caution. Despite existing limitations, there are important implications that can be drawn from the study findings. First, the findings revealed that young learners may benefit from educational contexts that provide ample

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

opportunities for communication and meaningful language use. Thus, educational programs are recommended to provide opportunities for young learners to immerse themselves in the L2 through increased exposure, sustained meaningful interaction, and engagement in English-mediated activities, even in lower exposure EFL settings. Second, given that assessment practices were related to young learners' development in different linguistic dimensions of speaking, language teachers may consider incorporating frequent and varied assessment opportunities that promote spoken language use and ongoing feedback (see Gass et al., 2020). Third, given the variation in the development of the linguistic features of speaking across the three learning contexts, language teachers and curriculum designers may consider adopting a multidimensional approach to speaking instruction and assessment that acknowledges that linguistic features of speaking may not develop uniformly across various instructional settings. Fourth, since young learners' proficiency predicts their developmental patterns in various linguistic features of speaking, language programs and institutions are recommended to administer placement/diagnostic tests to evaluate learners' proficiency levels and then offer level-specific objectives and learning outcomes. In fact, given that lower proficiency learners may demonstrate different developmental trajectories over time, language teachers are encouraged to provide different learning opportunities for this cohort to help them grow motivation and positive attitudes toward L2 study in general. Overall, the study findings add to the existing body of literature on young learners' L2 speaking development by adopting a longitudinal and multidimensional approach across three different learning contexts. This study also provides validity evidence for the TOEFL Primary speaking test by indicating that assessment is sensitive to and is likely to be influenced by learners' instructional experiences and exposure to English.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263126102046.

Funding statement. This study was sponsored by Educational Service's Young Student Series Program.

Competing interests. There are no competing interests.

References

Azpilicueta-Martínez, R. (2024). Exposure or age? The effect of additional CLIL instruction on young learners' grammatical complexity while performing an oral task. *Language Teaching Research*. https://doi.org/10.1177/13621688241303250

Azpilicueta-Martínez, R. (2025). Revising expectations: The effect of different levels of CLIL exposure on young learners' oral performance. *Journal of Immersion and Content-Based Language Education*, *13*(1), 99–122. https://doi.org/10.1075/jicb.24004.azp

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

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer (*Version 6.0.11*) [Computer software]*. https://www.fon.hum.uva.nl/paul/praat.html

Braun V., Clarke V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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

Cho Y., Blood I. A. (2020). An analysis of TOEFL® Primary™ repeaters: How much score change occurs? *Language Testing*, *37*(4), 503–522. https://doi.org/10.1177/0265532220927751

Cobb, T. (2024, May). *Web VP classic v.4 [Computer software]*. www.lextutor.ca/vp/eng/

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893–916. https://doi.org/10.1017/S0142716412000069

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition*, *23*(1), 171–185. https://doi.org/10.1017/S1366728918001062

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2021). Young learners' L2 English after the onset of instruction: Longitudinal development of L2 proficiency and the role of individual differences. *Bilingualism: Language and Cognition*, *24*(3), 439–453. https://doi.org/10.1017/S1366728920000747

Djigunović, J. M. (2016). Individual learner differences and young learners' performance on L2 speaking tests. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 243–261). Springer.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal*, *97*(1), 131–143. https://doi.org/10.1111/j.1540-4781.2013.01434.x

Educational Testing Service. (2023). *Handbook for the TOEFL Primary® tests*. https://www.ets.org/pdfs/toefl/toefl-primary-handbook.pdf

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications.

Freed, B. F., Dewey, D. P., Segalowitz, N. S., & Halter, R. H. (2004). The language contact profile. *Studies in Second Language Acquisition*, *26*(2), 349–356.

French, L. M., Gagné, N., & Collins, L. (2020). Long-term effects of intensive instruction on fluency, comprehensibility and accentedness. *Journal of Second Language Pronunciation*, *6*(3), 380–401. https://doi.org/10.1075/jslp.20026.fre

Gass, S. M., Behney, J., & Plonsky, L. (2020). Second language acquisition: An introductory course (5th ed.). Routledge. https://doi.org/10.4324/9781315181752

Gu, L., & Hsieh, C.-N. (2019). Distinguishing features of young English language learners' oral performance. *Language Assessment Quarterly*, *16*(2), 180–195. https://doi.org/10.1080/15434303.2019.1605518

Gu, L., Lockwood, J., & Powers, D. E. (2015). *Evaluating the TOEFL Junior® Standard test as a measure of progress for young English language learners* (Research Report No. RR-15-22). Educational Testing Service. https://doi.org/10.1002/ets2.12064

Gu, L., & So, Y. (2017). Strategies used by young English learners in an assessment context. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 118–135). Routledge.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Hsieh, C.-N. (2024). *Building a validity argument for the TOEFL Primary® tests* (TOEFL Research Report No. RR-24-16). ETS. https://www.ets.org/Media/Research/pdf/RR-24-16.pdf

Hsieh, C. N., & Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*, *36*(1), 27–50. https://doi.org/10.1177/0265532217734240

Huang, B. H., Bailey, A. L., Sass, D. A., & Shawn Chang, Y. H. (2021). An investigation of the validity of a speaking assessment for adolescent English language learners. *Language Testing*, *38*(3), 401–428. https://doi.org/10.1177/0265532220925731

Hunt, M. (2011). Learners' perceptions of their experiences of learning subject content through a foreign language. *Educational Review*, *63*(3), 365–378.

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

Jackel, N., Schurig, M., Florian, M., & Ritter, M. (2017). From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning*, *67*(3), 631–664. https://doi.org/10.1111/lang.12242

Johnstone, R. (2009). An early start: What are the key conditions for generalized success? In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: International perspectives* (pp. 31–41). Garnet Education.

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

Kang, O. (2012). Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (pp. 10–15). Ames, IA: Iowa State University. Retrieved from https://www.iastatedigitalpress.com/psllt/article/15198/galley/13721/view/

Kang, O., Thomson, R. I., & Moran, M. (2020). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*, *41*(4), 453–480. https://doi.org/10.1093/applin/amy053

Kang, O., Ahn, H., Yaw, K., & Chung, S.-Y. (2021). Investigation of relationship among learner background, linguistic progression, and score gain on IELTS. *IELTS Research Reports Online Series, No. 1*. British Council, Cambridge Assessment English and IDP: IELTS Australia. https://www.ielts.org/teaching-and-research/research-reports

Kang, O., & Kermad, A. (2020). Suprasegmental development of ESL learners in an immersion context. *Speak Out*, *63*, 43–50.

Kim, W. G., & García, S. B. (2014). Long-term English language learners' perceptions of their language and academic learning experiences. *Remedial and Special Education*, *35*(5), 300–312.

Lázaro-Ibarrola, A. (2012). Faster and further morphosyntactic development of CLIL vs. EFL Basque-Spanish bilinguals learning English in high-school. *International Journal of English Studies*, *12*(1), 79–96. https://doi.org/10.6018/ijes.12.1.125621

Lázaro-Ibarrola, A. (2024). What factors contribute to the proficiency of young EFL learners in primary school? Assessing the role of CLIL intensity, extramural English, non-verbal intelligence and socioeconomic status. *Language Teaching Research*, 13621688241292277. https://doi.org/10.1177/13621688241292277

Lázaro-Ibarrola, A. (2025). Intensity matters inside and outside primary school: Evidence from high-CLIL, low-CLIL, and non-CLIL learners. *TESOL Quarterly*, *59*(3), 1564–1595. https://doi.org/10.1002/tesq.3369

Lázaro-Ibarrola, A., & Roothooft, H. (2025). The impact of CLIL intensity and extramural English on the receptive skills of young EFL learners: CLIL intensity and extramural English. *Journal of Immersion and Content-Based Language Education*. https://doi.org/10.1075/jicb.24021.laz

Latif, M. W., & Wasim, A. (2024). Investigating EFL instructors' approaches to classroom-based assessment culture: An explanatory sequential mixed-method approach. *Language Testing in Asia*, *14*(1), 57. https://doi.org/10.1186/s40468-024-00332-x

Lee, J. H., Lee, H., & Lo, Y. Y. (2023). Effects of EMI-CLIL on secondary-level students' English learning: A multilevel meta-analysis. *Studies in Second Language Learning and Teaching*, *13*(2), 317–345.

Lorenzo, F., Granados, A., & Rico, N. (2021). Equity in bilingual education: Socioeconomic status and content and language integrated learning in monolingual Southern Europe. *Applied Linguistics*, *42*(3), 393–413. https://doi.org/10.1093/applin/amaa037

Merino, J. A., & Lasagabaster, D. (2018). The effect of content and language integrated learning programmes' intensity on English proficiency: A longitudinal study. *International Journal of Applied Linguistics*, *28*(1), 18–30. https://doi.org/10.1111/ijal.12177

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447–468. https://doi.org/10.1037/xge0000054

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, *35*, 1–21. https://doi.org/10.1093/applin/amu024

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Orosz, A. (2009). The growth of young learners' English vocabulary size. In M. Nikolov (Ed.), *Early learning of modern foreign languages* (pp. 181–195). Multilingual Matters.

Pérez Cañado, M. L., & Lancaster, N. K. (2017). The effects of CLIL on oral comprehension and production: A longitudinal case study. *Language, Culture and Curriculum*, *30*(3), 300–316. https://doi.org/10.1080/07908318.2017.1338717

Peters, E. (2018). The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. *ITL-International Journal of Applied Linguistics*, *169*(1), 142–168. https://doi.org/10.1075/itl.00010.pet

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press

Rallo Fabra, L., & Jacob, K. (2015). Does CLIL enhance oral skills? Fluency and pronunciation errors by Spanish-Catalan learners of English. In M. Juan-Garau & J. Salazar Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 163–177). Springer.

Redford, M. A., & Oh, G. E. (2016). Children's abstraction and generalization of English lexical stress patterns. *Journal of Child Language*, *43*(2), 338–365. https://doi.org/10.1017/S0305000915000215

Roessingh, H., & Elgie, S. (2009). Early language and literacy development among young English language learners: Preliminary insights from a longitudinal study. *TESL Canada Journal*, *26*(2), 24–45. https://doi.org/10.18806/tesl.v26i2.413

Ruiz de Zarobe, Y., & Banegas, D. L. (2024). *Content and language integrated learning in South America*. Springer.

Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics*, *37*(4), 813–840. https://doi.org/10.1017/S0142716415000259

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Suzuki, S., & Kormos, J. (2025). The moderating role of L2 proficiency in the predictive power of L1 fluency on L2 utterance fluency. *Language Testing*, *42*(1), 73–99. https://doi.org/10.1177/02655322241241851

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2), 435–463. https://doi.org/10.1111/modl.12706

Swanson, H. L. (2015). Growth in working memory and inhibition predicts literacy in English language learners: A cross-sectional and longitudinal study. *Memory*, *23*(5), 748–773. https://doi.org/10.1080/09658211.2014.927504

Sylvén, L. K., & Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL*, *24*(3), 302–321. https://doi.org/10.1017/S095834401200016X

Tabachnick, B., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal*, *104*(1), 169–191. https://doi.org/10.1111/modl.12620

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). John Benjamins.

Unsworth, S., Persson, L., Prins, T., & De Bot, K. (2014). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*, *36*(5), 527–548. https://doi.org/10.1093/applin/amt052

Wolf, M. K. (2024). Assessment of young language learners. In J. K. Antony (Ed.), *The concise companion to language assessment* (pp. 312–325). Wiley-Blackwell.

Yu, S., & Xu, H. (2021). Feedback and assessment in second language education during the COVID-19 and beyond. *The Asia-Pacific Education Researcher*, *30*(6), 483–486. https://doi.org/10.1007/s40299-021-00631-4

Cite this article: Dalman, M., Kang, O., & Gutierrez Arvizu, M. N. (2026). Exploring linguistic features in young foreign language learners' speech: *The interplay of learner background, educational context, and time in 6 months of development*. *Studies in Second Language Acquisition,* 1–28. https://doi.org/10.1017/S0272263126102046

https://doi.org/10.1017/S0272263126102046 Published online by Cambridge University Press