CAMBRIDGE
UNIVERSITY PRESS

ORIGINAL ARTICLE

Bilingual oral language development among dual language immersion students: Use of a Bayesian approach with language learning progressions

Ève Ryan[1] , Preston Botter[2] , Michelle L. Luna[3] , Alison L. Bailey[4] and Clémence Darriet[4]

[1]Department of Linguistics, University of Alaska, Fairbanks, Fairbanks, AK, USA, [2]School of Education, Indiana University, Bloomington, IN, USA, [3]Department of Psychology and the Institute for Educational Initiatives, University of Notre Dame, Notre Dame, IN, USA and [4]Department of Education, University of California, Los Angeles, Los Angeles, CA, USA
Corresponding author: Ève Ryan; Email: eryan9@alaska.edu

Abstract
The dual language development of dual language immersion (DLI) students, although often examined at the domain level (e.g., listening or reading), remains understudied for more specific skills (e.g., word, sentence, or discourse). This study examines the eleven-month progression of oral language skills in a picture description task in two languages (French and English) for early-elementary (Transitional Kindergarten through first grade) DLI students ($N = 42$). Using Bayesian methods, which estimate parameters using both the data and prior information, we describe French and English growth patterns as measured by learning progressions whose focus is on language features at the word, sentence, and discourse levels. For French oral language, we found evidence of meaningful positive linear growth for all language features, whereas for English oral language, meaningful linear positive growth was only detected for sophistication of topic vocabulary. Overall, coming from a French-speaking household was associated with steeper French oral language trajectories, but coming from an English-only household did not specifically impact English oral language trajectories. In both languages, grade level influenced the trajectories of some—but not all—features. We conclude with theoretical and practical implications, advocating for a language progression approach in instruction and research on bilingualism.

Keywords: Bayesian; bilingual; growth; immersion; picture description; progression

For children enrolled in a dual language immersion (DLI) program in the United States, instruction is delivered in English and another language, often referred to as the partner language. Research on DLI students has tended to focus on the development of their language(s) (i.e., English and/or the partner language) in different domains (e.g., listening or reading), without necessarily differentiating between different features of language within each domain (i.e., word, sentence, and discourse). In this study, we examine the progression of oral language skills in a picture description task in both languages for early-elementary students enrolled in a French-English DLI program, where daily instructional time is split equally between French and English.

Background

Language development in bilingual children

Language development can be understood as a complex adaptive system. This understanding follows a set of assumptions about language learning and use that considers the dynamic, adaptive, and nonlinear nature of how language is used and develops (Beckner et al., 2009). By taking this approach, language learning is seen within a framework that considers multiple features of language use as well as the social context of the learner (Bailey & Heritage, 2014). As young children develop language from initial recognition of sounds to oral production and beyond, a number of factors may come into play that influence, sometimes in dramatic ways, the course of language development. Listening and oral production are key in that they serve as precursors to later development of reading and writing (e.g., Dickinson, et al., 2003; Speece et al., 1999). A number of pattern-identifying skills help young children acquire an awareness of sounds, their meanings, and rules governing their association. This, in turn, leads to children's acquisition of vocabulary and the ability to form a variety of different sentence structures and increasingly complex and extended discourse such as conversation, narration, explanation, and argumentation (e.g., Bailey et al., 2015). These oracy skills are later applied to the decoding necessary for reading and writing, for which oral production is a necessary precursor (Dickinson et al., 2003).

For bilingual children, language development largely mirrors that of their monolingual peers, with some notable differences. Bilingual children acquire two (sometimes very different) systems, which interact and support the development of the other[1]. For example, phonological systems acquired by bilinguals may transfer positively or negatively between languages, with the phonological development in one language serving in the reading and development in the other, including for languages that do not share the same orthography (Hambly et al., 2013). Bilingual children's vocabulary develops as two systems; however, they may not necessarily acquire two words for one underlying concept, meaning that the size of their vocabulary may be smaller in one language but larger when considering both (Oller et al., 2007). Sentence production for bilinguals seems to occur at the same rate as monolinguals, though this depends on exposure to each language, and one language may serve as a transfer to the other, with a silent period of production in the other

language influenced by several external and internal factors (Gathercole & Thomas, 2009). For bilingual children, the development of narration, conversation, and argumentation is predicated on their development of word, sentence, and discourse features, which may differ between languages and are culturally embedded (e.g., Fiestas & Peña, 2004). Finally, bilingual children demonstrate pragmatic abilities not seen in the monolingual child, including code switching, which is the highly contextually dependent and rule-governed switch from one language to the other to convey meaning (Martínez & Mejía, 2019).

Language progressions of DLI students

Most existing studies on the effects of bilingual education in the U.S. focus on emergent bilingual populations, whose language skills in English and in the home language benefit from enrollment in such programs (e.g., Rolstad et al., 2005). For DLI students specifically, the first few years are characterized by an increased proficiency in listening and speaking for the partner language (Fortune & Tedick, 2015). Program enrollment also has long-term benefits on English language development (e.g., Steele et al., 2017). DLI students display increased levels of proficiency in the partner language from elementary to secondary grades (Watzinger-Tharp et al., 2021), and for elementary-aged students, progress is especially apparent in listening and reading (Watzinger-Tharp et al., 2018). Heritage language status has also been found to influence partner language trajectories, especially in partner language instruction (e.g., Lindholm-Leary & Hernández, 2011; Xu et al., 2015). For instance, at the preschool level, Spanish-dominant children had higher scores than bilinguals who had relatively comparable proficiencies in their languages at the beginning of the school year, and both groups displayed higher rates of growth in Spanish when receiving Spanish-only instruction as opposed to bilingual or English-only instruction (Durán et al., 2022). In a separate study with fourth-grade students in an English-Spanish DLI program, coming from a household where Spanish is spoken was associated with higher listening and writing performance, but similar reading and speaking performance compared to peers from non-Spanish-speaking households (Burkhauser et al., 2016). A recent study similarly found that home exposure to Gaelic in a Gaelic-English revitalization/immersion program was associated with higher vocabulary output and size, though this is also depended on time of exposure to the partner language (Chondrogianni et al., 2022).

While the aforementioned studies focused on the progression of all language domains (i.e., speaking, writing, listening, and reading), this study is specifically concerned with speaking and the separate trajectories for the development of word, sentence, and discourse skills. Research shows that for children in immersion settings, their oral language development in the partner language is often characterized by "lower-than-expected production skills in terms of grammatical accuracy, lexical variety, and sociolinguistic appropriateness" (Lyster & Tedick, 2022, p. 329). Fortune and Tedick (2015), analyzing Spanish oral fluency, grammar, vocabulary, and listening comprehension in English-dominant students enrolled in two-way immersion programs, found significant differences across elementary grade levels but not between fifth and eighth graders. In another study, Fortune and Ju (2017) found no significant differences between second and fifth grade students

on standardized assessments of Mandarin. However, differences emerged when conducting a follow-up linguistic complexity analysis of representative speech samples: progress was made in grammar but not lexicon. Such findings highlight the need for finer-grained measures of oral language development. Indeed, it remains to be seen whether different levels of oral language (i.e., word, sentence, and discourse) progress at different rates.

Prior use of language progressions in speaking and writing tasks has revealed various rates of development in discrete domains of language and on tasks of differing themes or contexts (e.g., Bailey, 2017; Bailey & Heritage, 2019). For multilingual students learning English in particular, language progressions can be influenced by factors such as mode (i.e., oral vs. written), type of task (i.e., academic vs. non-academic), and language feature (i.e., vocabulary, sentence structure and discourse coherence/cohesion). For instance, one study reported greater strengths in written explanations compared to oral explanations for topic vocabulary (Blackstock-Bernstein et al., 2022), while another study reported strengths in English oral vocabulary but slower development of discourse-level oral skills in elementary students' explanations (Bailey et al., 2015). However, to our knowledge, research on language progressions and the detailed advancements in students' language learning has yet to be conducted across different languages in the context of DLI programming.

Oral *language skills in picture description tasks*

This study focuses on early oral language skills used in picture description tasks, which have been linked to literacy development (Dickinson & Snow, 1987; Griffin et al., 2004; Snow et al., 1995). Picture description tasks are especially helpful to track language development "because they elicit samples of language that share the same content (. . .), thus facilitating direct comparisons of the linguistic characteristics of the samples", and because "the amount of textual input can be kept minimal and so learners need to rely on their own linguistic resources" (Boers, 2018, p. 375). Thus, having children all describe the same pictures allows us to control the content of their language while maximizing their opportunities to use their own language skills.

Picture description tasks have been used with various populations, including neurotypical adults (e.g., Boucher et al., 2019), individuals with language and communication disorders (e.g., Eisenberg & Guo, 2013; Vandenborre et al., 2018), adult L2 learners (e.g., Saito & Hanzawa, 2018), and bilingual preschoolers (e.g., Smith et al., 2021). For neurotypically developing children, De Temple and Beals (1991) identify the key characteristics of a successful oral decontextualized picture description task. These characteristics include quantity (i.e., amount of talk), specificity (i.e., "the degree of explicitness and correctness in the child's description," p. 475), density (i.e., "the linguistic complexity and elaboration" of the description, p. 475), main theme (i.e., whether the key elements of the picture appear in the description), and narrativity (i.e., the extent to which the child interpreted the task as a straight description or as a story-like narrative). In other words, a good picture description will be thorough, will contain sophisticated topic vocabulary, and will be organized in a coherent and cohesive way. Compared to other oral language tasks,

Nurss and Hough (1985) note that "single pictures are very useful for producing labels and descriptive words, that is, for increasing vocabulary. They are not particularly useful, however, for helping children gain a concept of a story or more complex narrative structure" (p. 284). In other words, as a task, single picture descriptions seem to highlight children's lexical skills, but not necessarily their discourse-level skills. In addition, children may adjust their language based on whether they are describing the picture in a contextualized or decontextualized condition (i.e., for a present or an absent audience). For example, in a recent study, Cho and Kim (2023) found that linguistic and discourse features varied based on context, as "second graders used more ENPs (*elaborated noun phrases*) and exhibited precise character introduction in the decontextualized setting, whereas higher lexical diversity and discourse beyond simple description occurred more often in the contextualized setting" (p. 10).

Within the context of bilingual programming, there is a dearth of research on the role played by home language status on oral language skills used in picture description tasks, with the exception of one study. The participants in Wu et al.'s (1994) study were fifty-two second- through fifth-grade children attending the United Nations International School, where the curriculum is in English, with French taught as a foreign language. Participants provided written and oral picture descriptions in French and English, both in contextualized and decontextualized conditions. Results suggested that household language background influenced the quality of the picture description in French, but not in English. Indeed, on the French task, children from monolingual English households performed less well in measures of quantity (counts of total words), specificity (explicitness and correctness), and main theme (key elements of the picture) than children from bilingual households where French was not spoken. However, there were no significant differences between children from monolingual French households and children from bilingual households where French was spoken. It remains to be seen whether such findings would hold at a younger age, and also in another bilingual programming context, such as DLI.

To summarize, while enrollment in a DLI program is associated with growth in speaking in both languages, there is limited research on the progression of more fine-grained domains of language at the word, sentence, and discourse levels and the developmental trajectories of the two languages. Indeed, as the literature shows, children use different oral language skills in picture description tasks, and for those enrolled in a bilingual program, their performance on such tasks may be influenced by their home language(s).

Theoretical framework

This research is guided by complex dynamic systems theory (CDST), "a post-structural metatheory with its own ontology and epistemology" (Larsen-Freeman & Todeva, 2021, p. 209). CDST highlights the complex synergy between experiences such as language exposure in the home, social interactions such as use of language in DLI classrooms, and cognitive processes at play in individual language systems (Beckner et al., 2009). Language development is viewed as a self-organizing process in which a complex multifaceted system is in constant interaction with the

environment. CDST is "fundamentally concerned with describing and tracing emerging patterns in dynamic systems in order to explain change and growth" (Larsen-Freeman, 2020, p. 248), thus viewing language development as a dynamic process that evolves over time, with variation expected both within and across individuals (De Bot et al., 2007; Larsen-Freeman, 2011).

The original design of the language learning progressions used in the current study was made on this developmental premise (Bailey & Heritage, 2014). This perspective emphasizes the role played by initial conditions and posits that subsequent language development is expected to happen in a nonlinear fashion (De Bot et al., 2007), which influences our approach to analysis of the longitudinal data collected for this study. CDST is also particularly relevant for research on multilingual individuals (Larsen-Freeman & Todeva, 2021) given its emphasis on the role played by the environment on the longitudinal development of languages, as well as on the complexity and fluidity afforded to multilinguals in deploying their languages and literacy practices depending on the context and interlocutor.

This study examines the French and English oral language range of performances and trajectories over time of early-elementary school children enrolled in a French-English DLI program in a large city of the southwestern US. This study explores the following questions:

RQ1a. What are the French growth patterns as measured by learning progressions for early-elementary DLI students across 11 months of programming?

RQ1b. What are the English growth patterns as measured by learning progressions for early-elementary DLI students across 11 months of programming?

RQ1c. What is the relationship between French and English growth patterns as measured by learning progressions for early-elementary DLI students across 11 months of programming?

RQ2a. How are French progressions associated with grade level and household language background?

RQ2b. How are English progressions associated with grade level and household language background?

Methods and data sources

Data for this study came from a larger project on the French–English language development of early-elementary students in a DLI program (see Ryan, 2021a, 2021b).

Participants

Participants included 42 children (21 male; 21 female) who attended a French–English two-way DLI program in a public school in the Southwestern United States, with time equally divided between French and English, every other day across all content areas. When the study began, the participants, whose ages ranged between

5 years; 1 month and 7 years; 6 months, were in transitional Kindergarten[2] (TK; $n = 4$), Kindergarten (K; $n = 20$), and first grade ($n = 18$). At the time of the study, the TK and Kindergarten students were placed in the same classroom and were thus merged together for analysis. According to a survey filled out by the parents, among the 24 TK/K participants, 10 (42%) came from multilingual households where French was spoken, seven (29%) came from multilingual households where French was not spoken, and seven (29%) came from monolingual English households. Among the 18 first-grade participants, six (33%) came from multilingual households where French was spoken, five (28%) came from multilingual households where French was not spoken, and seven (39%) came from monolingual English households.

Participants were assessed five times over an eleven-month period, including three times in the winter and spring, and two times in the fall, after the summer hiatus. Of the 42 participants, 39 took part in all five waves of data collection, one took part in three waves of data collection, and two took part in only one wave of data collection.

Procedures

Overview

Children were asked to provide picture descriptions to an imaginary friend who only spoke French (for the French session) or English (for the English session). Children were instructed to describe the picture in the target language (i.e., French or English) so that the imaginary friend could later draw it without looking at it. Participants were then presented with two different pictures: one depicting a school scene (the "school picture"), where a child is performing show-and-tell in a classroom, and the other depicting a domestic scene (the "home picture"), where a child is doing dishes in a kitchen. During the first wave of data collection, half of the students were presented with the school picture first, while the other half were presented with the home picture. The order of the pictures (i.e., school or home scene) was alternated at each of the five waves, as was the order of the target language (i.e., French or English). Students' responses were later transcribed by a research assistant and then verified by a researcher.

Coding

After language samples were collected, researchers rated students' descriptions using a modified version of the Linguistic Features Analysis Protocol (LFAP) of the Dynamic Language Learning Progressions (DLLP) approach (Bailey & Heritage, 2014; Bailey, 2017) for the progression of several language features including the sophistication of topic vocabulary (vocabulary choices keyed to adult responses to the picture tasks), the sophistication of sentence structure (increasingly complex and inclusion of more varied sentence structures), stamina (completeness of the picture descriptions), and coherence/cohesion of the description. With regard to the latter, as a form of discourse, descriptions may exceed the boundaries of a single sentence, requiring both coherence and cohesion among sentences and ideas. Together, coherence and cohesion assist the listener with understanding a student's

description. Coherence features included temporal discourse connectors (e.g., "first", or "next" in English; "d'abord," or "en suite" in French) and location markers (e.g., "on", or "next to" in English; "sur", or "à côté de" in French) to logically organize information in descriptions. Cohesion features included referential ties across and within sentences, such as use of pronouns to refer to previously introduced nouns and ellipses to create a parsimonious description. The presence of both coherence and cohesion markers was needed for a description to be placed beyond the *Early Emerging* phase, as explained below. After reviewing the French samples and given the low French proficiency of most participants, the phases of the original DLLP were expanded to include additional phases at the early stages of the progression. As can be seen in Appendix A (see OSF page), each linguistic feature except for coherence/cohesion could thus be placed at one of seven phases of development: *Not yet evident, Pre-emergent, Early emerging, Emerging, Early developing, Developing,* or *Controlled.* Coherence/cohesion could be placed at one of five phases: *Not yet evident, Early emerging, Emerging, Developing,* or *Controlled.*

Three raters proficient in both French and English compared the language features of each picture description transcript to the corresponding features on the language progression and selected the phase that best matched. Rater 1 served as the anchor rater and coded all the transcripts. Raters 2 and 3 each independently coded half of the transcripts. Consensus was reached whenever there was a disagreement between Rater 1 and the other rater. The number of agreements between the two ratings was divided by the total number of agreements and disagreements before consensus. The proportion of inter-rater agreements (IRA) between Raters 1 and 2 ranged from 80% to 90% in French and 73% to 95% in English. The proportion of IRA between Raters 1 and 3 ranged from 88% to 96% in French and 76% to 91% in English. The majority of IRA calculated comfortably surpass the threshold (75%) set as acceptable by Chaturvedi and Shweta (2015), particularly in the case of the large number of adjacent levels.

Analysis

Bayesian framework overview

Many perspectives exist that underlie the scientific rationale for using Bayesian methods (Levy & McNeish, 2021). For example, a model may be too complex for frequentist estimation methods, such as maximum likelihood. In such instances, researchers view prior specification as not something integral to their scientific question but rather just a necessary step to carry out Bayesian estimation. Other researchers instead see prior specifications as central to the scientific question at hand. The perspective we adopt here can be viewed as somewhere in the middle, as our incorporation of prior information is neither an afterthought nor is it central to answering our research questions. Instead, our use of prior information, which we discuss in Appendix B (see OSF page), is intended to (a) help our models converge and (b) protect our inferences from implausible findings. Priors used for this purpose are commonly referred to as weakly informative priors (e.g., Gelman et al., 2008). What constitutes a weakly informative prior varies based on the statistical model and the available substantive information. To illustrate, a critical aspect of our prior information incorporation involved applying half-Cauchy priors to

random effect standard deviations. This prior choice stems from our understanding that negative standard deviations are impossible. The use of such priors holds particular significance in our current study due to the frequent occurrence of negative standard deviations when employing frequentist estimation methods with limited sample sizes.

Before answering research questions, model comparisons were conducted. For research question 1, model comparisons helped us choose between the *random linear growth* and *random quadratic growth* models to then (a) inspect the point estimates and 95% credible intervals (CIs) and (b) visualize the growth trajectories. For research question 2, model comparisons helped us choose between models 1 and 2 for the inspection of point estimates and 95% CIs corresponding to the effects of *grade* and *home language.* In this paper, models were compared using the expected log pointwise predictive density (ELPD; Vehtari et al., 2017) a measure of overall fit and out-of-sample predictive accuracy (Bendixen & Purzycki, 2022). Using the *loo* package in *R*, the ELPD and its standard error were calculated using approximate leave-one-out cross-validation (Vehtari et al., 2018). We do not focus on the ELPD itself but rather on the ELPD difference and its standard error in order to facilitate model comparisons[3].

Besides model comparisons, each research question was assessed with reference to summaries of the posterior (i.e., posterior means and 95% CIs). Like frequentist point estimates, the Bayesian posterior mean is a single-number estimate of the strength of association between two variables. For example, γ_{01} is the effect of being in first grade as opposed to TK/K. Positive values of γ_{01} denote that being in first grade is associated with an increased probability of high ratings for each language feature, whereas negative values of γ_{01} denote that being in first grade is associated with a lower probability of high ratings on each language feature. Similarly, γ_{02} is the effect of home language background. For the English progressions, this means the effect of being an English-only household. However, for the French progressions, this means the effect of having any French spoken in the household. Positive values of γ_{02} denote that being from an English-only household is associated with an increased probability of high ratings for each language feature for pictures described in English. Positive values of γ_{02} denote that being from a French-speaking household is associated with an increased probability of high ratings for each language feature for pictures described in French. Like frequentist point estimates, Bayesian posterior means should be assessed jointly with measures of uncertainty (i.e., 95% CIs) to infer whether the association is "significant" or "meaningful." Random effect standard deviations and correlations are considered meaningful when their 95% CI does not overlap with zero. Because all other parameters are reported on the odds scale, they are considered meaningful when their 95% CI does not overlap with one. Using CIs, we can state that there is a 95% probability that a parameter's true value lies between the intervals' upper and lower bounds. In addition to being useful for hypothesis testing, CIs help assess the variability of an effect. The wider the CI, the less certain we are about the magnitude of the effect.

Models

All research questions were assessed using a combination of both models presented below, each representing some form of a hierarchical ordinal logistic regression model (Hedeker & Gibbons, 1994). Our use of ordinal regression is justified because DLLP progressions may be thought of as coarse representations of continuous underlying variables (Bauer & Sterba, 2011). In addition, treating ordinal outcomes as continuous may be problematic, and as the descriptives in Figure 1 show, our outcomes are not normally distributed (Ali et al., 2016; Bauer & Sterba, 2011; Hung & Huang, 2011). Finally, treating ordinal outcomes as continuous in scenarios when data are nested may exacerbate issues related to inflated or spurious estimates of random slope and random quadratic variances (see Bauer & Cai, 2009; and Bauer & Sterba, 2011, for more information), both of which are relevant to RQ1 concerning DLLP growth patterns.

Instead of assuming the outcome is continuous as in traditional linear regression models, the outcome in ordinal regression[4] models is treated as an ordered set of discrete responses (e.g., a Likert scale response). Changes in predictor variables result in shifts toward either end of the ordinal scale, which, in the context of our study, means that positive coefficients increase the probability of higher DLLP ratings on the learning progression (i.e., *Early developing, Developing*, and *Controlled*) as opposed to lower DLLP ratings (i.e., *Not yet evident, Pre-emergent, Early emerging,* and *Emerging*).

We use ordinal logistic regression (McCullagh, 1980), which is also referred to as the cumulative logit model or proportional odds model and is one of several models available for ordinal data. A benefit of ordinal logistic regression is that its coefficients may be transformed into cumulative odds ratios, which makes point estimates and CIs easier to interpret. An important assumption of this model is that all coefficients estimated are invariant to each ordinal outcome. Put differently, it is assumed that the effect of being in Grade 1 as opposed to TK/K is the same across all language progressions, or that the effect of home language background is the same across all language progressions.

Because our data are nested (i.e., assessment occasion nested within students), traditional ordinal regression would violate the independence assumption (e.g., Bauer & Sterba, 2011), hence our decision to use a hierarchical ordinal regression model, which can easily accommodate the complex growth models considered in our analyses.

Equations depicting the two models fit to each language feature can be seen in Table 1. Both models allow for student-level intercepts β_{0j} to vary, which reflects our prior belief that baseline propensity to receive a high rating varies considerably across this sample of students (see Ryan, 2021a, 2021b). Further, as can be seen in each model's level 2 equation, each student's baseline propensity to receive a high rating is a function of Grade (0 = TK/K; 1 = Grade 1) and Home Language effect (French-speaking or English-only-speaking household), denoted as γ_{01} and γ_{02}, respectively. Both γ_{01} and γ_{02} were used to answer RQ2[5]. Additionally, the coefficients β_{2j} (Model 1) and β_{3j} (Model 2) in the level 1 equations depict the change in expected propensity to receive a high rating when assessed with the school versus the home picture, which we could control for when interpreting coefficients
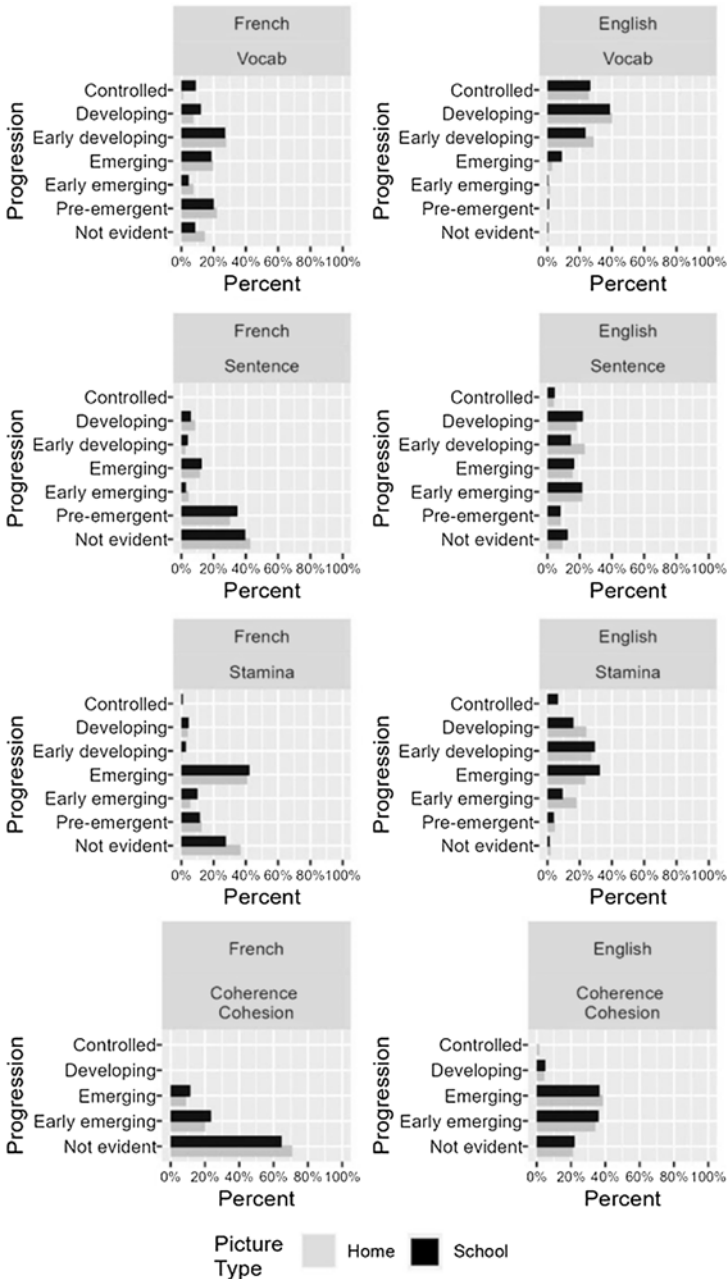
Figure 1. Overall empirical distribution of language progressions by language and picture.

relevant to our research questions. Finally, both models' level 2 equations for β_{0j} specify that student-level intercepts are allowed to vary. The average student intercept γ_{00} is a *fixed effect*, and random student deviations u_{0j} are *random effects*.

Table 1. Equations

		Model 1 – Random linear growth	Model 2 – Random quadratic growth
Level 1 Equation	Linear predictor for student j at time t	$\eta_{tj} = \beta_{0j} + \beta_{1j}MonthsElapsed_{tj} + \beta_{2j}PictureType_{tj}$	$\eta_{tj} = \beta_{0j} + \beta_{1j}MonthsElapsed_{tj} + \beta_{2j}MonthsElapsed^2_{tj} + \beta_{3j}PictureType_{tj}$
Level 2 Equation	Baseline proficiency for student j	$\beta_{0j} = \gamma_{00} + \gamma_{01}Grade_j + \gamma_{02}HomeLanguage_j + u_{0j}$	$\beta_{0j} = \gamma_{00} + \gamma_{01}Grade_j + \gamma_{02}HomeLanguage_j + u_{0j}$
	Linear growth rate for student j	$\beta_{1j} = \gamma_{10} + u_{1j}$	$\beta_{1j} = \gamma_{10} + u_{1j}$
	Quadratic growth rate for student j	*Not Applie*	$_2 = \gamma_{20} + u_{2j}$
	Variance covariance matrix	where $\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \right),$ & $\tau_{10} = \tau_{01}$	where $\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{10} & \tau_{11} & \tau_{12} \\ \tau_{20} & \tau_{21} & \tau_{22} \end{bmatrix} \right),$ & $\tau_{10} = \tau_{01}, \ \tau_{20} = \tau_{02}, \ \& \ \tau_{12} = \tau_{21}$

Note. η_{it} denotes the jth student's linear predictor at time t, where $j = 1\ldots42$ and $t = 1\ldots5$. With the help of $C = L_O - 1$ threshold parameters, v, or $v^1, \ldots v^c, \ldots v^C$, where L_0 is the number of levels observed in our data set (see Appendix B), the linear predictor can be used to predict the cumulative probability that the jth student at time t receives a rating of c or lower $\pi_{jt}^c = \frac{exp(v^c - \eta_{jt})}{1 + exp(v^c - \eta_{jt})}$. Because threshold parameters v and the average student intercept γ_{00} are not jointly identified, we set γ_{00} to zero in both models.

Instead of individual random effects, a variance component τ_{00} is estimated. High values of τ_{00} indicate that students differ greatly in baseline proficiency, whereas values of τ_{00} approaching zero indicate that student baseline proficiency is relatively homogeneous.

Differences between the two models can be seen in both the level-1 and level-2 equations in Table 1, as each model parameterizes growth differently. Model 1, the *random linear growth* model, contains three parameters that pertain to growth. The first is the average rate of linear growth γ_{10}, which is the average, per-month, linear rate of growth expected when controlling for all other variables in the model, i.e., *Grade*, *Home Language*, and *Picture Type*. This parameter may be found in Table 1, and it is added to student-level differences in linear growth u_{1j} to produce student-level growth rates β_{1j} (i.e., $\beta_{1j} = \gamma_{10} + u_{1j}$). With γ_{10} as the average, or expected, rate of linear growth, the student variation in linear growth is governed by the second growth parameter of the random linear growth model, $Var(u_{1j}) = \tau_{11}$. In other words, τ_{11} is the linear growth variance component, and it represents the variance in the rate of linear growth among students. Large values of τ_{11} indicate considerable student-level differences in growth, and low values indicate that most students' growth rates β_{1j} are close to γ_{10}. The third growth parameter of the random linear growth model, $Cov(u_{0j}, u_{1j}) = \tau_{01} = \tau_{10}$, is the covariance between student variation in baseline proficiency u_{0j} and linear growth u_{1j}. A positive value of τ_{01} indicates that students with a high initial proficiency β_{0j} also have high rates of linear growth β_{1j}, and a negative τ_{01} would indicate that students with a low β_{0j} tend to have higher β_{1j}.

In addition to the three growth parameters γ_{10}, τ_{11}, and τ_{01} found in model 1, model 2, the *random quadratic growth* model, contains four additional growth parameters, namely γ_{20}, τ_{22}, τ_{02}, τ_{12}. The first, γ_{20}, which is the population average quadratic effect, represents the degree to which growth over the 11 months of programming decelerates or accelerates. More specifically, positive values of γ_{20} indicate that average student growth accelerates over time, and negative values of γ_{20} indicate that student growth decelerates over time. The remaining three growth parameters can all be seen in the variance covariance matrix in Table 1. First, $Var(u_{2j}) = \tau_{22}$ represents the amount of between-student variation in levels of growth in terms of acceleration or deceleration. Finally, $Cov(u_{0j}, u_{2j}) = \tau_{02} = \tau_{20}$ represents the amount of covariation between student-level differences in the intercept and the quadratic effect, while $Cov(u_{1j}, u_{2j}) = \tau_{12} = \tau_{21}$ represents the amount of covariation between the student-level linear growth rate and the quadratic effect.

Results

Each of the two models was fit to each language (i.e., French and English) by language feature combination (i.e., vocabulary sophistication, sentence sophistication, stamina, and coherence/cohesion), resulting in 16 total sets of results. All models were estimated in *R* (R Core Team, 2021) using the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2022), which implements the No-U-turn (NUTS; Hoffman and Gelman, 2014; Betancourt, 2017), an adaptive variation of Hamiltonian Monte Carlo (HMC)

sampling. Additionally, the *R* package brms (Bürkner, 2017) was used to generate stan scripts. Each set of results was estimated using 6000 posterior samples, 2000 of which were warm-up draws[6]. Further, to check the validity of the posterior means and 95% CIs, four chains were used, and results indicated convergence. To ensure the model data fit, posterior predictive checks were performed and visually inspected. We found that all models fit the data adequately well. Finally, to ensure our findings are robust to our choice of priors, we conducted a sensitivity analysis and found that the use of an alternative set of priors had no impact on our results. (For more information on convergence information, posterior predictive checks, and our sensitivity analysis, see Appendices C, D, and E on the OSF page, which also contains the data and syntax used for our analyses.)

Bayesian analysis results

RQ 1a – French growth patterns

To assess growth patterns in French oral language across the duration of the study, we first inspected ELPD differences, which suggested that all four language features were best explained by a random linear growth model. Consequently, the random linear growth model was selected to answer all research questions pertaining to French oral language development.

Of the four language features evaluated in French, coherence/cohesion's average linear growth rate was the smallest (see Table 2, Random Linear Growth Model: $\widehat{\gamma}_{10}$ = 1.135). Nevertheless, its 95% CI contains only positive values, implying that students on average progressed in French coherence/cohesion over the eleven months of programming, though it is unclear how meaningful such growth was (95% CI = [1.003,1.283]).

The average linear growth rates estimated for sentence structure (see Table 3, Random Linear Growth Model: $\widehat{\gamma}_{10}$ = 1.326, 95% CI = [1.166, 1.527]), stamina (see Table 4, Random Linear Growth Model: $\widehat{\gamma}_{10}$ = 1.417, 95% CI = [1.278, 1.592]), and vocabulary (see Table 5, Random Linear Growth Model: $\widehat{\gamma}_{10}$ = 1.365, 95% CI = [1.252, 1.493]) were all noticeably larger than for coherence/cohesion. Indeed, as Figure 2 shows, from the beginning to the end of the study duration, the percentage of students receiving a higher rating increased. For example, as Figure 2 shows that at baseline, few (if any) students were expected to receive a rating of *emerging*, whereas the percentage of students receiving such a rating increased after 11 months. Similarly, at baseline, the majority of students were expected to receive a rating of *not evident* for all language features, unlike at the end of the study, except in the case of coherence/cohesion.

While meaningful positive growth was found in all four features in French, there were important differences in the relationship between individual differences in student baseline proficiency u_{0j} and linear growth u_{1j}. As can be seen in Tables 2 through 5, the correlations between u_{0j} and u_{1j} were negative for all four language features, though this negative relationship was only found to be meaningful for sophistication in sentence structure (see Table 3, Random Linear Growth Model: $\widehat{\rho}_{01}$ = −0.62, 95% CI = [−0.849, −0.269]) and for vocabulary (see Table 5, Random Linear Growth Model: $\widehat{\rho}_{01}$ = −0.72, 95% CI = [−0.928, −0.385]). In other words, while both positive and meaningful growth was observed in French overall, sentence

Table 2. Coherence/Cohesion posterior means and 95% credible intervals

Parameter	French				English			
	Random linear growth		Random quadratic growth		Random linear growth		Random quadratic growth	
	Posterior mean	95% CI	Posterior mean	95% CI	Posterior mean	95% CI	Posterior mean	95% CI
Average growth rates								
Linear growth $\widehat{\gamma}_{10}$	1.135*	1.003, 1.284	1.586*	1.117, 2.296	1.016	0.917, 1.129	1.048	0.813, 1.353
Quadratic growth $\widehat{\gamma}_{20}$			0.967*	0.935, 1			0.997	0.972, 1.022
Level-2 fixed effects								
Grade $\widehat{\gamma}_{01}$	2.728	0.618, 13.241	3.084	0.611, 16.301	3.215	0.633, 16.71	3.258	0.617, 16.85
Home language $\widehat{\gamma}_{02}$	8.135*	1.745, 39.404	9.251*	1.871, 49.053	0.418	0.075, 2.172	0.434	0.077, 2.342
Random effect standard deviations								
Intercept $\widehat{\tau}_{00}^{1/2}$	8.361*	3.685, 24.646	9.865*	4.103, 33.568	12.617*	6.317, 29.833	12.495*	6.177, 30.357
Linear growth $\widehat{\tau}_{11}^{1/2}$	1.244*	1.083, 1.446	1.327*	1.023, 2.142	1.283*	1.158, 1.443	1.241*	1.016, 1.613
Quadratic growth $\widehat{\tau}_{22}^{1/2}$			1.022*	1.001, 1.072			1.022*	1.001, 1.05
Random effect correlations								
Intercept – Linear growth $\widehat{\rho}_{01}$	−0.092	−0.626, 0.569	−0.165	−0.726, 0.556	−0.153	−0.559, 0.32	0.033	−0.593, 0.693
Intercept – Quadratic growth $\widehat{\rho}_{02}$			0.024	−0.679, 0.718			−0.183	−0.772, 0.503
Linear growth – Quadratic growth $\widehat{\rho}_{21}$					−0.262	−0.922, 0.648	−0.185	−0.852, 0.664

Note. All parameter estimates are on the odds scale except for random effect standard deviations and random effect correlations. Posterior means corresponding to parameters on the odds scale are denoted with an asterisk when accompanied with a 95% that does not contain one. Posterior means corresponding to random effect standard deviations and random effect correlations are denoted with an asterisk when accompanied with a 95% CI that does not contain zero.

Table 3. Sophistication of sentence structure posterior means and 95% credible intervals

Parameter	French				English			
	Random linear growth		Random quadratic growth		Random linear growth		Random quadratic growth	
	Posterior Mean	95% CI	Posterior Mean	95% CI	Posterior Mean	95% CI	Posterior Mean	95% CI
Average growth rates								
Linear growth $\widehat{\gamma}_{10}$	1.325*	1.165, 1.524	2.111*	1.49, 3.069	0.909*	0.826, 0.998	0.793	0.527, 1.186
Quadratic growth $\widehat{\gamma}_{20}$			0.957*	0.928, 0.987			1.013	0.974, 1.053
Level-2 fixed effects								
Grade $\widehat{\gamma}_{01}$	13.548*	2.314, 91.64	15.523*	2.337, 111.619	1.827	0.318, 10.692	1.846	0.241, 14.459
Home language $\widehat{\gamma}_{02}$	70.396*	11.032, 500.633	83.079*	11.643, 678.39	0.5	0.079, 2.926	0.507	0.06, 4.351
Random effect standard deviations								
Intercept $\widehat{\tau}_{00}^{1/2}$	32.951*	12.186, 121.513	48.899*	15.311, 247.242	15.012*	7.497, 36.465	29.149*	11.843, 88.61
Linear growth $\widehat{\tau}_{11}^{1/2}$	1.386*	1.229, 1.604	1.536*	1.143, 2.383	1.263*	1.151, 1.407	3.07*	2.12, 4.714
Quadratic growth $\widehat{\tau}_{22}^{1/2}$			1.029*	1.001, 1.082			1.113*	1.072, 1.162
Random effect correlations								
Intercept – Linear growth $\widehat{\rho}_{01}$	−0.62*	−0.848, −0.271	−0.573*	−0.881, −0.012	−0.072	−0.46, 0.356	−0.288	−0.583, 0.063
Intercept – quadratic growth $\widehat{\rho}_{02}$			−0.055	−0.755, 0.669			0.268	−0.09, 0.573
Linear growth – quadratic growth $\widehat{\rho}_{21}$			−0.304	−0.891, 0.627			−0.951*	−0.982, −0.889

Note. All parameter estimates are on the odds scale except for random effect standard deviations and random effect correlations. Posterior means corresponding to parameters on the odds scale are denoted with an asterisk when accompanied with a 95% that does not contain one. Posterior means corresponding to random effect standard deviations and random effect correlations are denoted with an asterisk when accompanied with a 95% CI that does not contain zero.

Table 4. Stamina posterior means and 95% credible intervals

Parameter	French				English			
	Random linear growth		Random quadratic growth		Random linear growth		Random quadratic growth	
	Posterior Mean	95% CI	Posterior Mean	95% CI	Posterior Mean	95% CI	Posterior Mean	95% CI
Average growth rates								
Linear growth $\widehat{\gamma}_{10}$	1.416*	1.275, 1.588	1.356*	1.03, 1.8	1.064	0.964, 1.175	1.19	0.918, 1.559
Quadratic growth $\widehat{\gamma}_{20}$			1.004	0.978, 1.032			0.989	0.966, 1.012
Level-2 fixed effects								
Grade $\widehat{\gamma}_{01}$	10.936*	2.104, 61.48	10.953*	2.009, 59.536	7.298*	1.853, 29.483	7.5*	1.842, 31.745
Home language $\widehat{\gamma}_{02}$	23.88*	4.355, 137.47	25.137*	5.006, 147.708	1.323	0.307, 5.445	1.257	0.279, 5.585
Random effect standard deviations	17.13*	7.602, 48.197						
Intercept $\widehat{\tau}_{00}^{1/2}$	1.243*	1.103, 1.412	16.248*	7.446, 45.794	11.502*	6.076, 25.686	12.936*	6.388, 31.263
Linear growth $\widehat{\tau}_{11}^{1/2}$			1.197*	1.01, 1.568	1.293*	1.18, 1.446	1.529*	1.193, 2.319
Quadratic growth $\widehat{\tau}_{22}^{1/2}$			1.018*	1.001, 1.045			1.024*	1.001, 1.066
Random effect correlations								
Intercept – Linear growth $\widehat{\rho}_{01}$	−0.464	−0.81, 0.026	−0.237	−0.805, 0.56	−0.438*	−0.718, −0.064	−0.424	−0.735, 0.006
Intercept – Quadratic growth $\widehat{\rho}_{02}$			−0.243	−0.793, 0.549			0.131	−0.67, 0.735
Linear growth – Quadratic growth $\widehat{\rho}_{21}$			−0.174	−0.869, 0.68			−0.59	−0.975, 0.468

Note. All parameter estimates are on the odds scale except for random effect standard deviations and random effect correlations. Posterior means corresponding to parameters on the odds scale are denoted with an asterisk when accompanied with a 95% that does not contain one. Posterior means corresponding to random effect standard deviations and random effect correlations are denoted with an asterisk when accompanied with a 95% CI that does not contain zero.

Table 5. Sophistication of topic vocabulary posterior means and 95% credible intervals

Parameter	French				English			
	Random linear growth		Random quadratic growth		Random linear growth		Random quadratic growth	
	Posterior Mean	95% CI	Posterior Mean	95% CI	Posterior Mean	95% CI	Posterior Mean	95% CI
Average growth rates								
Linear growth $\widehat{\gamma}_{10}$	1.368*	1.257, 1.497	1.784*	1.385, 2.32	1.108*	1.026, 1.202	1.062	0.832, 1.362
Quadratic growth $\widehat{\gamma}_{20}$			0.975*	0.952, 0.998			1.004	0.981, 1.028
Level-2 fixed effects								
Grade $\widehat{\gamma}_{01}$	5.031*	1.506, 17.212	5.05*	1.455, 17.737	3.215	0.633, 16.71	3.258	0.617, 16.85
Home language $\widehat{\gamma}_{02}$	19.41*	5.515, 72.508	19.382*	4.907, 79.794	0.418	0.075, 2.172	0.434	0.077, 2.342
Random effect standard deviations								
Intercept $\widehat{\tau}_{00}^{1/2}$	11.861*	6.189, 25.575	13.422*	6.474, 34.531	6.793*	3.883, 13.521	6.989*	3.854, 14.95
Linear growth $\widehat{\tau}_{11}^{1/2}$	1.2*	1.103, 1.317	1.273*	1.033, 1.892	1.173*	1.047, 1.303	1.217*	1.025, 1.64
Quadratic growth $\widehat{\tau}_{22}^{1/2}$			1.016*	1.001, 1.056			1.009*	1, 1.035
Random effect correlations								
Intercept – Linear growth $\widehat{\rho}_{01}$	−0.72*	−0.919, −0.381	−0.613	−0.915, 0.04	−0.435	−0.784, 0.09	−0.407	−0.808, 0.258
Intercept – Quadratic growth $\widehat{\rho}_{02}$			−0.031	−0.786, 0.733			0.053	−0.718, 0.782
Linear Growth – Quadratic Growth $\widehat{\rho}_{21}$			−0.325	−0.946, 0.628			−0.282	−0.951, 0.645

Note. All parameter estimates are on the odds scale except for random effect standard deviations and random effect correlations. Posterior means corresponding to parameters on the odds scale are denoted with an asterisk when accompanied with a 95% that does not contain one. Posterior means corresponding to random effect standard deviations and random effect correlations are denoted with an asterisk when accompanied with a 95% CI that does not contain zero.
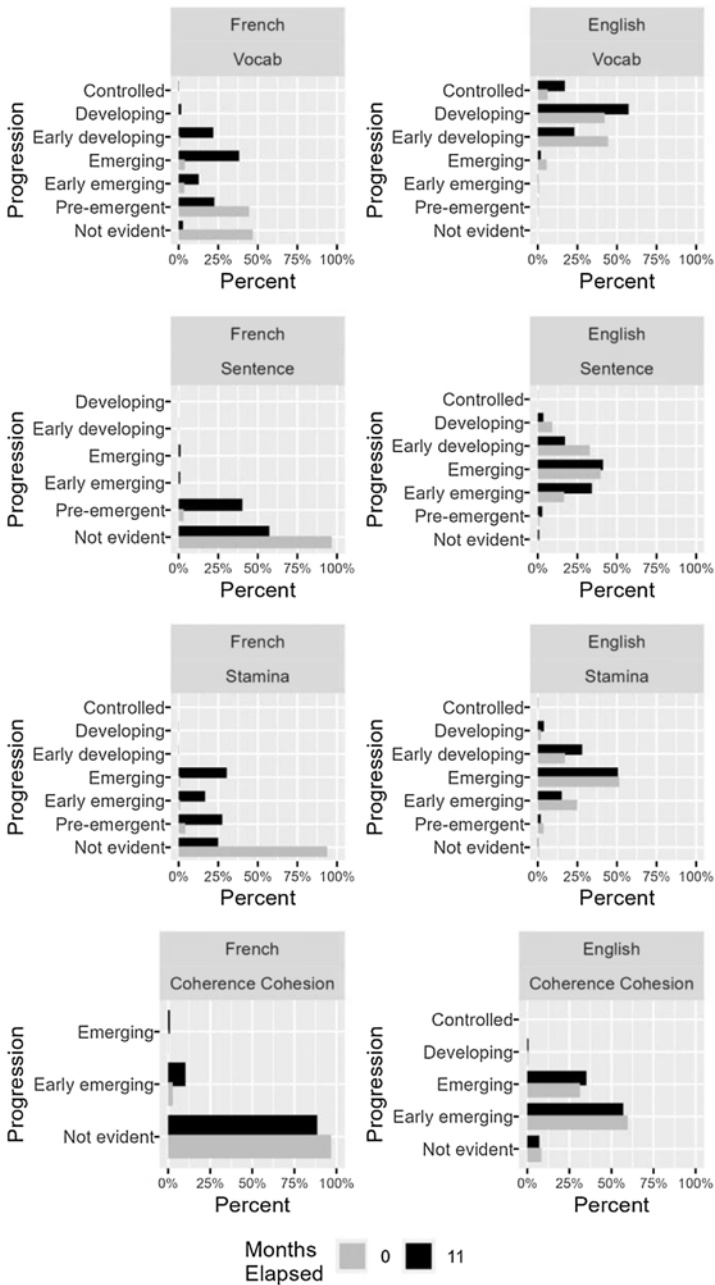
Figure 2. Model predicted distribution of language progressions with time as a covariate.

structure and vocabulary growths were different in that students with higher baseline ratings in these two features tended to progress at slower rates than students with lower baseline ratings.

RQ 1b – English growth patterns
To assess growth patterns in English oral language across the duration of the study, we first inspected ELPD differences, which suggested that a random linear growth model best explained coherence/cohesion, stamina, and vocabulary sophistication. In contrast, sentence sophistication was best explained by a random quadratic growth model. Given this, the random linear growth model was selected to answer research questions relevant to all language features except sentence sophistication, where the random quadratic growth model was used.

Of the four language features assessed in English, vocabulary is the only one with a meaningful positive linear growth trajectory parameter, indicating that with each month of programming, the odds that students received a higher rating compared to all lower ratings combined increased by an average of 10.8% (see Table 5, Random Linear Growth Model: $\widehat{\gamma}_{10} = 1.108$, 95% CI = [1.026, 1.202]). By comparison, the average linear growth for coherence/cohesion (see Table 2, Random Linear Growth Model: $\widehat{\gamma}_{10} = 1.016$, 95% CI = [0.917, 1.129]) and stamina (see Table 4, Random Linear Growth Model: $\widehat{\gamma}_{10} = 1.064$, 95% CI = [0.964, 1.175]) were both positive but not meaningful (i.e., their CIs overlapped with 1). Finally, as mentioned earlier, sophistication of sentence structure was the only feature best explained by the quadratic growth model, with a negative average linear growth rate and a positive average quadratic growth rate (see Table 3, Quadratic Growth Model: $\widehat{\gamma}_{10} = 0.793$, 95% CI = [0.527, 1.185]; $\widehat{\gamma}_{20} = 1.013$, 95% CI = [0.974, 1.053]). In other words, the odds of receiving a low rating compared to all higher ratings combined increased slightly over the study duration, though the rate at which this occurred decelerated over time.

As Figure 2 shows, the percentage of students receiving a higher rating from the beginning to the end of the study duration increased only for the vocabulary feature. At baseline, the majority of students were expected to receive a rating of *Emerging* and *Early Developing*, whereas by the end of the study, the expected percentage of students receiving a rating of *Emerging* decreased considerably, and the expected percentage of students receiving a rating of *Early Developing* and *Controlled* noticeably increased.

RQ 1c – Relationship between French and English growth patterns
We used Spearman product moment correlations with the SpearmanCI package (de Carvalho, 2018) to investigate the relationship between French and English across all four language features. Table 6 shows the five sets of correlations between French and English ratings at each individual wave. Correlations are accompanied by 95% confidence intervals, and when intervals do not overlap with 0, correlations are denoted with an asterisk. As can be seen, both vocabulary and coherence/cohesion appear to have a weak to moderate association between languages at waves 2, 3, and 5. Stamina in French is correlated with stamina in English at waves 2–5. Sentence structure, on the other hand, only appears to be correlated between French and English at the last two waves. While these results do not point to definitive positive relationships across all language features, overall, these results point to a possible weak to moderate positive correlation between French and English,

Table 6. Spearman correlations between French and English LFAP ratings by data collection wave

Language Feature	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Vocabulary	0.12 (−0.11, 0.35)	0.37* (0.18, 0.57)	0.39* (0.20, 0.59)	0.19 (−0.04, 0.42)	0.50* (0.33, 0.68)
Sentence	−0.08 (−0.30, 0.15)	0.12 (−0.12, 0.35)	0.18 (−0.05, 0.41)	0.39* (0.18, 0.60)	0.53* (0.38, 0.69)
Stamina	−0.02 (−0.22, 0.26)	0.34* (0.11, 0.56)	0.29* (0.06, 0.51)	0.27* (0.06, 0.50)	0.29* (0.03, 0.57)
Coherence Cohesion	0.12 (−0.10, 0.35)	0.32* (0.12, 0.52)	0.29* (0.09, 0.50)	0.05 (−0.19, 0.30)	0.43* (0.24, 0.63)

Note. Correlations are denoted with an asterisk when accompanied with a 95% confidence interval that does not contain zero.

suggesting that growth in language features in one language is generally associated with growth in the other language.

RQ 2a – Associations between French progressions and grade level/household language background

After selecting a model based on the ELPD difference for each language feature assessed in French, we then inspected the posterior mean and 95% CIs of the level-2 fixed effects for Grade $\widehat{y}_{01}$ and Home Language $\widehat{y}_{02}$. For all language features assessed in French, coming from a French-speaking household (i.e., Home Language = 1) was associated with greater odds of receiving a higher rating when controlling for other variables in the model (i.e., grade and linear growth). By far, the effect of having French spoken in the home was largest for sophistication of sentence structure (see Table 3, Random Linear Growth Model: $\widehat{y}_{02} = 70.396$, 95% CI = [11.032, 500.633]). However, CIs suggest the increase in odds due to having French spoken in the household could plausibly range from a value as low as 11.032 to one as large as 500.633. Similarly, wide 95% CIs were observed for coherence/cohesion (see Table 2, Random Linear Growth Model: $\widehat{y}_{02} = 8.135$, 95% CI = [1.175, 39.404]), stamina (see Table 4, Random Linear Growth Model: $\widehat{y}_{02} = 23.88$, 95% CI = [4.355, 137.47]), and sophistication of topic vocabulary (see Table 5, Random Linear Growth Model: $\widehat{y}_{02} = 19.41$, 95% CI = [5.515, 72.505]).

Similarly, being in first grade (i.e., Grade = 1) was meaningfully associated with greater odds of receiving a high rating in most features in French after controlling for home language and linear growth. This can be seen in the sophistication of sentence structure feature (see Table 3, Random Linear Growth Model: $\widehat{y}_{01} = 13.548$, 95% CI = [2.314, 91.64]), the stamina feature (see Table 4, Random Linear Growth Model: $\widehat{y}_{01} = 10.936$, 95% CI = [2.104, 61.48]), and the sophistication of topic vocabulary feature (see Table 5, Random Linear Growth Model: $\widehat{y}_{01} = 4.927$, 95% CI = [1.538, 15.94]). However, being in first grade was not meaningfully associated with greater odds of receiving a high rating in coherence/cohesion (see Table 2, Random Linear Growth Model: $\widehat{y}_{01} = 2.728$, 95% CI = [0.618, 13.214]).

To summarize, with the exception of coherence/cohesion, where only home language was meaningful, both grade and home language were meaningful predictors in all French language progressions. However, the patterns of association differed across the two predictors. Specifically, across all language progressions, the increased odds of receiving a higher rating due to being in first grade were noticeably smaller than the increased odds associated with having French spoken at home. Figures 2, 3 and 4 depict the model-predicted distributions conditional on values of time, grade, and home language. For example, with regards to sentence sophistication in French, first graders were roughly five times more likely to be placed at *pre-emergent* on the DLLP than TK/K students, and coming from a French-speaking household increased by 12% the chances of being placed at *pre-emergent* as opposed to *not evident* on the DLLP.
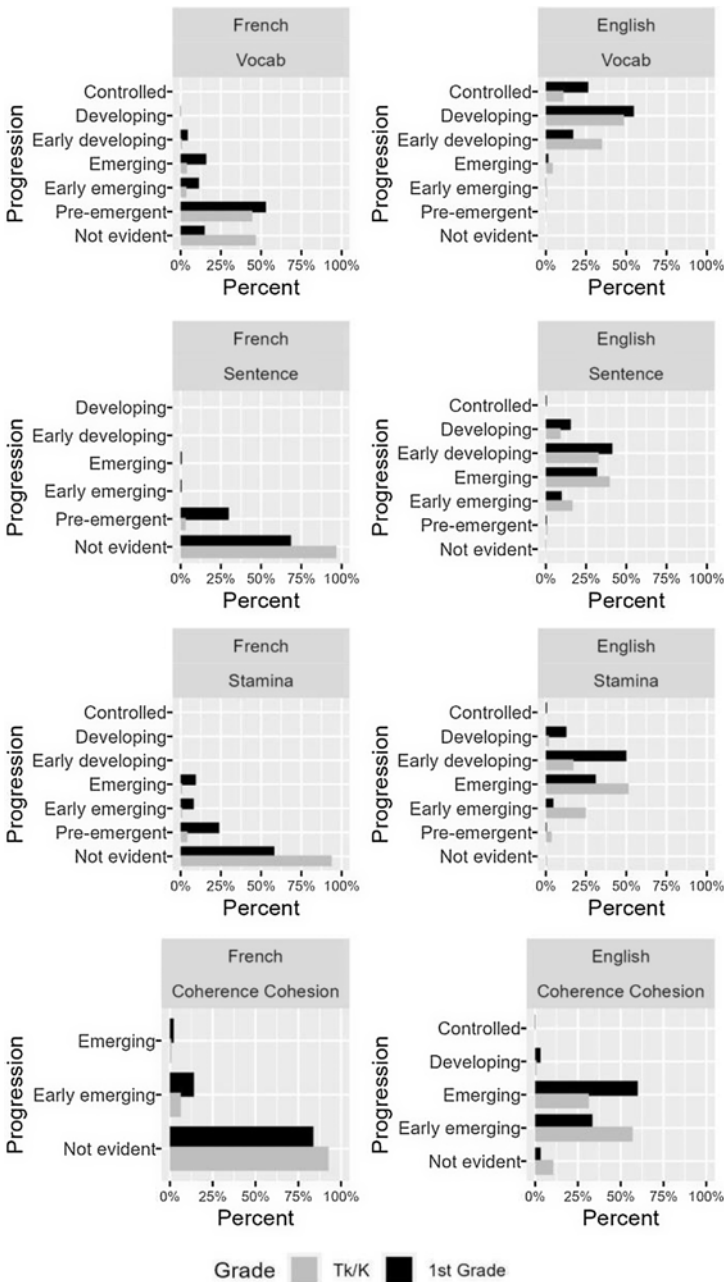
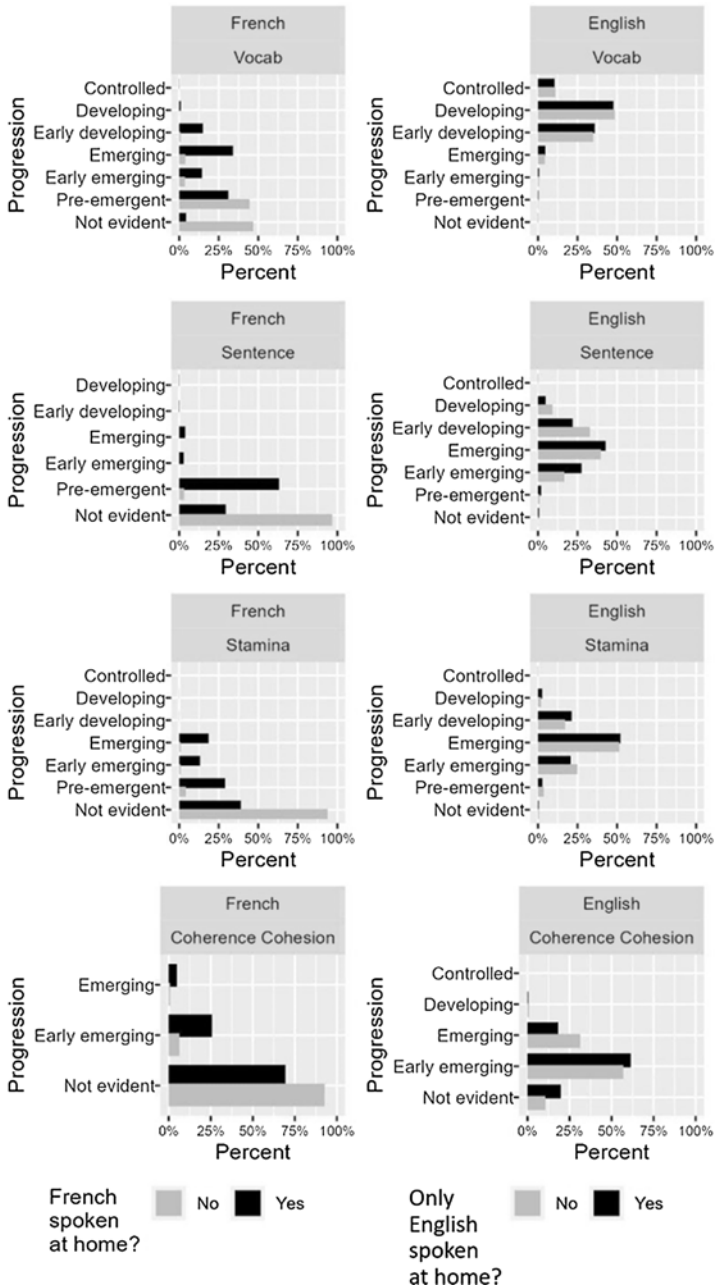**Figure 3.** Model predicted distribution of language progressions with grade as a covariate.

Figure 4. Model predicted distribution of language progressions with home language as a covariate.

RQ 2b – Associations between English progressions and grade level/household language background

After selecting a model based on the ELPD difference for each language feature assessed in English, we inspected the posterior mean and 95% CIs of the level-2 fixed effects for Grade $\widehat{y}_{01}$ and Home Language $\widehat{y}_{02}$. Figures 3 and 4 depict the model-predicted distributions conditional on the values of grade and home language. Coming from an English-only household (i.e., Home Language = 1) did not play a meaningful role in English progressions. In contrast, being in first grade (i.e., Grade = 1) was meaningfully associated with greater odds of receiving a high rating in stamina (see Table 4, Random Linear Growth Model: $\widehat{y}_{02} = 7.298$, 95% CI = [1.853, 29.483]).

Results summary

To answer RQ1, the Random Linear Growth model was selected for all language by DLLP feature combinations, except for English sentence structure, where a random quadratic growth model was selected. For all language features assessed in French, estimates of average linear growth trajectories γ_{10} are all positive and have 95% CIs that do not overlap with one, which suggests that, on average, participants' probability of receiving higher ratings increased over the duration of the study. In contrast, in English, meaningful growth was only found in vocabulary, which suggests that, on average, participants' probability of receiving higher ratings over the duration of the study only increased for vocabulary.

To answer RQ2, a random linear growth model was selected for all language features assessed in French. Coefficients corresponding to the effect of grade and home language were all positive and had 95% CIs containing only positive values, implying that being in first grade and coming from a French-speaking household were both meaningfully associated with a greater probability of receiving high ratings on all language features assessed in French. For English, a random linear growth model was selected for coherence/cohesion, sophistication of topic vocabulary, and stamina. A random quadratic growth model was selected for sophistication of sentence structure. Coming from an English monolingual household was not meaningfully associated with the probability of receiving higher ratings on any of the language features assessed in English. However, for the stamina feature, coefficients corresponding to the effect of grade were positive and had 95% CIs containing only positive values, implying that being in first grade was meaningfully associated with a greater probability of receiving high ratings on stamina in English.

Discussion

In this study, we examined French and English oral language trajectories for early-elementary DLI students across 11 months of programming based on a picture description task. For French oral language, we found evidence of meaningful positive linear growth for all features and picture scene combinations with the smallest growth for coherence/cohesion. In contrast, for English oral language, meaningful positive growth was only detected for sophistication of topic vocabulary. Overall, coming from a French-speaking household improved French oral language

trajectories, but coming from an English-only household did not improve English oral language trajectories. In both languages, grade level influenced the trajectories of some – but not all – features.

It is worth noting that the nature of the task itself may explain the varying degrees of growth on different features. Indeed, compared to other language elicitation tasks, Nurss and Hough (1985) found that picture descriptions favored lexical skills over discourse-level skills. Similarly, in our study, we note that for both languages, growth is detected on word-level features (i.e., sophistication of topic vocabulary), but growth is much smaller on discourse-level features (e.g., coherence/cohesion), which makes sense considering that describing a picture itself does not call for discourse complexity. The small growth for coherence/cohesion echoes findings from previous research with elementary school children who were either multilingual (predominantly Spanish-English) speakers or monolingual English speakers (Bailey, 2017). Given the young age of the participants, there may be a developmental explanation. Early elementary children may not yet have sufficient forms and structures to be formulating language at the level of more sophisticated multi-utterance organization that coherence among utterances and cohesive devices within them requires. Another potential explanation is that the progression for coherence/cohesion is not precise enough to detect growth at such early phases, which implies that there are precursor aspects of the feature that are not represented in the current progression.

Findings from our research suggest that French and English language growth within individual children is not uniform across features. The fact that grade level was meaningfully associated with greater progression for most features in French confirms that there is growth in the partner language between the first and second year of the DLI program, corroborating trends that were identified with DLI students in later grades (e.g., Watzinger-Tharp et al., 2018; 2021). However, unlike previous studies with DLI students, growth was not detected for most features in English, except for sophistication of topic vocabulary and stamina. One hypothesis is that the eleven-month time frame was simply too short to detect meaningful growth in sentence structure and coherence/cohesion in English, the students' dominant language. Furthermore, at least for the vocabulary feature, the skewed distribution points to the possibility that students performed at ceiling levels. Other DLI studies have found a comparatively slow growth in English for students who start at more advanced proficiency levels (e.g., Watzinger-Tharp et al., 2021)[7].

This study also suggests great variability in baseline scores and growth rates, in line with CDST, which highlights the high level of variability that exists within and across individuals. For example, both in English and in French, individual students varied more in their baseline proficiency than in their growth rates. Additionally, there is more variability in baseline proficiency in French compared to English for vocabulary, sentence structure, and stamina. These findings suggest that there is a lot of individual variability in where students start off, especially for the nondominant language. Taken together with the findings that speaking French at home predicted improvements in French language progression, we hypothesize that the source of this variability stems from individual differences in French language experiences at home (Paradis, 2023). Furthermore, across both French and English, variability in growth rates is fairly similar, suggesting that the effect of the

education program is the same across students. Finally, for French vocabulary and sentence structure and English stamina, participants with higher baseline scores tended to experience less growth over the study duration than peers with lower baseline scores (and vice versa). It is possible that children who were already scoring high at the beginning of the task had little room to improve given the nature of the task combined with their language abilities at this developmental time point.

Furthermore, coming from a French-speaking household was meaningfully associated with greater progression for most language features in French, whereas coming from a monolingual English household was not meaningfully associated with greater progression in English. Indeed, previous studies with the same participants also showed a significant advantage in French receptive and expressive vocabulary based on levels of French exposure in the household (input, output, or literacy), an advantage that was not shared in English for children from English-only households (Ryan, 2021a; 2021b). Other studies in DLI contexts also show greater performance in the partner language for heritage language students, at least for some skills (e.g., Burkhauser et al., 2016; Lindholm-Leary & Hernández, 2011; Xu et al., 2015). It is worth noting that English was the dominant language for most children in the sample, including those from French-speaking households (Ryan, 2023). Consequently, considering their high levels of English proficiency at baseline, students may have been demotivated in English, finding the task repetitive and excessively easy.

Limitations, future work, and conclusion

As explained earlier, given the exploratory nature of this study and our choice of analyses, we used weakly informative priors that give less weight to the data than traditional uninformative priors. Using a Bayesian framework allowed us to run more complex analyses despite the small sample size of the dataset, letting models converge and protecting our inferences from impossible findings. However, despite the benefits provided by a Bayesian approach, including the flexibility to examine the results more comprehensively, having a larger dataset would still be beneficial. Additionally, the analyses we conducted treated English and French language progressions separately. However, children are learning both languages at once. Thus, a promising direction for future research could be to examine how the knowledge of each language taken together in one model shapes children's concurrent bilingual language development. For example, the use of a latent variable model such as factor analysis may help determine whether the aforementioned correlations between French and English growth indicate interdependence between the two linguistic systems. Such an investigation can be motivated and guided by existing theories of multilingual development that would suggest a common underlying proficiency (Cummins, 1980) or unitary linguistic system (e.g., Vogel & Garcia, 2017). Indeed, common underlying proficiency theory (Cummins, 1980) argues that knowledge (e.g., cognition, literacy skills) shared across two or more languages can be accessed and utilized for the acquisition of all of a speaker's languages and transfers to the acquisition of additional languages, thus supporting a learner's multilingual development overall. On the other hand, the notion of a unitary linguistic system central to translanguaging implies that language users do

not possess separate linguistic systems for each of the languages they know, rather they "select and deploy particular features from a unitary linguistic repertoire to make meaning and to negotiate particular communicative contexts" (Vogel & Garcia, 2017, p. 1).

Finally, we did not conduct formal tests to analyze disparities in growth patterns by different language features. Such comparisons can be conducted within the Bayesian framework, but we deliberately refrained from pursuing them in our current study. The rationale behind this decision is that, in order to detect such differences, we would likely require either a larger sample size, more informative prior information, or both.

In conclusion, this study helps fill a gap in research on DLI education by describing the progression of fine-grained features of language at the word, sentence, and discourse levels and the developmental trajectories of two languages, namely French and English, for early-elementary students. Theoretically, this research adds to the literature on CDST as it illustrates the dynamic and adaptive ways in which multiple features of two languages develop longitudinally in the context of bilingual schooling. In addition, this study offers practical implications for bilingual educators by providing them with a language progression approach that can help untangle deviance from delay in children's bilingual language development. In order to obtain a more informed picture of DLI outcomes, we encourage future research to apply such a language progression approach to a broader range of contexts, by examining additional languages, domains (e.g., writing), grades, or content areas (e.g., mathematics).

Replication package. Data and code have been made publicly available on OSF at the following link: https://osf.io/4ugk3/.

Competing interests. The authors have no conflicts of interests to disclose.

Notes

1 The degree of independence and the very notion of transfer between two systems is questioned by translanguaging scholars, who argue that multilingual individuals draw from a unitary repertoire based on the communicative needs of a situation (García et al., 2021).

2 Transitional Kindergarten is the first year of a two-year Kindergarten program for children who, at the time of the study, had birthdays shortly after the September 2nd Kindergarten enrollment cutoff.

3 For more information on model comparison results, see Appendix F on the OSF page.

4 Our use of ordinal regression is in line with language assessment studies (e.g., Gujord, 2023) that examine the extent to which (if at all) there is a meaningful association between ordinal outcomes and various independent variables.

5 For example, a positive point estimate of γ_{01} accompanied by a 95% CI with only positive values would indicate that being in first grade as opposed to Kindergarten is meaningfully associated with greater odds of receiving a higher rating. Similarly, a negative point estimate of γ_{01} accompanied by a 95% CI with only negative values would indicate that being in first grade as opposed to Kindergarten is meaningfully associated with a lower odds of receiving a higher rating.

6 Initially, far fewer posterior samples were drawn. As can be seen on the OSF page, the number of posterior draws was increased to 4500, a fairly high number, due to convergence issues. Also, to help with early convergence issues related to a high number of divergent samples, an adapt_delta of 0.99 was chosen. Divergences are regions of the posterior of high curvature that are difficult to explore, and even small numbers of divergences can invalidate results; therefore our use of an adapt_delta of 0.99 was critical.

7 As one reviewer pointed out, the deceleration of proficiency attainment is expected at higher levels, as the inverted ACTFL (n.d.) proficiency scale symbolizes.

References

ACTFL. (n.d.). *ACTFL Proficiency Guidelines*. https://www.actfl.org/educator-resources/actfl-proficiency-guidelines

Ali, S., Ali, A., Khan, S. A., & Hussain, S. (2016). Sufficient sample size and power in multilevel ordinal logistic regression models. *Computational and Mathematical Methods in Medicine*, **2016**. https://doi.org/10.1155/2016/7329158

Bailey, A. L. (2017). Progressions of a new language: Characterizing explanation development for assessment with young language learners. *Annual Review of Applied Linguistics*, **37**, 241–263. https://doi.org/10.1017/S0267190517000113

Bailey, A. L., Blackstock-Bernstein, A., & Heritage, M. (2015). At the intersection of mathematics and language: Examining mathematical strategies and explanations by grade and English learner status. *The Journal of Mathematical Behavior*, **40**, 6–28. https://doi.org/10.1016/j.jmathb.2015.03.007

Bailey, A. L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly*, **48**(3), 480–506. https://doi.org/10.1002/tesq.176

Bailey, A., & Heritage, M. (2019). *Progressing students' language day by day*. Corwin Press.

Bailey, A. L., Osipova, A., & Kelly, K. R. (2015). Language development. In L. Corno & E. M. Anderman (Eds.), *Handbook of Educational Psychology* (pp. 213–226). Routledge. https://doi.org/10.4324/9781315688244

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, **34**(1), 97–114. https://doi.org/10.3102/1076998607310504

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, **16**(4), 373. https://doi.org/10.1037/a0025813

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D. & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, **59**, 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Bendixen, T., & Purzycki, B. G. (2022). Cognitive and cultural models in psychological science: A tutorial on modeling free-list data as a dependent variable in Bayesian regression. https://doi.org/10.31234/osf.io/4n2jm

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. https://doi.org/10.48550/arXiv.1701.02434

Blackstock-Bernstein, A., Woodbridge, A., & Bailey, A. (2022). Comparing elementary students' explanatory language across oral and written modes. *The Elementary School Journal*, **122**(3), 315–340. https://doi.org/10.1086/718077

Boers, F. (2018). Picture prompts and some of their uses. *Language Teaching Research*, **22**(4), 375–378.

Boucher, J., Slegers, A., & Brambati, S. M. (2019). Cross-sectional analysis of picture descriptions of healthy young and older adults. *Neuropsychologie Clinique et Appliquée – Applied and Clinical Neuropsychology*, **3**, 132–145. https://doi.org/10.1080/13854046.2020.1817563

Burkhauser, S., Steele, J. L., Li, J., Slater, R. O., Bacon, M., & Miller, T. (2016). Partner-language learning trajectories in dual-language immersion: Evidence from an urban district. *Foreign Language Annals*, **49**(3), 415–433. https://doi.org/10.1111/flan.12218

Bürkner P (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, **80**(1), 1–28. https://doi.org/10.18637/jss.v080.i0

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chaturvedi, S. R. B. H., & Shweta, R. C. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, **41**(3), 20–27.

Cho, M., & Kim, Y. S. G. (2023). Do second graders adjust their language by discourse context? *Language, Speech, and Hearing Services in Schools*, 1–15.

Chondrogianni, V., Judge-Clayden, F., & Butcher, M. (2022). Majority language vocabulary and nonword repetition skills in children attending minority language immersion education. *Applied Psycholinguistics*, 1–35. https://doi.org/10.1017/S0142716422000285

Cummins, J. (1980). The construct of language proficiency in bilingual education. In *Current issues in bilingual education. Georgetown University Round Table on Language and Linguistics (GURT) 1980* (pp. 81–103). Georgetown University Press.

De Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, **10**(1), 7–21. https://doi.org/10.1017/S1366728906002732

de Carvalho, M. (2018). spearmanCI: Jackknife Euclidean / Empirical Likelihood Inference for Spearman Rho. R package version 1.0.

De Temple, J. M., & Beals, D. E. (1991). Family talk: Sources of support for the development of decontextualized language skills. *Journal of Research in Childhood Education*, **6**(1), 11–19. https://doi.org/10.1080/02568549109594818

Dickinson, D. K., McCabe, A., Anastasopoulos, L., Peisner-Feinberg, E. S., & Poe, M. D. (2003). The comprehensive language approach to early literacy: The interrelationships among vocabulary, phonological sensitivity, and print knowledge among preschool-aged children. *Journal of Educational Psychology*, **95**(3), 465–481. https://doi.org/10.1037/0022-0663.95.3.465

Dickinson, D. K., & Snow, C. E. (1987). Interrelationships among prereading and oral language skills in kindergartners from two social classes. *Early Childhood Research Quarterly*, **2**(1), 1–25. https://doi.org/10.1016/0885-2006(87)90010-X

Durán, L., Wackerle-Hollman, A., Miranda, A., Chávez, C., Pentimonti, J., Zyskind, K., & Rodriguez, M. C. (2022). Spanish and English oral language growth rates of bilingual preschoolers: The effect of language of instruction. *Learning Disabilities Research & Practice*, **37**(3), 175–188. https://doi.org/10.1111/ldrp.12287

Eisenberg, S. L., & Guo, L. Y. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools*, **44**(1), 20–31. https://doi.org/10.1044/0161-1461(2012/11-0089)

Fiestas, C. & Peña, E. (2004). Narrative discourse in bilingual children. *Language, Speech, and Hearing Services in Schools*, **35**(2), 155–168. https://doi.org/10.1044/0161-1461(2004/016)

Fortune, T. W., & Ju, Z. (2017). Assessing and exploring the oral proficiency of young Mandarin immersion learners. *Annual Review of Applied Linguistics*, **37**, 264–287. https://doi.org/10.1017/S0267190517000150

Fortune, T. W., & Tedick, D. J. (2015). Oral proficiency assessment of English-proficient K-8 Spanish immersion students. *The Modern Language Journal*, **99**(4), 637–655. https://doi.org/10.1111/modl.12275

García, O., Flores, N., Seltzer, K., Wei, L., Otheguy, R., & Rosa, J. (2021). Rejecting abyssal thinking in the language and education of racialized bilinguals: A manifesto. *Critical Inquiry in Language Studies*, **18**(3), 203–228.

Gathercole, V., & Thomas, E. (2009). Bilingual first-language development: Dominant language takeover, threatened minority language take-up. *Bilingualism: Language and Cognition*, **12**(2), 213–237. https://doi.org/10.1017/S1366728909004015

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Griffin, T. M., Hemphill, L., Camp, L., & Wolf, D. P. (2004). Oral discourse in the preschool years and later literacy skills. *First language*, **24**(2), 123–147. https://doi.org/10.1177/0142723704042369

Gujord, A. K. H. (2023). Who succeeds and who fails? Exploring the role of background variables in explaining the outcomes of L2 language tests. *Language Testing*, **40**(2), 227–248.

Hambly, H., Wren, Y., McLeod, S., & Roulstone, S. (2013). The influence of bilingualism on speech production: A systematic review. *International Journal of Language & Communication Disorders*, **48**, 1–24. https://doi.org/10.1111/j.1460-6984.2012.00178.x

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**(4), 933–944. https://doi.org/10.2307/2533433

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**(1), 1593–1623. https://doi.org/10.48550/arXiv.1111.4246

Hung, L. F., & Huang, C. M. (2011). Effect of the number of categories, number of time-points, and sample size on the recovery of random-effect ordinal regression model parameters. *Journal of Statistics and Management Systems*, **14**(6), 1175–1198. https://doi.org/10.1080/09720510.2011.10701607

Larsen-Freeman, D. (2011). A complexity theory approach to second language development/acquisition. In D. Atkinson (Ed.), *Alternative Approaches to Second Language Acquisition* (pp. 48–72). Routledge. https://doi.org/10.4324/9780203830932

Larsen-Freeman, D. (2020). Complex dynamic systems theory. In B. VonPatten, & J. Williams (Eds.), *Theories in second language acquisition* (pp. 248–270). Routledge. https://doi.org/10.4324/9780429503986

Larsen-Freeman, D., & Todeva, E. (2021). A sociocognitive theory for plurilingualism: Complex dynamic systems theory. In E. Piccardo, A. Germain-Rutherford, & G. Lawrence (Eds.), *The Routledge handbook of plurilingual language education* (pp. 209–224). Routledge. https://doi.org/10.4324/9781351002783

Levy, R., & McNeish, D. (2021). Perspectives on Bayesian inference and their implications for data analysis. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000443

Lindholm-Leary, K., & Hernández, A. (2011). Achievement and language proficiency of Latino students in dual language programmes: Native English speakers, fluent English/previous ELLs, and current ELLs. *Journal of Multilingual and Multicultural Development*, **32**(6), 531–545. https://doi.org/10.1080/01434632.2011.611596

Lyster, R., & Tedick, D. J. (2022). Oral language development in immersion and dual language classrooms. In W. Cardoso, D. T. Derwing, M. Munro, & R. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking* (pp. 328–343). Routledge.

Martínez, R. A., & Mejía, A. F. (2019). Looking closely and listening carefully: A sociocultural approach to understanding the complexity of Latina/o/x students' everyday language. *Theory into Practice*, **59**(1), 1–11. https://doi.org/10.1080/00405841.2019.1665414

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **42**(2), 109–127. https://www.jstor.org/stable/2984952

Nurss, J. R., & Hough, R. A. (1985). Young children's oral language: Effects of task. *The Journal of Educational Research*, **78**(5), 280–285. https://doi.org/10.1080/00220671.1985.10885616

Oller, D., Pearson, B., & Cobo-Lewis, A. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics*, **28**(2), 191–230. https://doi.org/10.1017/S0142716407070117

Paradis, J. (2023). Sources of individual differences in the dual language development of heritage bilinguals. *Journal of Child Language.* Advance online publication. https://doi.org/10.1017/S0305000922000708

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rolstad, K., Mahoney, K., & Glass, G. V. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy*, **19**(4), 572–594. https://doi.org/10.1177/0895904805278067

Ryan, È. (2023). Parents' investment in a French-English dual language immersion program in the United States. *Journal of Language, Identity & Education*, **22**(1), 51–65. https://doi.org/10.1080/15348458.2020.1832497

Ryan, È. (2021a). French vocabulary development of early-elementary students in a dual language immersion program: The role of out-of-school input and output. *International Journal of Bilingualism*, **25**(5), 1200–1213. https://doi.org/10.1177/13670069211000849

Ryan, È. (2021b). The impact of home literacy on bilingual vocabulary development. *Bilingual Research Journal*, **44**(1), 108–123. https://doi.org/10.1080/15235882.2021.1905105

Saito, K., & Hanzawa, K. (2018). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*, **22**(4), 398–417. https://doi.org/10.1177/1362168816679030

Smith, B. E., Pacheco, M. B., & Khorosheva, M. (2021). Emergent bilingual students and digital multimodal composition: A systematic review of research in secondary classrooms. *Reading Research Quarterly*, **56**(1), 33–52. https://doi.org/10.1002/rrq.298

Snow, C. E., Tabors, P. O., Nicholson, P. A., & Kurland, B. F. (1995). SHELL: Oral language and early literacy skills in kindergarten and first-grade children. *Journal of Research in Childhood education*, **10**(1), 37–48. https://doi.org/10.1080/02568549509594686

Speece, D. L., Roth, F. P., Cooper, D. H., & De La Paz, S. (1999). The relevance of oral language skills to early literacy: A multivariate analysis. *Applied Psycholinguistics*, **20**(2), 167–190.

Stan Development Team. (2022). Stan. https://mc-stan.org/

Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on student achievement: Evidence from lottery data. *American Educational Research Journal*, **54**(1), 282S–306S. https://doi.org/10.3102/0002831216634463

Vandenborre, D., Visch-Brink, E., van Dun, K., Verhoeven, J., & Mariën, P. (2018). Oral and written picture description in individuals with aphasia. *International Journal of Language & Communication Disorders*, **53**(2), 294–307. https://doi.org/10.1111/1460-6984.12348

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. *R package version*, **2**(0), 1003. https://mc-stan.org/loo/, https://discourse.mc-stan.org

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5), 1413–1432. https://doi.org/10.48550/arXiv.1507.04544

Vogel, S. & Garcia, O. (2017). Translanguaging. In G. Noblit & L. Moll (Eds.), *Oxford research encyclopedia of education* (pp. 1–21). Oxford University Press.

Watzinger-Tharp, J., Swenson, K., & Mayne, Z. (2018). Academic achievement of students in dual language immersion. *International Journal of Bilingual Education and Bilingualism*, **21**(8), 913–928. https://doi.org/10.1080/13670050.2016.1214675

Watzinger-Tharp, J., Tharp, D. S., & Rubio, F. (2021). Sustaining dual language immersion: Partner language outcomes in a statewide program. *The Modern Language Journal*, **105**(1), 194–217. https://doi.org/10.1111/modl.12694

Wu, H. F., De Temple, J. M., Herman, J. A., & Snow, C. E. (1994). "l'animal qui fait oink! oink!": Bilingual children's oral and written picture descriptions in English and French under varying instructions. *Discourse Processes*, **18**(2), 141–164. https://doi.org/10.1080/01638539409544889

Xu, X., Padilla, A. M., & Silva, D. M. (2015). Learner performance in Mandarin immersion and high school world language programs: A comparison. *Foreign Language Annals*, **48**(1), 26–38. https://doi.org/10.1111/flan.12123