Applied Linguistics, 2024, 45, 364–382
https://doi.org/10.1093/applin/amad031
Advance access publication 13 June 2023
Article

The Effects of CLIL and Sources of Individual Differences on Receptive and Productive EFL Skills at the Onset of Primary School

Adriana Soto-Corominas*, Helena Roquet, and Marta Segura

Department of Applied Linguistics, Universitat Internacional de Catalunya, Barcelona, Catalunya, Spain
*E-mail: asotoc@uic.cat

Research on the implementation of CLIL at the onset of primary school is limited and has largely overlooked the role of other sources of individual differences. This study investigated the effects of the CLIL approach to English learning, together with the effects of out-of-school exposure to the language through media and other sources of individual differences, in a sample of Grade 1 students in Catalonia (Spain) using a longitudinal design. Participants (N = 176) from 14 different schools completed a test battery at the beginning and end of Grade 1 that assessed receptive and productive English skills. Results revealed that abilities at the onset of Grade 1 were the best predictor of abilities at the end of the year, and that CLIL was not associated with additional advantages in the students that followed the approach. In addition, certain characteristics of the linguistic and family background of participants predicted additional gains during the academic year: participants who engaged in more English extracurricular activities and participants with more educated mothers performed better at the end of Grade 1.

Introduction

Content and Language Integrated Learning (CLIL) is a teaching approach that aims to foster both Foreign Language (FL) and content learning in an integrated way (Merino and Lasagabaster 2018). As a tool to promote multilingualism, CLIL has experienced a rapidly increasing implementation in schools within the latest decades (Dalton-Puffer 2008). Parallel to the growing expansion of CLIL in schools, research in this field has also gained ground.

A number of studies have analyzed which linguistic skills may be enhanced by CLIL. However, most of those studies have focused on secondary education students (e.g. Merino and Lasagabaster 2018), whose literacy skills are fully developed. Research with primary education students is still rather limited and has reported contradictory results. In addition, this body of research has often failed to consider the sources of individual differences that may affect children's FL development and that may thus act as a confounding factor when CLIL effects are analyzed.

Empirical research that considers CLIL effects together with sources of individual variation is needed to determine whether the economic/human investment that goes into the implementation of CLIL in earlier grades translates into comparable gains in FL development.

CLIL research in primary education in Spain

Since the implementation of CLIL may yield largely different results depending on the country where it is applied due to important contextual differences (Sylvén 2013), we focus our discussion of CLIL outcomes on research conducted in Spain, where the current study took place. The limited number of studies on CLIL at the primary level have yielded contradictory results. On the one hand, certain studies have found CLIL students to outperform those who follow EFL only. For example, Jiménez-Catalán *et al*. (2006) examined vocabulary profiles measured through students' reading and writing abilities, and reported higher results in CLIL students. One of the most ambitious studies to date, Pérez-Cañado's (2018), similarly found significant advantages for CLIL students in vocabulary, grammar, speaking, and listening abilities from a sample of six-graders from 53 schools.

Other studies, however, have found CLIL students performing on par with or worse than their non-CLIL counterparts. For speaking and listening skills, Serra (2007) showed no significant differences between CLIL and non-CLIL students in their growth between Grades 1–6, in line with Pladevall-Ballester and Vallbona (2016), who found that non-CLIL students actually outperformed their CLIL counterparts in listening comprehension skills in their two-year longitudinal study. Agustín-Llach (2015) and Agustín-Llach and Canga Alonso (2014), who investigated lexical development between Grades 4–6, found that students who had followed a CLIL approach since Grade 1 performed similarly to their non-CLIL counterparts and showed similar growth trends over time.

Finally, some studies have found CLIL advantages in some domains but not in others in the same group of students. In Nieto's (2016) study using census data from schools in the province of Castilla-La Mancha, Grade 4 learners' oral and written production and comprehension were examined. Students who had followed CLIL since Grade 1 outperformed their non-CLIL peers in oral production only, but not in the other abilities. In line with Nieto (2016), Gayete (2022) compared CLIL and non-CLIL learners from the same school in the Valencian Community. Significantly better results were reported in Grade 2 CLIL students in oral production only, while in listening comprehension it was the non-CLIL group that outperformed the CLIL group.

In summary, the limited research on receptive and productive abilities in primary school shows conflicting results. However, extant research has two main caveats. First, most studies included samples from only one or two schools. This poses an important limitation, since these results are only relevant for the specific context where such studies were conducted. This makes results difficult to generalize, as characteristics intrinsic to the school, their teachers or their students could influence results. Secondly, none of these studies have directly addressed the effects of sources of individual differences on the development of the FL. Factors such as frequency of participation in extracurricular FL activities, FL input richness in the home or socioeconomic status, among others, have sometimes been used only as a measure to ensure homogeneity between groups (e.g. Pérez-Cañado 2018), while their association with primary students' FL development in the CLIL context has yet to be studied.

Explaining conflicting results in CLIL

Contradictory results in previous studies may be explained by the following two hypotheses by Muñoz (2015): first, a minimum number of hours may be required to reap the advantages of CLIL. This is shown in studies that compare children with the same age but different number of hours of English exposure (e.g. Xanthou 2011; Housen 2012). Secondly, CLIL approaches may be more advantageous in older children, as shown in studies that compare children with the same number of hours of CLIL instruction at different ages (Lorenzo *et al*. 2010; Bret 2011; Canga Alonso 2015).

The studies related to the first hypothesis lead to the conclusion that increased exposure to the English language through CLIL leads to proficiency advantages, but the amount of exposure that is necessary remains unclear (Muñoz 2015). However, results point towards advantages for CLIL students not being apparent from the early stages of CLIL implementation.

Regarding the second hypothesis, studies suggest that the acquisition rate of older students in CLIL is faster than that of younger students. The nature of this older age advantage could be based on maturational effects; older learners may be better able to benefit from the cognitive-academic skills developed in their L1(s) and use them in the CLIL subject to their advantage. Alternatively, proficiency thresholds may be at the root of the older age advantage. Hypothetically, a higher proficiency in the CLIL language at the onset of instruction may facilitate FL gains. Thus, benefits of CLIL instruction may emerge faster in older learners, whose starting level is typically higher than that of younger learners. However, studies that have investigated proficiency thresholds in CLIL implementation, albeit in university, do not lend support to this theory (Aguilar and Muñoz 2014).

Outside-of-school English exposure

While the implementation of CLIL may play a prominent role in how a FL is learned at school, children have vastly different experiences engaging with the target language outside of school that could impact how the language is learned (Peters 2018; De Wilde *et al*. 2022). However, because research on the effects of CLIL rarely considers such experiences, it is unclear the degree to which conflicting results could be explained through these experiential factors.

Most of the research on individual differences on second language (L2) development has been done on L2-community learners (i.e. children who acquire the community language as an L2; e.g. Paradis 2019), but growing research shows that variations in the FL input may lead to differential rates of development in learners in primary and secondary school.

Frequency of FL reading has been shown to be associated with abilities in the FL in primary- and secondary-school-age learners, including productive and receptive vocabulary (Peters 2018; De Wilde *et al*. 2020, 2022) and oral proficiency (Sundqvist 2009). However, few studies have investigated the association between FL reading and FL skills due to the low frequency with which young children engage in reading (e.g. Lindgren and Muñoz 2013).

Technology, such as watching TV/videos online in the FL, also offers the potential for children to engage with the FL from home. Studies that have considered this type of input have found positive associations between engagement with these activities and FL outcomes, such as in listening comprehension (Lindgren and Muñoz 2013) and vocabulary (Sundqvist 2009; Kuppens 2010; Peters 2018) in primary and secondary school. Especially relevant for our study are the results from Muñoz *et al.* (2018), who tested the receptive abilities in vocabulary and grammar in L2-English by Spanish/Catalan and Danish children at age 7 and 9 and found that exposure to movies in English only predicted performance in the older group.

As opposed to TV watching, playing videogames offers the possibility for learners to actively engage in interaction with fluent or native speakers of the L2 (Ryu 2013), which could lead to gains in the development of the FL (Mackey and Goo 2007). For example, De Wilde *et al*. (2020) found that the frequency of videogame playing in English was positively associated with several measures of L2-English, including vocabulary, in Dutch children aged 10–12. Similar results were reported for Danish and Swedish children learning L2-English in primary school in Hannibal Jensen (2017) and Sylvén and Sundqvist (2012), respectively. Lindgren and Muñoz (2013), who collected information on a variety of exposure factors, found that while playing videogames more frequently in the FL was associated with better outcomes in listening comprehension, other predictors (such as TV viewing) bore a stronger association.

Finally, an additional source of out-of-school FL exposure is extracurricular activities in English. These extracurriculars could be English language classes or other types of classes (e.g. crafts, sports, theater) conducted strictly in English. A survey conducted in 2021 found that 41.4 per cent of primary school students in Spain attended extracurricular FL classes, making them an additional, and frequent, source of FL input (Franco Hidalgo-Chacón *et al*. 2022). Most of these

classes have reduced class groups that allow them to be more interactive than English classes at school.

Importantly, not all studies find a positive association between out-of-school exposure to the FL and skills in that language. Unsworth *et al.*'s (2015) study on the L2-English development of Dutch children ages 4–6 found no such relation. In fact, studies investigating the contribution of out-of-school input on FL development have found that secondary school students may benefit from it more greatly than primary school students (Van Mensel and Galand 2022).

Other potential sources of variation

FL development has been shown to be influenced by factors indirectly related to linguistic experience. For example, maternal education, used often as a proxy for family socioeconomic status, bears an association with the quantity and quality of linguistic input children receive (Hoff 2006). Children with more educated mothers tend to have better linguistic outcomes, regardless of whether a language is used to communicate between the mother and child (Paradis 2019). Indeed, maternal education has been shown to have a positive association with FL vocabulary outcomes in primary and secondary students (Van Mensel and Galand 2022), though this has not been a consistent finding (Lingdren and Muñoz 2013).

The role of gender has been investigated in terms of how it may modulate engagement with the FL outside of school, on the one hand, and FL development more broadly, on the other. Regarding the first line of research, some studies have noted gender differences in how learners engage with English materials outside of school, with male students generally engaging with more videogame playing than females (Sundqvist and Sylvén 2014; Sundqvist and Wikström 2015) and female students watching more TV/movies than males (Muñoz 2020). However, once other differences are accounted for, most studies have not found an advantage for either gender (e.g. De Wilde and Eyckmans 2017).

Age of onset of acquisition (AOA) of the FL has played a pivotal role in the field of L2 acquisition in the debate on maturational constraints and ultimate attainment. However, research on AOA effects in community-L2 learners has been unjustly generalized to the setting of FL instruction (Muñoz 2011). While the former body of research has found that a younger AOA may be advantageous in the long run, studies on the development of FL skills have generally failed to find similar results (e.g. Muñoz 2011). In the present study, we control for English AOA given that the participants are at the very onset of formal instruction (Grade 1). As such, fluctuations in AOA could be expected to play a stronger role than in studies that have tested samples of older participants.

Present study

With the increasing popularity of CLIL, schools are implementing CLIL approaches earlier on to boost FL skills. However, such decisions are not always grounded on empirical research, which is limited in primary schools in general and practically non-existent at the onset of primary school. In addition, the existent body of research has yielded findings with conflicting results. Importantly, extant research has not considered sources of individual variation, and has often failed to include diverse samples of CLIL and non-CLIL participants. As such, we address these gaps by testing a sample of students coming from different primary schools in the region of Catalonia (Spain). We asked the following two questions:

(1) Does following a CLIL approach predict gains in English receptive/productive abilities between the beginning and end of Grade 1 once other sources of variation are accounted for?

We hypothesized that null results were likely given Muñoz's (2015) double hypothesis that a minimum number of hours may be necessary for CLIL to show significant benefits and that older children may be more likely to benefit from CLIL than younger children.

(2) What are the best predictors of gains over this period of time?

The lack of studies investigating individual variation at the onset of primary schooling forced us to extrapolate from results with samples of older children. Given that previous studies had found that younger children were less likely to engage in English-rich activities (reading, TV/video watching, videogame playing, and formal extracurricular activities), we predicted that individual variation in these activities may not be enough to show associations with their English skills. Regarding maternal education, we expected it to play a role by being positively associated with English skills. Finally, even though English AOA and gender were controlled for, we did not expect either of these variables to play a significant role.

Background context: Catalonia

Catalonia is an autonomous province in northeast Spain where Spanish and Catalan have official status, and where bilingualism is historical and widespread. Obligatory schooling in Catalonia starts at age 6 (Grade 1), though over 94 per cent of children are schooled at age 3 (IDESCAT 2020).

The schooling system implemented in Catalonia is often referred to as *Catalan immersion*, because Catalan is the primary language of instruction in public and charter schools. However, by the end of obligatory schooling (Grade 10), students must be able to use both Spanish and Catalan fluently in both oral and written communication. In addition, English is part of the curriculum (Generalitat de Catalunya 2018), and by the end of obligatory schooling students must have attained a B1 level of English. Instruction of English may begin as early as kindergarten (prior to age 6) in many schools, or may be delayed until the onset of primary school (Generalitat de Catalunya 2018). As such, the goal of the educational model implemented in Catalonia is to foster trilingual abilities in Catalan, Spanish, and English. Spain being one of the European countries with the lowest skills in English (EF 2022: 18), many Spanish and Catalan schools have embraced CLIL approaches to boost students' skills in the language (Codó 2022).

Whereas the development of Catalan and Spanish for the majority of students occurs in naturalistic contexts, for most, the development of English happens at school. The opportunities for English exposure outside of formal contexts are limited, as English is not present in the community and TV/movies are often dubbed into or are produced in Spanish or Catalan. Extracurricular activities provide opportunities for more hours of English exposure and thus are often chosen by parents who want to enhance their children's early language learning. In addition, parents may also expose children to English in the home, even if their own skills are limited, by providing access to media in English (Alexiou 2015).

Methods

Study design

This study presents the data of the first two times of English data collection of an ongoing longitudinal study that assesses the linguistic abilities in English, Catalan, and Spanish of the same sample of participants. The first data collection (Time 1) took place in October/November 2021, at the onset of primary schooling (Grade 1), with Time 2 taking place in May/June 2022.

Participants

At Time 1, 190 participants (97 males, 93 females) took part in the study from 14 schools within the province of Barcelona. Of this initial sample, we do not consider the data from eight participants whose parents reported speaking English in the home. Furthermore, we do not consider the data of three participants with a diagnosis of autism spectrum disorder and of two additional participants who had had a diagnosis of a language delay earlier in life. We also do not consider the data from a participant whose home language was Spanish but lived in Switzerland until the age of 6, since he may have had some community exposure to English.

Our final sample thus comprises 176 participants (89 males, 87 females). All participants attended Grade 1 and had an average age of 6;4 (SD = 0;4) at Time 1. Of the 176 participants, nine

had been born outside Catalonia. A total of 14 participants spoke a language in the home in addition to or instead of Catalan/Spanish. Of these, two spoke German, eight spoke a Romance language (e.g. French, Galician), and the rest spoke a language that was not Romance or Germanic (Arabic, Chinese, Punjabi, and Russian).

A total of 16 participants of the 176-participant sample were not tested in English at Time 1. Most of these participants were not tested due to their absence the day they were scheduled to be tested ($N = 14$). The other two declined to participate. At Time 2, one of the 14 schools was not available for testing, which reduced the sample to 142 participants (69 males, 73 females).

Schools

The 14 participating schools were part of 75 randomly selected schools in the province of Barcelona (Catalonia). Many schools were contacted as it was anticipated that most would not be willing to participate in the study, given that the academic year of 2021–22 was the first year following the COVID-19 pandemic.

Of the 14 schools, 7 were public and 7 chartered. Four public schools implemented a CLIL approach and 3 did not, whereas 3 of the 7 chartered schools implemented a CLIL approach, as opposed to 4 that did not. Importantly, when a school implemented a CLIL approach in Grade 1, *all* students (and hence all participants from that school) followed the same approach. All CLIL schools used only English for the CLIL subjects.

There were differences in terms of the hours of English instruction between the schools that implemented CLIL and those that did not, as shown in Figure 1. While CLIL schools had more English instruction than non-CLIL schools overall ($M_{CLIL} = 5.36$, $SD_{CLIL} = 4.09$; $M_{Non\text{-}CLIL} = 3.36$, $SD_{Non\text{-}CLIL} = 1.18$), there was overlap between the two types of schools. This is properly accounted for in the statistical modeling.

Within the CLIL schools there was variation in terms of the CLIL subjects. Arts was taught in English in three schools. Music, Science, and Physical education were each taught in English in

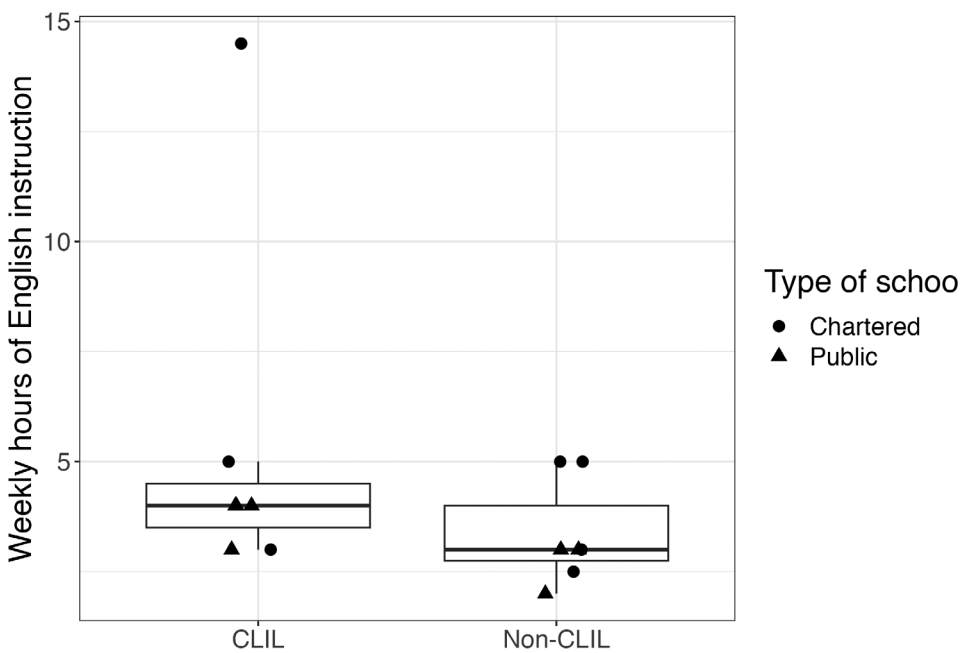

Figure 1: Boxplot showing the number of hours of English instruction per week of the 14 schools according to whether they were public or chartered. Lines in the middle of the boxes indicate medians, not means.

two schools, and Robotics, Drama, Dance, and Computer science were taught in English in one school (note that some schools offered more than one CLIL subject). Similarly, there was variation across schools in the number of CLIL hours: four schools offered 1 h of CLIL per week, two offered 2 h, and one offered 11.5 h. We chose not to eliminate the last school from our sample for three reasons: first, this type of school exists in Catalonia and happened to be sampled, therefore constituting a legitimate part of the studied population. Secondly, our statistical analysis controlled for any variability arising from individual schools (see Data Analysis section). As such, data from this school could not be argued to strongly bias results. Finally, we verified the previous claim by rerunning the analysis excluding data from this school and the interpretation of the results did not meaningfully change.

Instruments and reliability

Parents or primary caregivers were sent a questionnaire. In order to test participants' receptive and productive abilities in English we administered three tests: the *Peabody Picture Vocabulary Test*–5th edition (PPVT; Dunn 2019), the *Test for Reception of Grammar 2* (TROG; Bishop 2003), and the *Multilingual Assessment Instrument for Narratives* (MAIN; Gagarina *et al*. 2019).

Background questionnaire.

Parents who agreed to participate (see Procedures section) were asked to complete a background questionnaire to collect information on the participants' demographic and linguistic background. They were given the option of completing the questionnaire online, over the phone, or in person.

Crucially, the questionnaire prompted parents to indicate the average number of hours per week that the participants engaged in reading activities in English (including time dedicated to English homework and time of joint reading with their caregivers), extracurricular English language classes, extracurricular classes (e.g. arts and crafts or soccer) in English, and TV viewing or video game playing in English. The hours of the two types of extracurricular classes were combined into the variable *Weekly extracurriculars* given the similarity of the two constructs and the overall low frequency of both.

PPVT.

This test measures receptive lexical abilities. Participants are shown an array of four pictures and are asked to select the picture that matches the word spoken by the experimenter. This test has 240 items. When administered, this test was discontinued when participants made six errors in a group of eight items. Raw (i.e. non-standardized) scores are employed in this study, since the test was not normed on the population in which it was used. Research assistants scored each item during test administration. The Cronbach's alpha coefficient of internal consistency was 0.98 at Time 1 and 0.97 at Time 2.

TROG.

This test measures receptive grammatical abilities. Participants are shown an array of four pictures and are asked to select the picture that matches the statement given by the experimenter. Though the original test has 80 items, with 4 items evaluating 20 grammatical structures (e.g. negative statements, relative clauses), the piloting of this test showed it was too long to be part of the test battery in its full version. As such, it was reduced to 40 items (2 items for each of the 20 grammatical structures). It should be noted that, while the purpose of the original (i.e. full) test is to pinpoint constructions that represent areas of difficulty for the participant (Bishop 2003), the test is employed here as an overall measure of grammatical ability, and we refrain from discussing results regarding individual structures. Our administration of the TROG was discontinued when participants made six errors in a group of eight items. The Cronbach's alpha for this test was 0.94 at Time 1 and 0.91 at Time 2.

This test employs a restricted set of high-frequency vocabulary of nouns, verbs, and adjectives (Bishop 2003). Nevertheless, it does rely, to a certain extent, on vocabulary abilities. For

this reason, participants were first asked to take a preliminary test to determine whether they knew the content words in the task. This test was not scored, and its mechanics were that of the vocabulary test above, with the only difference being that participants had arrays of 8 pictures to choose from. No participant that knew less than half of the words took the test. However, this only affected one participant in the entire sample. Words that were not known by the participant were taught by the experimenter and retested. If necessary, words were then taught again prior to administering the test.

MAIN.

We employed the *Dog* story of the MAIN to measure listening comprehension and productive skills. This story has six full-color pictures and is presented in a printed format.

The MAIN was administered as a story retell. That is, research assistants first told the participant the story by following a fold-out presentation mode, reading the story provided by the MAIN instructions (Gagarina *et al*. 2019), and then prompted participants to retell the story. Participants' output was audio recorded for posterior transcription and analysis.

Since English was a FL for all participants and it was anticipated that participants' skills would be highly limited at Time 1, in addition to the standard protocols for the administration of the MAIN, one additional consideration was followed during data collection. If participants started their story retell entirely in English or with some code switching between English and Catalan/Spanish, the experimenter would not interrupt. If participants produced more than two utterances entirely in Catalan/Spanish, the experimenter interrupted by saying 'in English?' If participants did not understand the question, the experimenter asked 'will you explain it in English?' in Catalan/Spanish. Participants were not interrupted again if they kept narrating the story in a non-target language.

The measure of productive abilities we employ for this study is word types (i.e. the number of different words produced during the retell), which measures participants' productive vocabulary. Other measures were more affected by the high percentage of code switching and repetition in participants' production.

After the story retell, experimenters administered the 10 open-ended comprehension questions of the MAIN (Gagarina *et al*. 2019), which we use as a measure of listening comprehension. Questions were never translated for participants, but correct responses provided in Catalan/Spanish were considered correct.

All stories for Times 1 and 2 were transcribed by the same trilingual transcriber. Twenty-five per cent of the stories at Time 1 and 29 per cent at Time 2 were transcribed from scratch by a second transcriber. Word-for-word percentage agreement at Time 1 was 97.2 per cent and at Time 2 it was 96.9 per cent. The same transcriber who transcribed all the audios also scored the comprehension questions. A second rater scored the comprehension questions for 25 per cent of the participants at Time 1, with an agreement rate of 90.3 per cent. At Time 2, the agreement rate was 93.9 per cent. The Cronbach's alpha for the comprehension test at Time 1 was 0.88 and 0.77 at Time 2.

Procedures

Ethical considerations.

The protocols for this study were revised and approved by the ethics board at the Universitat Internacional de Catalunya. Participating schools shared the invitation to participate in the study with students' parents in one of two formats depending on the typical mode of communication between the schools and families: either online (via email or through the school's own online platform) or on paper.

Data collection.

Data were collected at school during the school day. Participants were removed from class and tested individually in a quiet space. In total, 12 research assistants collected the data. Two of the

research assistants were native speakers of English, one had a B2 level of English, and the rest had a C1 or C2 level of English.

The three English tasks presented here are part of a larger battery that further included two literacy tests. The order of the five tests was randomized across participants. All tests were administered in the same session, which lasted a maximum of 50 min and an average of 30 min, including breaks.

Data analysis.

All descriptive and inferential tests were run in R (R Core Team 2022). We addressed both research questions with the same analyses. For each of the four outcome variables (vocabulary, grammar, listening comprehension, and word types in narratives), we ran the descriptive statistics with the relevant paired-samples and independent-samples Wilcoxon tests. When the Wilcoxon tests were significant, we obtained the Cohen's *d* effect size using the package *lsr* (Navarro 2015).

Subsequently, we fit a Generalized Linear Mixed-Effects (GLMER) model with a Poisson distribution using the *lme4* package (Bates *et al*. 2015), where the outcome variable was the score at Time 2. The predictors were: the total number of hours of English instruction participants had taken at school between Times 1 and 2 (*Hours of School English*), the number of weekly hours of extracurricular English activities (*Weekly extracurriculars*), of reading English activities in the home (*Weekly reading*), and of English TV viewing or videogame playing at home (*Weekly TV/videogames*), the years of maternal education (*Maternal education*), whether the school participants attended implemented CLIL or not (*CLIL*), their gender (*Gender*), their AOA in English (*English AOA*), and, crucially, participants' score in the same test at Time 1 (*Time 1 score*). The Time 1 score predictor served as an autoregressor, accounting for all the variability at Time 2 that could be explained by Time 1 abilities. All predictors that were numerical (i.e. all but *CLIL* and *Gender*), were scaled and centered. A random intercept was added for *School* to control for the variability explained by the fact that participants attended different schools.

Backward selection was followed for the predictors. Predictors that did not contribute significantly to the model were eliminated, one at a time, to reach the most parsimonious model. Reduced models were compared to their fuller counterparts by means of likelihood ratio tests. Since the effect of *CLIL* was central in answering research question 1, we did not eliminate this factor, even when non-significant.

All models were inspected for overdispersion and multicollinearity, and diagnostics of the residuals were run with the *DHARMa* package (Hartig 2020). When necessary, adjustments were made to the model and are explained in the Results section.

Results

Participant characteristics

Given the central role of the CLIL variable in this study, we present participant characteristics in Table 1 according to whether they attended a school that implemented CLIL or not. Participants in CLIL and Non-CLIL schools were similar in all the dimensions of interest except for the weekly number of hours of TV viewing and videogame playing in English, since non-CLIL students engaged in more than double the hours on average (which was due to some extreme values in this group).

Vocabulary

The results for the vocabulary test at Times 1 and 2 appear in Figure 2 for those participants who took the test both times, shown separately for participants attending CLIL and non-CLIL schools.

Table 2 shows the descriptive statistics for the entire sample (i.e. even those students who did not take the test one of the two times). For these tables (see also Tables 3–5), we employ the median and interquartile range (henceforth, IQR) as measures of central tendency instead of the mean and standard deviation since many of the test results were not normally distributed. The

Table 1: Participant characteristics, divided according to whether their school follows a CLIL approach or not

	CLIL (*N* = 99)		Non-CLIL (*N* = 77)	
	M	**SD**	**M**	**SD**
Age (months)	75.79	3.30	76.30	3.46
Age of English onset (months)	29.40	14.70	28.90	18.70
Hours of School English	108.20	64.65	85.88	31.12
Weekly extracurriculars	0.65	0.69	0.69	0.87
Weekly reading	0.77	1.50	0.84	1.73
Weekly TV/video games	1.00	1.52	2.44	5.52
Maternal education (years)	15.47	2.66	15.81	2.64

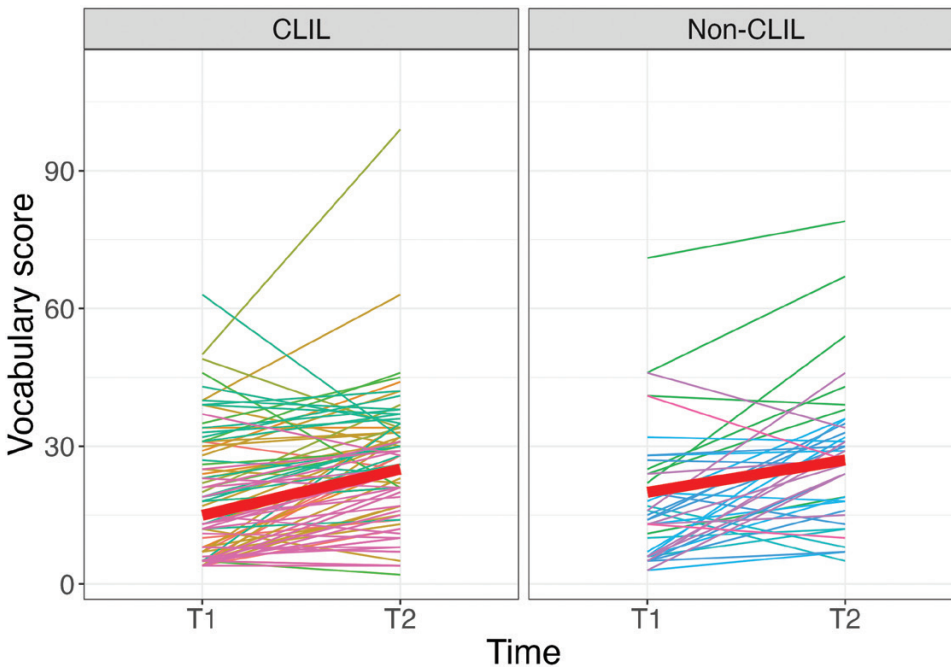

Figure 2: Vocabulary test results for participants who completed the test at the two time points. The x-axis represents the testing time (Times 1 and 2), and the y-axis represents the vocabulary score (range 0–240). Each line is one participant, with each color indicating the school of the participant. Scores are faceted according to whether the school followed or not a CLIL approach. The thick red line indicates the trend for the CLIL and Non-CLIL groups separately using the group median.

median and IQR are less susceptible to extreme outliers and asymmetrical distributions. Table 2 also includes the results of Wilcoxon tests. Specifically, two paired-samples Wilcoxon tests compared the performance of the CLIL and Non-CLIL participants, separately, at Times 1 and 2. As shown in Table 2, both tests were significant, demonstrating that both groups made significant vocabulary gains over time. Considering the Cohen's *d* effect size was medium in both groups, the extent of the gains was similar in both groups. In addition, Table 2 also presents independent-samples Wilcoxon tests comparing the performance of CLIL and Non-CLIL participants at the two

Table 2: Descriptive statistics for vocabulary test, together with Wilcoxon tests and, when relevant, Cohen's d effect size

	Time 1		Time 2		Paired-samples Wilcoxon test
	Median	IQR	Median	IQR	
CLIL	15	22	25	16	$p < .001$; $d = 0.622$ (medium)
Non-CLIL	20	21.5	27	18	$p < .001$; $d = 0.742$ (medium)
Independent-samples Wilcoxon test	$p = .031$; $d = 0.409$ (medium)		$p = .602$		

time points. At Time 1, this test was significant ($p = .031$), indicating that at Time 1, participants in Non-CLIL schools outperformed their CLIL counterparts. At Time 2, however, there was no evidence of such a difference.

Next, we discuss the statistical modeling to address our research questions. Since the initial Poisson GLMER model was found to be overdispersed, a negative binomial model was fit (Winter 2019: 227). The output of this model appears in Supplementary Appendix A. In terms of the effects of CLIL, the model found that this factor did not contribute significantly to the model ($p = .105$). However, other predictors were found to be associated with Time 2 vocabulary scores. As could be expected, Time 1 scores were strongly and positively associated with Time 2 scores ($p < .001$), suggesting that vocabulary abilities at Time 1 were strongly predictive of Time 2 abilities. In addition, the amount of English hours at school between Time 1 and 2 were also predictive of Time 2 vocabulary scores ($p = .008$). That is, participants who had engaged in more hours of English instruction at school between Times 1 and 2 performed better at Time 2. Two other predictors showed a positive association with vocabulary scores at Time 2: *Weekly extracurricular hours* ($p = .027$) and *Maternal education* ($p = .015$). This suggests that participants who took more hours of English classes outside of school and those with more educated mothers had higher vocabulary scores at Time 2.

Grammar

The results for grammar scores at Times 1 and 2 appear in Figure 3. As for the group results for the receptive grammar test (Table 3), participants in both CLIL and Non-CLIL schools made significant improvements between the two times. Differences between the two groups were not statistically significant at Time 1 or 2, though they trended towards significance for Time 1 ($p = .088$), in favor of Non-CLIL participants.

The initial Poisson GLMER showed overdispersion. As such, we fit a negative binomial GLMER. Similarly to the model for vocabulary, *CLIL* did not contribute to the model significantly ($p = .617$). However, Time 1 grammar scores ($p < .001$) were strongly associated with Time 2 abilities. In addition, there were two predictors that trended towards significance and were left in the model since a model without either of them was a marginally worse fit to the data. These two predictors were *Maternal education* ($p = .074$) and *Weekly extracurriculars* ($p = .071$), and they both were positively associated with Time 2 grammatical abilities. The output of this model appears in Supplementary Appendix B.

Listening comprehension

Results for the listening comprehension test, which could range between 0 and 10, appear in Figure 4 for CLIL and Non-CLIL participants. Group results are shown in Table 4. Participants in both CLIL and Non-CLIL schools made significant improvements between Times 1 and 2, and differences between the two groups were not statistically significant at either time.

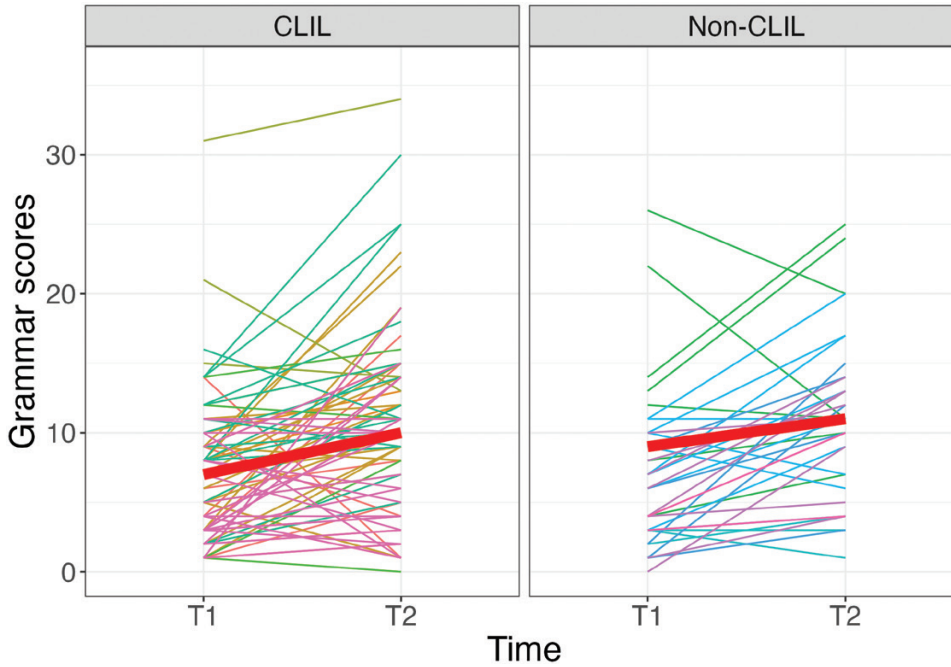

Figure 3: Grammar test results for participants who completed the test at the two time points. The x-axis represents the testing time (Times 1 and 2), and the y-axis represents the grammar score (range 0–40).

Table 3: Descriptive statistics for grammar test, together with Wilcoxon tests and, when relevant, Cohen's d effect size

	Time 1		Time 2		Paired-samples Wilcoxon test
	Median	IQR	Median	IQR	
CLIL	7	7	10	9.5	$p < .001$; $d = 0.554$ (medium)
Non-CLIL	9	10	11	6.75	$p < .001$; $d = 0.726$ (medium)
Independent-samples Wilcoxon test	$p = .088$		$p = .255$		

As shown in Figure 4, many participants scored 0 for narrative comprehension at Time 1. The negative binomial GLMER model, suitable for overdispersed data, found that the *CLIL* factor did not contribute to the model significantly ($p = .226$). Instead, Time 1 listening comprehension scores ($p < .001$) were the best predictor of Time 2 performance. One more predictor made a contribution to the model that was marginally significant: *Maternal education* ($p = .074$). The full output of this model appears in Supplementary Appendix C.

Word types in narrative production

The last outcome variable of interest was the number of word types participants used in the story retell of the *Dog* story of the MAIN. These results appear visualized in Figure 5. In terms of the group scores (Table 5), participants in both CLIL and Non-CLIL groups made significant improvements between Times 1 and 2. At Time 1, Non-CLIL participants produced significantly more types than CLIL participants, but this was not true at Time 2.

The initial Poisson GLMER model for the number of types in the narration was overdispersed and had singularity issues (i.e. the random intercept for *School* predicted no variance). As such,

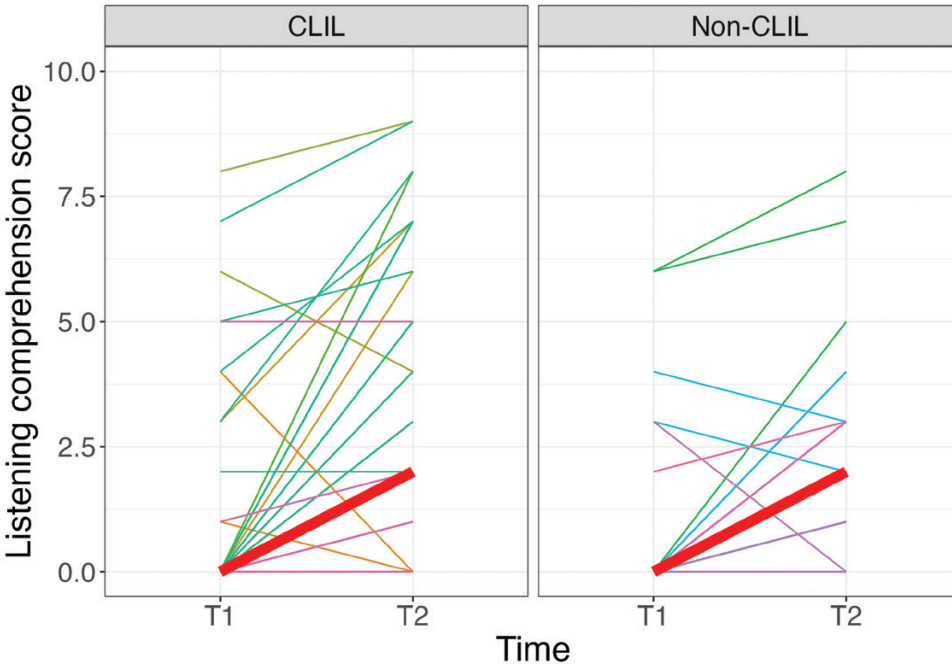

Figure 4: Listening comprehension test results for participants who completed the test at the two time points. The x-axis represents the testing time (Times 1 and 2), and the y-axis represents the listening comprehension score (range 0–10).

Table 4: Descriptive statistics for narrative comprehension test, together with Wilcoxon tests and, when relevant, Cohen's d effect size

	Time 1		Time 2		Paired-samples Wilcoxon test
	Median	IQR	Median	IQR	
CLIL	0	0	2	4	$p < .001$; $d = 0.739$ (medium)
Non-CLIL	0	2	2	3	$p < .001$; $d = 0.795$ (medium)
Independent-samples Wilcoxon test	$p = .290$		$p = .641$		

we fit a negative binomial GLM model without a random intercept. The full output of the optimal model appears in Supplementary Appendix D. As we found for the other three outcome variables, *CLIL* was not a significant predictor of Time 2 performance ($p = .842$). The only two predictors that were found to contribute significantly, and positively, to the model were *Word types at Time 1* ($p < .001$) and *Maternal education* ($p = .011$).

Discussion

This study is one of the first to consider the effects of CLIL implementation together with other potential sources of individual variation on the development of FL English receptive and productive abilities. We asked two main questions: first, whether following a CLIL approach at school was beneficial to the development of English abilities during Grade 1. Secondly, whether characteristics of the family and linguistic background that have been shown to influence FL skills in older children would similarly affect individual variation in this young sample of children.

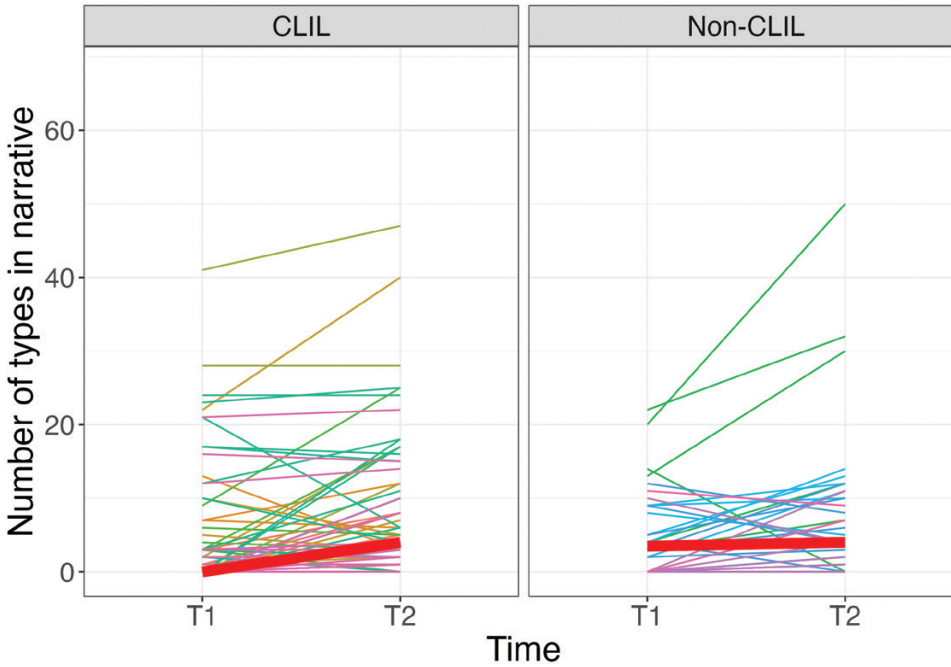

Figure 5: Number of types in narrative production for participants who completed the test at the two time points. The x-axis represents the testing time (Times 1 and 2), and the y-axis represents the number of types.

Table 5: Descriptive statistics for word types in narrative production, together with Wilcoxon tests and, when relevant, Cohen's d effect size

	Time 1		Time 2		Paired-samples Wilcoxon test
	Median	IQR	Median	IQR	
CLIL	0	4.25	4	9	$p < .001$; $d = 0.494$ (medium)
Non-CLIL	3.5	10.75	4	10	$p = .009$; $d = 0.420$ (medium)
Independent-samples Wilcoxon test	$p = .014$; $d = 0.394$ (small)		$p = .660$		

CLIL effects at the onset of primary schooling

When results of CLIL and Non-CLIL students were compared, separately for Times 1 and 2, it was found that Non-CLIL students were significantly better than CLIL students in vocabulary and word types in story retells, and marginally better in grammar at Time 1. However, at Time 2, none of the comparisons yielded significant results, suggesting that initial differences between CLIL and Non-CLIL students at the onset of Grade 1 had been neutralized by the end of that same year.

Modeling Time 2 results by including Time 1 scores as an autoregressor, together with other sources of individual differences, did not find evidence that following a CLIL approach yielded any particular advantages once other variables were accounted for. As such, these results could be interpreted cautiously as not providing support for the early implementation of CLIL at the onset of primary schooling.

As demonstrated by our review of the literature, this study is far from being the first one to not find advantages for a CLIL approach in primary school (Serra 2007; Agustín-Llach 2015;

Pladevall-Ballester and Vallbona 2016), though it is one of the first studies that have tested the results of this approach at the very onset of primary.

Even though CLIL was not found to be a significant predictor in and of itself, the model for receptive vocabulary found that the hours of English instruction at school were positively associated with Time 2 abilities. That is, participants who had spent more time in English classes at school (CLIL or not), were more likely to obtain higher scores at Time 2. It was the case that CLIL students overall spent more time learning English at school than Non-CLIL students (see Figure 1). Thus, effects of quantity of exposure to the FL were apparent at least for receptive vocabulary.

The question remains, however, why CLIL did not confer an advantage to students who followed such an approach at Grade 1. To explain this lack of effect, we invoke Muñoz's (2015) double hypothesis. First, it is possible that the CLIL participants had not received a sufficient amount of 'extra' input than the Non-CLIL participants. Such limited extra input may not be sufficient for the CLIL approach to yield advantages in Grade 1. Very young learners, such as those in early primary, benefit from implicit learning in naturalistic and immersion FL learning contexts (DeKeyser 2000; Paradis *et al*. 2021). Thus, the application of CLIL may need to go hand in hand with massive/increased FL exposure for young learners to benefit from it. In addition, it is possible that children at Grade 1 may be simply too young to benefit from a CLIL approach. Previous research comparing the implementation of CLIL at different ages has shown that older students may benefit more from this approach than younger ones (Muñoz 2015). Whether the older age advantage is rooted in older learners' increased cognitive/academic maturity or in their higher proficiency level at the onset of CLIL experiences is, however, difficult to disentangle.

Best predictors of Time 2 performance

Of all the potential sources of individual variation, skills at Time 1 were the most robust predictor of Time 2 skills. This was unsurprising given previous studies with a similar design (Unsworth *et al*. 2015; Van Mensel and Galand 2022). Contrary to our initial hypothesis, we found that participants' performance at Time 2 was associated with their out-of-school engagement with English. Specifically, participants with a higher frequency of extracurricular English classes tended to have higher levels of vocabulary and grammar. It is possible that since extracurricular activities tend to have more reduced groups of students than classes at school, they are more conducive to English learning.

We did not find evidence that the frequency of engagement with English reading or TV/videogames was associated with better performance at Time 2. As stated in our review of the literature, it is possible that the sample was overall too young to engage with these activities with such a frequency that would lend itself to robust findings. It is possible that with increasing age, children will engage in more out-of-school experiences with English so that a larger effect becomes apparent (e.g. Unsworth *et al*. 2015; Sundqvist and Sylvén 2014; Muñoz *et al*. 2018; Van Mensel and Galand 2022).

One of the most robust predictors was maternal education. Children with more educated mothers had better outcomes for the four abilities at Time 2. Other studies have found this association for older students as well (Van Mensel and Galand 2022). The robustness of this finding for our current sample brought us to probe further into the association between maternal education and FL skills. A series of Pearson's correlations did not find any significant correlation between maternal years of education, on the one hand, and children's frequency of engagement with English reading ($r = 0.006$, $p = .935$), TV/video games ($r = –0.088$, $p = .248$), extracurricular activities ($r = 0.003$, $p = .996$), or English AOA ($r = –0.052$, $p = .493$). However, an ordinal model found that more educated mothers reported higher levels of English proficiency ($p < .001$). Since none of the mothers in the sample used English to communicate with their children, the implications of this finding are unclear. It is possible that more educated mothers find ways to support their children's English development that were not controlled for in this study (e.g. helping with English homework).

Finally, and despite this study testing children at the very onset of formal schooling, we found no evidence that the gender of the participants, nor their English AOA, affected their performance

in any of the abilities at Time 2. This study, then, is in line with previous ones that have found similar null results (e.g. Muñoz 2011; De Wilde and Eyckmans 2017).

Conclusions and limitations

The main findings of this study bear specific implications for FL instruction. First, CLIL was not found to be a significant predictor of Time 2 performance. It is possible that CLIL students had not had a sufficient amount of added English exposure at school in terms of intensity per week to make a difference, that they were too young to benefit from the CLIL approach, or that the time span to which they were exposed to CLIL was too short. Regardless, these results, together with that of other studies with similar findings, suggest that Grade 1 may not be the optimal time to introduce CLIL. Longitudinal studies that follow CLIL learners for a longer period of time will be able to determine when students start benefiting from CLIL significantly. However, the timing is key. It is expected that CLIL students will eventually show advantages over Non-CLIL students due to increased FL exposure. But, if it is found that students who start CLIL in mid or late primary catch up to their counterparts who have followed CLIL since Grade 1, delaying the onset of CLIL implementation would lead to the optimization of school resources and be altogether more cost-effective.

Secondly, skills at Time 1 (beginning of Grade 1) were the best predictor of skills at Time 2 (end of Grade 1). This finding suggests that disparities in children's FL skills at the onset of primary may remain or even increase as time progresses. Since this was not a retrospective study, it was not a goal to determine what may cause these initial differences prior to the onset of formal schooling. However, FL teachers may find it useful to assess children's skills at the early stages of primary to find out what students may be in need of extra support.

Thirdly, children who engage in extracurricular activities in English seem to have some advantages, at least with respect to vocabulary and grammar. Unfortunately, these activities may not be accessible for families with limited resources. As such, encouraging parents to enroll their children in such activities should be done with caution.

Finally, having a more educated mother predicted increased gains in all abilities, though the mechanics underlying such an association are unclear. While maternal education is not malleable, FL instructors at school should be sensitive to the fact that variations in the educational level of mothers can have implications for the students' progress in class.

The conclusions from this study should be considered together with its two main limitations. First, development was measured at the beginning and end of one academic year (around 8 months). As such, this timespan may have been insufficient for gains to emerge in these very young CLIL students. A longitudinal study that follows students for a longer time period (e.g. Grades 1–6) would be able to determine when the extra FL input conferred by CLIL may meaningfully improve measurable outcomes and inform the debate on when the optimal time to start CLIL is. In addition, as one anonymous reviewer pointed out, we were not able to analyze the impact of specific aspects of the CLIL programs implemented in each school (e.g. the learning activities that teachers employed, the specific subjects that were taught as CLIL). Combining 14 schools increased the generalizability of our results but inevitably limited the granularity of our analyses and findings. Thus, we believe that in order to develop a comprehensive understanding of the effects of CLIL in primary schools we need to combine larger-scale studies such as the present one with others that narrow in on school-specific aspects of CLIL implementation.

Despite these limitations, the present study contributes to our limited knowledge of the effects of CLIL instruction at early stages of primary education in a multilingual context while controlling for other sources of individual differences that should not be neglected in this type of research.

Supplementary data

Supplementary material is available at *Applied Linguistics* online.

Notes on Contributor

Adriana Soto-Corominas is a researcher at the Department of Applied Linguistics at Universitat Internacional de Catalunya. She is interested in individual variation in bilingual and multilingual development in childhood. She currently holds a Marie Sklodowska-Curie action grant to investigate the development of Catalan, Spanish, and English in multilingual children schooled in Catalunya. *Address for correspondence*: Department of Applied Linguistics, Universitat Internacional de Catalunya, Barcelona, Catalunya, Spain. <asotoc@uic.cat>

Helena Roquet is an assistant professor at the Department of Applied Linguistics at Universitat Internacional de Catalunya, where she is the Director of the Institute for Multilingualism. Her research focuses on the effects of the implementation of Content and Language Integrated Learning (CLIL) approaches during the childhood and teenage years.

Marta Segura is a researcher at the Department of Applied Linguistics at Universitat Internacional de Catalunya. She investigates the effects of CLIL in preschoolers.

Acknowledgements

We thank all participants, their parents, and schools for their participation.

Funding

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101027137.

References

Aguilar, M. and C. Muñoz. 2014. 'The effect of proficiency on CLIL benefits in Engineering students in Spain,' *International Journal of Applied Linguistics* 24/1: 1–18.

Agustín-Llach, M. P. 2015. 'Age and type of instruction (CLIC vs traditional EFL) in lexical development,' *International Journal of English Studies* 16/1: 75–96. doi:10.6018/ijes/2016/1/220691.

Agustín-Llach, M. P. and A. Canga Alonso. 2014. 'Vocabulary growth in young CLIL and traditional EFL learners: Evidence from research and implications for education,' *International Journal of Applied Linguistics* 26/2: 211–27. doi:10.1111/ijal.12090.

Alexiou, T. 2015. 'Comic series and "Peppa Pig": A hidden treasure in language learning' in M. Tzakosta (ed): *Language Learning and Teaching in Multi-cultural Environments* (pp. 187–206). Gutenberg.

Bates, D., *et al*. 2015. 'Fitting linear mixed-effects models using lme4,' *Journal of Statistical Software* 67/1: 1–48. doi:10.18637/jss.v067.i01.

Bishop, D. V. M. 2003. *The Test for Reception of Grammar, Version 2 (TROG-2)*. Harcourt Assessment.

Bret, A. 2011. 'Implementing CLIL in a primary school in Spain: The effects of CLIL on l2 English learners' oral production skills'. Master's thesis. Universitat Autònoma de Barcelona.

Canga Alonso, A. 2015. 'Receptive vocabulary of CLIL and non-CLIL primary and secondary school learners,' *Complutense Journal of English Studies* 23: 59–77. doi:10.5209/revCJES.2015.v23.51301.

Codó, E. 2022. 'The dilemmas of experimental CLIL in Catalonia,' *Journal of Multilingual and Multicultural Development* 43/4: 341–57.

Dalton-Puffer, C. 2008. 'Outcomes and processes in Content and Language Integrated Learning (CLIL): Current research from Europe' in W. Delanoy and L. Volkmann (eds): *Future Perspectives for English Language Teaching*. Carl Winter.

De Wilde, V., M. Brysbaert, and J. Eyckmans. 2022. 'Formal versus informal L2 learning: How do individual differences and word-related variables influence French and English L2 vocabulary learning in Dutch-speaking children?,' *Studies in Second Language Acquisition* 44/1: 87–111.

De Wilde, V., M. Brysbaert, and J. Eyckmans. 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: 171–85. doi:10.1017/S1366728918001062.

De Wilde, V. and J. Eyckmans. 2017. 'Game on! Young learners' incidental language learning of English prior to instruction,' *Studies in Second Language Learning and Teaching* 7/4: 673–94. doi:10.14746/ssllt.2017.7.4.6.

DeKeyser, R. M. 2000. 'The robustness of the critical period effects in second language acquisition,' *Studies in Second Language Acquisition* 22/4: 499–533. doi:10.1017/S0272263100004022.

Dunn, L. M. 2019. *The Peabody Picture Vocabulary Test* (5th edn). NCS Pearson.

EF. 2022. EF English Proficiency Index [Electronic resource]. https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2022/ef-epi-2022-english.pdf

Franco Hidalgo-Chacón, J. P., I. Rodríguez-Arteche, and M. M. Martínez-Aznar. 2022. '¿Qué hacen los estudiantes de Educación Primaria españoles fuera del horario académico? Actividades extraescolares,' *Revista Complutense de Educación* 33/3: 459–74.

Gagarina, N., *et al*. 2019. 'MAIN: Multilingual assessment instrument for narratives – Revised,' *ZAS Papers in Linguistics* 63.

Gayete, G. 2022. 'The effects of CLIL on L3 students' oral production and comprehension in a Primary school context' in M. Pallarés-Renau, F.J. Vellón, and P. Salazar-Campillo (eds): *Investigacions transversals i integradores en Ciències Humanes i Socials [Transversal and Integrated Research in Human and Social Sciences]*. Emergents, 3.

Generalitat de Catalunya. 2018. *El model lingüístic del sistema educatiu de Catalunya*. Departament d'Ensenyament.

Hannibal Jensen, S. H. 2017. 'Gaming as an English language learning resource among young children in Denmark,' *CALICO Journal* 34/1: 1–19. doi:10.1558/cj.29519.

Hartig, F. 2020. 'DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models,' R package version 0.3.3.0, available at https://CRAN.R-project.org/package=DHARMa

Hoff, E. 2006. 'How social contexts support and shape language development,' *Developmental Review* 26/1: 55–88.

Housen, A. 2012. 'Time and amount of L2 contact inside and outside the school – Insights from the European Schools' in C. Muñoz (ed): *Intensive Exposure Experiences in Second Language Learning* (pp. 111–138). Multilingual Matters.

IDESCAT. 2020. 'Taxa neta d'escolarització. Alumnes de 2 i 3 anys,' Accessed 20 February 2023. https://www.idescat.cat/indicadors/?id=anuals&n=10369&tema=educa

Jiménez-Catalán, R. M., Y. Ruíz de Zarobe, and J. Cenoz. 2006. 'Vocabulary profiles of English foreign language learners in English as a subject and as a vehicular language,' *Views* 15/3: 23–6.

Kuppens, A. H. 2010. 'Incidental foreign language acquisition from media exposure,' *Learning, Media and Technology* 35/1: 65–85.

Lindgren, E. and C. Muñoz. 2013. 'The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension,' *International Journal of Multilingualism* 10/1: 105–29.

Lorenzo, F., S. Casal, and P. Moore. 2010. 'The effects of content and language integrated learning in European education: Key findings from the Andalusian bilingual sections Evaluation project,' *Applied Linguistics* 31/3: 418–42.

Mackey, A. and Goo, J. 2007. 'Interaction research in SLA: A meta-analysis and research synthesis' in A. Mackey (ed): *Conversational Interaction and Second Language Acquisition. A Series of Empirical Studies* (pp. 377–419). Oxford University Press.

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

Muñoz, C. 2011. 'Input and long-term effects of starting age in foreign language learning,' *IRAL* 49/2: 113–33. doi:10.1515/iral.2011.006.

Muñoz, C. 2015. 'Time and timing in CLIL: A comparative approach to language gains,' In Maria Juan-Garau & Joana Salazar-Noguera (eds.), *Content-based language learning in multilingual educational environments* 87–104. Berlin: Springer.

Muñoz, C. 2020. 'Boys like games and girls like movies: Age and gender differences in out-of-school contact with English,' *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 33/1: 171–201. doi:10.1075/resla.18042.mun.

Muñoz, C., Cadierno, T. and Casas I. 2018. '*Different starting points for English language learning: A comparative study of Danish and Spanish young learners,' Language Learning* 68/4: 1076–109. doi:10.1111/lang.12309.

Navarro, D. J. 2015. *Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners (Version 0.6)*. University of New South Wales.

Nieto, E. 2016. 'The impact of CLIL on the acquisition of L2 competences and skills in primary education,' *International Journal of English Studies* 16/2: 81–101. doi:10.6018/ijes/2016/2/239611.

Paradis, J. 2019. 'English second language acquisition from early childhood to adulthood: The role of age, first language, cognitive, and input factors' in *Proceedings of the BUCLD* (vol. 43, pp. 11–26).

Paradis, J., Genesee, F., and Crago, M. B. 2021. *Dual Language Development and Disorders* (3rd edn). Paul H. Brookes Publishing Co.

Pladevall-Ballester, E. and A. Vallbona. 2016. 'CLIL in minimal input contexts: A longitudinal study of primary school learners' receptive skills,' *System* 58: 37–48. doi:10.1016/j.system.2016.02.009.

Pérez Cañado, M. L. 2018. 'CLIL and educational levels: a longitudinal study on the impact of CLIL on language outcomes,' *Porta Linguarium* 29: 51–70. doi:10.30827/Digibug.54022.

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: 142–68.

R Core Team. 2022. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/

Ryu, D. 2013. 'Play to learn, learn to play: Language learning through gaming culture,' *ReCALL* 25/2: 286–301.

Serra, C. 2007. 'Assessing CLIL at primary school: A longitudinal study,' *International Journal of Bilingual Education and Bilingualism* 10/5: 582–602. doi:10.2167/beb461.0.

Sundqvist, P. 2009. *Extramural English Matters: Out-of-school English and its Impact on Swedish Ninth Graders' Oral Proficiency and Vocabulary*. Karlstad University. Diss.

Sundqvist, P. and L. K. Sylvén. 2014. 'Language-related computer use: Focus on young L2 English learners in Sweden,' *ReCALL* 26/1: 3–20.

Sundqvist, P. and P. Wikström. 2015. 'Out-of-school digital gameplay and in-school L2 English vocabulary outcomes,' *System* 51: 65–76.

Sylvén, L. K. 2013. 'CLIL in Sweden–why does it not work? A metaperspective on CLIL across contexts in Europe,' *International Journal of Bilingual Education and Bilingualism* 16/3: 301–20.

Sylvén, L. and P. Sundqvist. 2012. 'Gaming as extramural English L2 learning and L2 proficiency among young learners,' *ReCALL* 24/3: 302–21.

Unsworth, S., *et al*. 2015. 'An investigation of factors affecting early foreign language learning in the Netherlands,' *Applied Linguistics* 36/5: 527–48.

Van Mensel, L. and B. Galand. 2022. 'Testing the predictive power of executive functions, motivation, and input on second language vocabulary acquisition: a prospective study,' *European Journal of Psychology of Education* 1–20. doi:10.1007/s10212-022-00597-x.

Winter, B. 2019. *Statistics for Linguists: An Introduction using R*. Routledge.

Xanthou, M. 2011. 'The impact of CLIL on L2 vocabulary development and content knowledge,' *English Teaching: Practice and Critique* 10/4: 116–26.