tesol QUARTERLY
tesol international association

Intensity Matters Inside and Outside Primary School: Evidence from High-CLIL, Low-CLIL, and Non-CLIL Learners

AMPARO LÁZARO-IBARROLA
Institute for Advanced Social Research I-COMMUNITAS, Public University of Navarre (UPNA)
Pamplona, Spain

Abstract

Research involving secondary school EFL learners has demonstrated that greater intensity of exposure, via CLIL lessons, yields notable benefits. However, studies in primary school are scarce and less optimistic. Furthermore, little is known about the effects of different degrees of CLIL intensity and of learners' exposure to Extramural English (EE) through formal or informal out-of-school activities. To address these gaps, this study examines the impact of CLIL and EE on the proficiency of 180 primary school learners of English (aged 10–11 years) divided into a high-CLIL ($N = 78$), a low-CLIL ($N = 56$), and a non-CLIL ($N = 46$) group. Results showed some advantages among high-CLIL learners, including higher scores and less intragroup variability, while no significant differences were observed between low-CLIL and non-CLIL learners. EE activities were common, with several weak but positive correlations found between EE and proficiency, particularly involving watching TV and reading, and mainly affecting non-CLIL learners.
doi: 10.1002/tesq.3369

INTRODUCTION

The issue of intensity lies at the core of second language acquisition (SLA) research and is of great relevance in school contexts.

Understanding "time and timing" in L2 acquisition (Muñoz, 2015; Swain, 1981), that is, when to introduce languages (timing) and how intensely (time) is essential to understand the potential and limitations of language acquisition in instructed settings. At present, most schools in European contexts have become quite homogeneous in terms of "timing," offering a very early introduction of L2 lessons. However, there are important differences in the amount of time (intensity) devoted to teaching the L2s, especially as regards the amount of content and language integrated learning (CLIL) lessons.

Before CLIL, while learners were in their schools they either received a lot of exposure (immersion contexts) to the target language, or very little (foreign language (FL) contexts). The umbrella term CLIL (Dalton-Puffer, Llinares, Lorenzo, & Nikula, 2014) defines a quantity of exposure somewhere between these two extremes, with learners receiving an undetermined number of subjects in the target language. This means that the number of CLIL lessons can vary a great deal from context to context, that is, not all learners labeled as "CLIL" receive the same amount of exposure. In spite of this, the "CLIL" label has been simply used in opposition to "non-CLIL," and claims have been made regardless of CLIL intensity, with very few exceptions (Merino & Lasagabaster, 2018; Ruiz de Zarobe, 2008; Somers & Llinares, 2021).

On the other hand, extramural English (EE), that is, exposure to English outside the walls of the classroom, formally or informally, and with or without the intention of improving the language, is becoming more and more frequent among young learners in FL contexts (Muñoz *et al.* 2018; Sundqvist, 2009; Sundqvist & Sylvén, 2016). To date, however, very few authors have included measures of EE in their research (e.g., Muñoz, Cadierno, & Casas, 2018; Soto-Corominas, Roquet, & Segura, 2023; Sunqvist and Sylvén, 2016).

Last but not least, CLIL research suffers from a significant sampling bias (Godfroid & Andringa, 2023), as the vast majority of studies have been conducted in secondary schools. Consequently, very little is known about the impact of CLIL on younger children (under the age of 12 years) who are learning the language in primary schools, where, paradoxically, CLIL programs are widely implemented (Lorenzo, Granados, & Rico, 2021; Muñoz, 2015). Besides, the findings from the available studies are inconclusive, so much so that it is still unclear whether primary school learners benefit from CLIL or not and, if they do, what aspects or skills are affected (Gálvez Gómez, 2021; Housen, 2012).

Within this context, the present study constitutes an attempt to understand the effectiveness of CLIL intensity and EE among primary school learners. To do this, the impact of CLIL exposure and EE on

the proficiency of 180 children (aged 10–11 years) learning English as a foreign language (EFL) in primary school was analyzed.

BACKGROUND LITERATURE

CLIL Versus Non-CLIL in Primary School

The majority of studies, whether small scale or large scale, have found advantages for students in CLIL programs compared to students in non-CLIL programs. This is especially true for adolescent learners in secondary school contexts, as outlined in several reviews (Dalton-Puffer, Nikula, & Smit, 2010; Graham, Choi, Davoodi, Razmeh, & Dixon, 2018; Muñoz, 2015; Pérez Cañado, 2021). The most positive results have been reported in Spain, which researchers attribute to the country's relatively low proficiency in English as a Foreign Language (Goris, Denessen, & Verhoeven, 2019).

Studies comparing CLIL and non-CLIL groups with younger learners—that is, with children under 12 years who are still in primary school—are much scarcer, and their findings are less conclusive. While advantages for CLIL children have been reported in aspects such as vocabulary (Jiménez Catalán & Ruiz de Zarobe, 2009) or certain dimensions of L2 motivation (Lázaro-Ibarrola & Azpilicueta-Martínez, 2024; Pladevall-Ballester, 2018), most studies reveal more similarities than differences when CLIL and non-CLIL groups are compared. For instance, Moreno and de Diezmas (2016) carried out a large-scale study with Spanish children in primary school in CLIL and non-CLIL strands. Her results showed advantages in speaking and interaction for CLIL learners, but no differences in reading, listening, and writing, which the author partly attributed to the young age of the learners (9–10-year-olds) and the limited amount of extra exposure in the CLIL group (250 h in total). With considerably younger students (6–7-year-olds), Soto-Corominas et al. (2023) analyzed the impact of CLIL on receptive and productive skills, and no advantages in CLIL learners were found. Pladevall-Ballester and Vallbona (2016) found that a group of CLIL learners, after 54 h of CLIL exposure, had no advantages in reading and were, in fact, inferior to EFL learners in their listening skills. The authors concluded that EFL-only exposure had been more effective than CLIL in the development of listening skills.

Several CLIL studies have also demonstrated that the differences between CLIL and non-CLIL learners increase as the students get older and, also, that other factors might be overweighing the impact of CLIL, for instance, school type (Pérez Cañado, 2018) or initial aptitudes

(Soto-Corominas et al., 2023). Pérez Cañado (2018) conducted a large-scale longitudinal study across various communities of Spain. Her results for primary school students showed statistically significant differences in favor of the bilingual (CLIL) group, especially regarding speaking. However, effect sizes were low and some differences waned when considering other variables (e.g., charter vs public strands). In contrast, her results from secondary school showed much more pronounced differences. In another study, Gálvez Gómez (2021) found very clear differences in speaking when comparing CLIL and non-CLIL learners in secondary school, but no differences among learners in primary education (year 6). In a similar vein, Agustín-Llach and Canga Alonso (2016) found that CLIL and non-CLIL groups were similar in their receptive vocabulary size in the years of primary school, but this aspect increased significantly in favor of CLIL students as they got older.

Outside Spain, a large-scale study by Housen (2012) compared, among other groups, Italian learners of English (grades 2–5) in an EFL context with learners in an elitist CLIL context in Italy (European School context). His research showed very similar results in the two groups in terms of rate and L2 learning outcomes. Housen (2012) concluded that during the first 3 years of primary school, the extra exposure was not enough for students to benefit from it. Along the same lines, Stotz and Meuter (2003) measured listening and speaking in 178 CLIL and non-CLIL learners in Zurich. In this case, the CLIL learners fared better in listening but the results regarding oral production revealed no differences.

In sum, research on the effects of CLIL is scarce in primary education and the few available studies present learners exposed to varying degrees of CLIL intensity. The results regarding the different skills also vary from one study to another, but finding few or no differences between CLIL and non-CLIL cohorts seems to be the prevalent trend.

CLIL Intensity

A lot of variability can be observed in terms of weekly and accumulated exposure in CLIL contexts. For instance, CLIL students received 1 extra weekly lesson in the study by Pladevall-Ballester (2018) and by Pladevall-Ballester and Vallbona (2016); 2 lessons in Lázaro-Ibarrola and Azpilicueta-Martínez (2024); 2–4 lessons in the large-scale study by Pérez Cañado (2018); 5–6 in one of the groups in Ruiz de Zarobe (2008); and 13 in one of the groups in the study carried out by Somers and Llinares (2021). As for accumulated CLIL exposure, CLIL learners had received 250–270 h more than non-CLIL learners in the study by Housen (2012) and Moreno and de Diezmas (2016),

but only 54 h in the study by Pladevall-Ballester and Vallbona. To make matters worse, not all studies describe intensity of exposure in sufficient detail and it is often hard to know the exact amount of weekly and/or accumulated exposure. In her review study, Muñoz (2015) concluded that differences between CLIL and non-CLIL groups tended to emerge when additional exposure to CLIL surpassed 300 h in secondary school and suggested that this amount should probably be larger in primary school. All of this emphasizes the need to explore CLIL intensity. Nevertheless, to date, this issue is rarely present in the otherwise abundant CLIL literature (Merino & Lasagabaster, 2018; Ruiz de Zarobe, 2008; Somers & Llinares, 2021).

Specifically connecting CLIL intensity and linguistic proficiency, there are very few studies, and all of them have been conducted in the context of secondary school in Spain. Merino and Lasagabaster (2018) measured the proficiency of 393 students enrolled in programs of three different intensities: non-CLIL, CLIL- (3.4 h per week), and CLIL+ (8.4 h per week). These authors tested the students in the four skills at two points in time, when they were in grade 7 (11–12 year olds) and when they were in grade 8 (12–13 year olds). Their findings revealed a significant advantage in the four skills in the high-CLIL group. On the other hand, they found no significant differences between non-CLIL and CLIL-. They suggested that low intensity is not effective while high intensity seems to enhance proficiency. In two small-scale studies, Ruiz de Zarobe (2008, 2010) measured the proficiency of higher and lower intensity groups (to which she referred to as CLIL1 and CLIL2), although she did not control for potential differences before testing. The CLIL1 group received one subject in English (Science), which meant 3–4 weekly hours; the CLIL2 group received two subjects (Science and Literature), which meant 5–6 weekly hours; and the non-CLIL group did not receive any CLIL lessons. Ruiz de Zarobe (2008) analyzed the oral narratives of the learners at three points in time (aged 14, 15, and 17 years) considering the following categories: pronunciation, vocabulary, grammar, fluency, and content. When comparing CLIL1 and CLIL2 at time 1 (aged 14 years), after only 1 year of CLIL exposure (35 extra hours), there was no significant difference between CLIL1 and CLIL2. However, the non-CLIL group showed lower scores, which the author attributed to their lack of CLIL exposure. At time 2, when students were 15, CLIL2 scored higher in four categories. A study of the written production of these same groups of learners showed similar results (Ruiz de Zarobe, 2010), but the advantages of CLIL learners were smaller, which the author attributed to a lack of practice of the writing skill. To the best of our knowledge, there are no studies addressing CLIL intensity in the primary school context.

Extramural Exposure

As Sundqvist and Sylvén (2016) explained, the term "extramural" comes from the Latin words "extra," meaning outside, and "mural," meaning walls. Thus, extramural English (EE) is used to refer to the English that learners come in contact with outside the walls of the classroom. In the present study, we have classified EE into two types: informal and formal. In informal EE, the learners simply perform activities in English, such as reading, listening to music, digital gaming, or navigating on the Internet. The main goal the learners have when engaging in these informal activities is to convey or understand messages, that is, learners do not engage in the activities in the pursuit of improving their proficiency in English, although this may also be a desirable objective for them. On the contrary, formal EE refers to out-of-school language lessons, which are very frequent in some contexts, such as Spain (Lancaster, 2018; Muñoz, 2015).

As highlighted by several authors, English learners frequently have experiences with the target language outside of school and this contact can affect their learning (Kuppens, 2010; Lindgren & Muñoz, 2013; Muñoz et al., 2018). In fact, it has been demonstrated that children achieve a certain level of English proficiency through out-of-school exposure prior to receiving English lessons (De Wilde, Brysbaert, & Eyckmans, 2022). Furthermore, this exposure can be more relevant than classroom exposure at an early age (Soto-Corominas et al., 2023). Yet, with few exceptions, information on EE is missing in the CLIL literature. In particular, very little is known about the impact of EE with children in FL contexts (Muñoz et al., 2018). Most of our knowledge on this issue comes from a series of studies with young learners in Sweden, a society offering many opportunities to use English outside school (Sundqvist & Sylvén, 2016). The results suggest that EE has a positive impact on language achievements (Olsson & Sylvén, 2015; Sundqvist, 2009; Sylvén & Sundqvist, 2015). For instance, Sundqvist (2009) conducted a study with Swedish learners of English (aged 15–16 years) and found a strong correlation between EE, oral proficiency and vocabulary. In a later study with 76 Swedish children (aged 10–11 years), Sylvén and Sunqvist (2015) concluded that young Swedish learners benefit from EE and that their language skills and EE are mutually reinforcing. In another study, Olsson and Sylvén (2015) investigated progress in academic vocabulary use in 230 Swedish students (aged 16–19 years). Although they did not find any significant impact of EE on progress of academic vocabulary over time, the authors found that CLIL students devoted more time to EE than their non-CLIL peers.

Turning to other countries, in a very large-scale study, Lindgren and Muñoz (2013) found that out-of-school exposure, and especially watching TV, had a positive impact on the listening and reading outcomes of 865 young learners (aged 10–11 years) from seven different European contexts. Along the same lines, Muñoz et al. (2018) also considered out of school exposure in a study analyzing the receptive abilities in vocabulary and grammar in English by Spanish/Catalan and Danish children (aged 7 and 9 years). Their analyses showed that exposure to films in English positively correlated with listening comprehension in the older group. In Belgium, and in the context of secondary school, Peters, Noreillie, Heylen, Bulté, and Desmet (2019) analyzed, among other aspects, the impact of extramural exposure to foreign language input on the receptive vocabulary knowledge in two foreign languages (French and English) of Flemish learners. Their findings showed that, despite fewer years of English instruction, learners' vocabulary knowledge in English was considerably larger than their French *vocabulary* knowledge. The authors suggested that the reason for this could be the learners' large amounts of out-of-school exposure to English language input. In particular, learners' online activities seemed to have a positive effect on vocabulary knowledge. However, the authors also acknowledged the fact that linguistic distance had not been considered in their study and English was much closer to Flemish than French.

As for formal EE in the form of extra lessons outside school, there is very little research. Lancaster (2018) conducted a quantitative study in Andalusia (Spain) to discern the impact of CLIL and extra lessons. To do this, Lancaster (2018) controlled several variables (motivation, verbal intelligence, socioeconomic status, or the CLIL program itself) and examined data from 156 CLIL learners and 162 EFL-only students. The results showed that secondary CLIL students without extra lessons (to which the author refers as formal instruction) outperformed non-CLIL learners with extra lessons, suggesting that taking part in CLIL programs is more conducive to language learning than attending private academies. In her study, Lancaster (2018) also highlighted the need to explore the impact of private lessons in the context of Spain, where learners so often attend private language academies.

THE STUDY

CLIL programs in primary schools are widespread in Spain, but research on their impact remains very limited. In addition, CLIL intensity and exposure to the language outside school have been largely ignored in the literature. In order to address these gaps, we examined

the impact of CLIL intensity and EE in 180 primary school learners of English (aged 10–11 years) divided into three groups: a high-CLIL group, with 5–7 h of CLIL; a low-CLIL group, with 2–3 h of CLIL; and a non-CLIL group, with only EFL lessons. Our research questions are the following:

RQ1. Are there differences between the scores obtained by high-CLIL, low-CLIL, and non-CLIL learners in the linguistic tests?

a. Are there differences in the mean score?
b. Are there differences in each of the skills (listening, reading, speaking, and writing)?

RQ2. Does EE have an impact on students' proficiency?

a. Is there a correlation between formal EE (extracurricular lessons) and the test scores?
b. Is there a correlation between informal EE (TV viewing, listening to music, video games, etc.) and the test scores?

METHODS

Participants

Our original data set comprised 264 students (aged 10–11 years) in primary school in a bilingual community (Basque/Spanish) in the north of Spain. The learners belonged to eight state schools in the same city. All schools were state schools located in working-class neighborhoods. In the area, students primarily choose their school based on their choice of vehicular language (Basque or Spanish) an on proximity to their residence. All schools offered 5-weekly EFL lessons from the beginning of schooling. The CLIL schools also offered CLIL lessons from the first grade of primary. The CLIL lessons comprised Art and Craft, Science, and Social Sciences. In all cases, CLIL was the only program in the CLIL schools, meaning that all the learners participated in the CLIL programs, thereby reducing social or academic selection (Mearns, de Graaff, & Coyle, 2020).

To address some of the shortcomings present in research regarding the lack of control of certain variables (Dalton-Puffer et al., 2010), questionnaires gathering biographical, linguistic, and socioeconomic information were administered prior to the study. The questionnaires were administered in the learners' L1s (Spanish and Basque). The teachers helped the children understand all the items on the

questionnaires, and the students were also allowed to take them home to complete some of the questions with the help of their parents.

Based on the biographical and linguistic information gathered, children who used English at home or with friends, who had started school later than the first year of primary education (when aged 6–7 years), or who had spent more than 2 weeks in an English-speaking country were excluded from the study. The questionnaire also provided information on attendance of extracurricular English lessons. In this respect, and to make a clear distinction between students receiving and not receiving extra lessons, only students who (i) had never received extra lessons and (ii) students who had received extra lessons for at least 2 years (including, at least, the present or the past year) were included. That is, all students who did not fit neatly into one of these categories were excluded from the sample.

Our final sample comprises 180 children divided into three groups depending on the number of hours of CLIL they received in their schools (Table 1).

As Table 1 shows, the learners in the non-CLIL group, who only received 5 FL lessons per week, had accumulated a total exposure of around 1000 h at the time of data collection. The learners in the low-CLIL group, who received 5 EFL lessons and also 2–3 CLIL lessons, had accrued a total exposure of around 1500 h (approximately 1000 h of EFL lessons and 500 h of CLIL lessons). The high-CLIL group, with learners receiving 5 EFL lessons and 5–7 CLIL lessons per week, reached around 2200 h of exposure (1000 h of EFL and 1200 h of CLIL). This means that, at the time of data collection, the low-CLIL group had received more than 500 extra hours of exposure compared to the non-CLIL group, and the high-CLIL group had received an additional exposure of over 700 h compared to the low-CLIL group, and over 1200 h compared to the non-CLIL group.

TABLE 1
Participants in the Study and Intensity of Exposure in School

	Mean age	EFL lessons	CLIL lesson	Weekly exposure	Accumulated exposure
High-CLIL ($N = 78$)	11	5	5–7	10–12	2200 h
Low-CLIL ($N = 56$)	10.4	5	2–3	7–8	1500 h
Non-CLIL ($N = 46$)	10.9	5	0	5	1000 h

TABLE 2
SES in the Three Groups

Group	*N*	Mean	SD
High-CLIL	78	11.15	3.20
Low-CLIL	56	11.58	2.62
Non-CLIL	46	11.52	3.19
Statistics KW	H = .876; p = .645		

Finally, the students also completed an official socioeconomic status (SES) test. This test, with a maximum score of 20 points, included information on parental education, occupation and income, space in the house, and access to resources (books, culture, and technology). This is important given that SES levels seem to correlate with L2 performance (Butler & Le, 2018; Lorenzo et al., 2021). The results revealed that SES levels were very similar in the three groups (Table 2) and the small differences were not significant (p = 0.645).

Ethical Issues

The project under which this study was conducted obtained institutional approval and the research was carried out in accordance with the institution's approved research protocol. All the tests were supervised and approved by the school teachers and administered with their help. Authorized consent forms were signed by the parents or legal guardians of the participants prior to data collection. As for the children themselves, the teachers explained the objectives of the research to them, and they were allowed to decide whether they wanted to participate or not. All the children decided to participate.

Research Instruments

The 180 participants had to complete an EE questionnaire and four linguistic tests. All the data were collected between February and May, and the questionnaire and tests were administered in the same order in all the schools (see details in Table 3). During the first week, participants completed the EE questionnaire, providing information about their exposure to English outside school. In subsequent weeks, they took the proficiency tests. First, they completed listening and reading tests during their respective language sessions at school. Next, they wrote the compositions during one English session and, finally, they

TABLE 3
Test Administration

Week	Test	Duration	Administration
Week 1	EE questionnaire	30 min	In class with the teacher
Weeks 2–3	Listening test (English)	40 min	Online in class with the center teacher
	Reading test (English)	40 min	Online in class with the teacher
Weeks 4–5	Writing test (English)	30 min	On paper in class with the teacher
Weeks 6–9	Speaking test (Spanish)	10 min	Individually with the teacher in a s eparate classroom

took the oral tests. For these oral assessments, teachers recorded students individually, taking them out of the classroom to conduct the test. In the following subsections, more details on the tests are provided.

Proficiency Tests

To measure proficiency in the four skills, the students completed four proficiency tests. For listening and reading comprehension, a Cambridge Flyers Test (A2) from the Cambridge English Qualifications Papers for Young Learners series (Cambridge, 2018) was employed. The tests were administered in class by the school teachers. These tests are designed for learners in primary school and correspond to the A2 level of the CEFR. In the tests, the participants had to choose the right answer among multiple options or match elements. The reading test comprised 28 questions, and the listening test comprised 19 questions. Correct answers were given 1 point, and incorrect answers received no points.

For speaking and writing, the students had to narrate a story following six pictures (Heaton, 1966; see Appendix A) chosen in agreement with the school teachers. The oral production of the students was recorded individually. The compositions were written by hand during one school session. For evaluating the speaking test, we adopted the rubric developed by Merino and Lasagabaster (2018). This assessment tool examines five key areas: pronunciation, grammar, fluency, coherence, and vocabulary. The final score is calculated as the average of these five components. To evaluate the written compositions, we utilized the Cambridge writing assessment subscales designed for the A2 Key for Schools exam. This rubric focuses on three dimensions: content, organization, and language. The final score is derived from the mean of these three equally weighted areas. Both tests were rated by

one independent rater who was an experienced English teacher familiar with the language production of children of this age. A second rater, who was a member of the research team, evaluated 20% of the data. Reliability was excellent for both skills, with an intraclass correlation coefficient (ICC) of .92 for speaking and .96 for writing.

Extramural Exposure Questionnaire

This questionnaire was designed to measure EE. It comprised a set of items related to the amount of time devoted to out-of-school exposure to English. This exposure comprised (i) formal exposure in the form of private lessons at home or in language schools; and (ii) informal exposure.

The questionnaire about private lessons asked learners whether they had ever received private lessons. When the answer was positive, the questionnaire asked learners to choose the school years in which they had taken private lessons and the number of weekly hours.

The questionnaire about informal activities comprised six items, and the value for Cronbach's alpha showed acceptable internal consistency reliability for the survey ($\alpha = .738$). The six activities were the following: TV viewing, reading, listening to music, navigating on the Internet, playing video games, and performing activities such as sports in English. For each of the activities, the learners had to choose the approximate number of hours devoted to them every week (0 h, 1–2 h, 3–4 h, 5–6 h, and more than 7 h). This questionnaire comprised the same questions used by Sundqvist and Sylvén (2016). However, unlike theirs, our questionnaire did not include use of English with parents or friends, given that the students who had this type of exposure were not included in our sample.

Statistics

As for statistical analysis, tests for normality were conducted (Kolmogorov–Smirnov and Shapiro–Wilk). As a non-normal distribution of the data ($p < 0.01$) was found, statistical analyses between groups for the linguistic scores were performed using KW tests. Significance was set at $p \leq 0.05$ and the confidence level was 95%. Where significant differences were detected, post hoc tests (Bonferroni) were conducted. Also, eta squared (η^2) was calculated to measure effect size when considering the three groups and Cohen's d when differences between two groups emerged. To compare the performance of learners with and without formal exposure, the Mann–Whitney test

was used and Spearman's ρ was calculated to observe correlations. To interpret the magnitude of effect sizes, we followed the alternate field-specific scale proposed by Plonsky and Oswald (2014) for L2 research (d values around .40 are small, around .70 medium, and around 1.00 large; for correlation coefficients, ρs close to .25 are small, close to .40 medium, and close to .60 large).

RESULTS

Test Results

As Table 4 shows, when the mean score of the four skills was considered, the high-CLIL group obtained the highest mark (6.59) and presented the greatest homogeneity in score distribution. The low-CLIL group ranked second (6.13) and the non-CLIL group obtained the lowest score (5.9) and, besides, it also presented the greatest variability of scores (see Boxplots in Figure 1). The statistical analysis confirmed significant differences ($p = .028$) with a moderate effect size ($\eta^2 = .041$). Post hoc tests showed that the high-CLIL group was superior to the non-CLIL group ($p = .045$). When the effect size for these two groups (high-CLIL and non-CLIL) was calculated the result showed that it was large (d = .838). In contrast, there were no significant differences between the scores of the high-CLIL and the low-CLIL group ($p = .149$) or between the scores of the low-CLIL and the non-CLIL group ($p = 1$) (see results in Appendix B).

When considering the four skills separately (Table 5), the scores were, in general, quite similar in the three groups, with minimal differences in the receptive skills and slightly larger in the productive skills.

Starting with the receptive skills, the three groups obtained extremely similar scores in reading (see Boxplots in Figure 2) and the post hoc analyses confirmed that there were no statistically significant differences between groups ($p = .413$). In listening (see Boxplots in Figure 3), the high-CLIL and low-CLIL groups were fairly similar while

TABLE 4
Mean Score and Statistical Analyses in the Linguistic Tests for the Three Groups

	Mean score	SD
High-CLIL	6.59	1.09
Low-CLIL	6.13	1.31
Non-CLIL	5.9	1.7
KW H = 7.118; $p = .028$ Eta squared $\eta^2 = .041$		

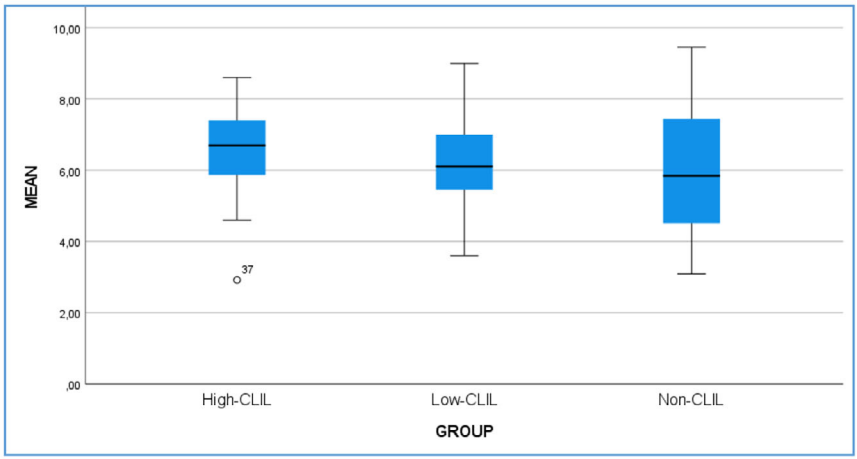


FIGURE 1. Mean Scores in the non-CLIL, low-CLIL, and high-CLIL groups.

TABLE 5

Scores and Statistical Analyses in the Linguistic Tests for the Four Skills of the Three Groups

	Reading		Listening		Speaking		Writing	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
High-CLIL	6.92	1.66	6.18	1.46	6.19	1.95	7.09	1.94
Low-CLIL	6.69	2.17	6.48	1.52	5.28	1.70	6.59	2.17
Non-CLIL	6.45	1.95	5.71	1.75	5.90	2.47	5.76	2.57
KW	$H(2) = 1.769$		$H(2) = 3.805$		$H(2) = 6.261$		$H(2) = 8.824$	
Eta squared	$p = .413$ $\eta^2 = .010$		$p = .149$ $\eta^2 = .034$		$p = .044$ $\eta^2 = .036$		$p = .012$ $\eta^2 = .056$	

the non-CLIL group obtained slightly lower scores. However, the differences between groups did not reach statistical significance ($p = .149$).

In contrast, some significant differences emerged between groups in the productive skills, with small effect sizes for speaking ($\eta^2 = .036$) and a bit larger for writing ($\eta^2 = .056$). The high-CLIL group obtained the highest scores in the oral test (6.19), followed by the non-CLIL group (5.90). In this case, the low-CLIL group obtained the lowest score (5.28). There was also great variability among the scores of non-CLIL learners in contrast with greater homogeneity in the CLIL groups (see Boxplots in Figure 4). Post hoc tests (Appendix B) showed that significant differences only appeared when the high-CLIL and the low-CLIL groups were compared ($p = .037$). The effect size calculated

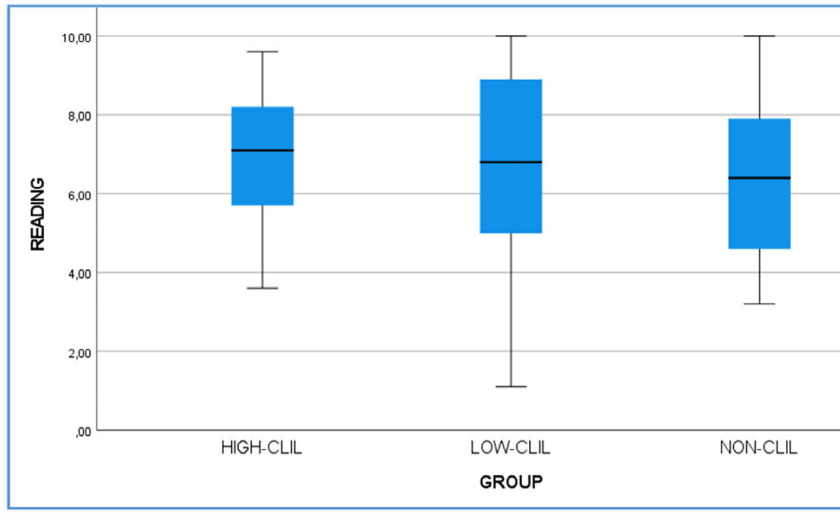


FIGURE 2. Reading scores.

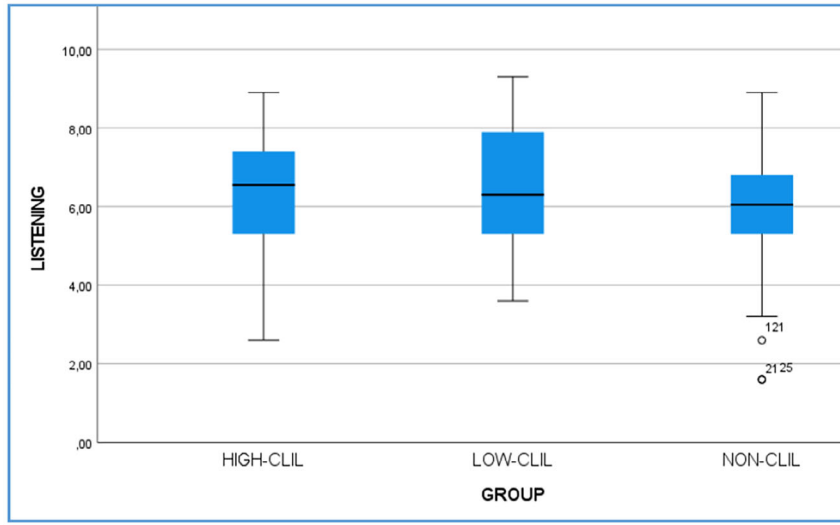


FIGURE 3. Listening scores.

for these two groups in speaking was large (d = 836). The difference between the high- and the non-CLIL groups did not reach statistical significance (p = .911), and neither did the score of the non-CLIL group when compared to the low-CLIL group (p = .644).

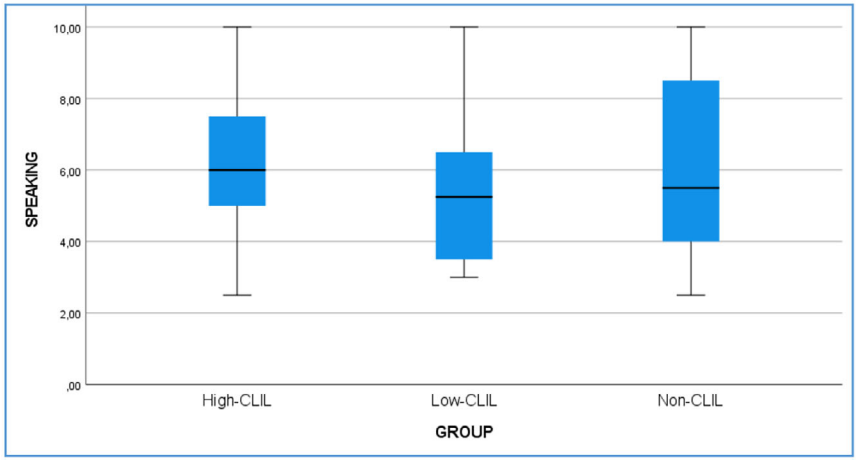


FIGURE 4. Speaking scores.

Finally, the writing skills presented the largest differences in terms of raw scores (high-CLIL: 7.09; low-CLIL: 6.59; and non-CLIL: 5.76) (see Boxplots in Figure 5). Post hoc analyses (Appendix B) revealed significant differences when the high-CLIL group was compared to the non-CLIL group ($p = .009$). The effect size calculated considering only these two groups was large ($d = .974$). Aside from this, there were no significant differences between the high- and the low-CLIL group ($p = .415$), or between the low- and the non-CLIL group ($p = .439$).

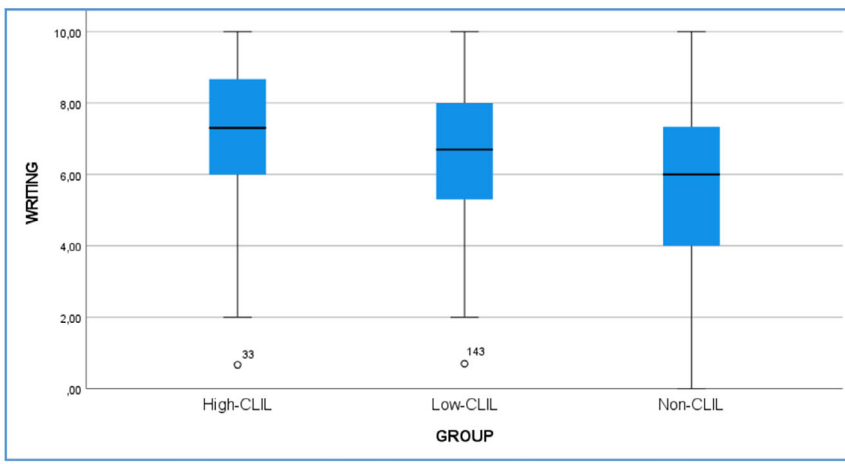


FIGURE 5. Writing scores.

Extramural Exposure

Formal extramural exposure: Extracurricular lessons. Regarding formal EE, 57 students (around 32%) received extracurricular English lessons and 123 (around 68%) had never received extra lessons. The distribution of private lessons slightly varied across groups. Around 22% of the learners (21 children) were taking private lessons in the high-CLIL group; around 30% (14 children) in the non-CLIL group; and around 40% (22 children) in the low-CLIL group.

When the scores of learners with and without extra lessons were compared (Tables 6–8), the results were extremely similar in the high-CLIL group and quite similar in the low-CLIL group. In the non-CLIL group, however, the scores of the learners with extra lessons were slightly higher and, in fact, there was a significant difference in favor of the learners with extra lessons in the writing skill ($p = .048$). The differences in the other skills and in the mean did not reach statistical significance, but they were close to it in reading, listening and the mean in the non-CLIL group, and effect sizes were medium and large. A Spearman test corroborated a positive and moderate association between extra lessons and higher scores in writing in the non-CLIL group ($\rho = .295$; see Appendix C).

Informal extramural exposure. Only 13 children out of our sample of 180 did not report engaging in any informal activity in English outside school. When looking at the specific activities, around 80% of learners listened to music in English (146 children); around 45% watched TV (79 children), read books or comics (82 children), and played videogames (84 children); and only around 33% of the learners engaged in other activities, such as sport or music (59 children). The students in the high-CLIL group devoted more hours to EE (Table 9), specifically, 8.71 h a week on average, while low-CLIL students devoted 6.5 and non-CLIL students 6.2. This difference was significant when

TABLE 6
Mean Scores of Learners with and without Extra Lessons in the High-CLIL Group

	With extra lessons		Without extra lessons		Statistics	
	M	SD	M	SD	*p*	*d*
Reading	7.05	1.60	6.68	1.74	.48	.19
Speaking	6.16	1.94	6.11	1.93	.94	.05
Writing	7.00	2.06	7.20	1.55	.69	.07
Listening	6.22	1.41	6.06	1.64	.79	.11
Mean	6.60	1.05	6.51	1.22	.73	.09

TABLE 7
Mean Scores of Learners with and Without Extra Lessons in the Low-CLIL Group

	With extra lessons		Without extra lessons		Statistics	
	M	SD	M	SD	*p*	d
Reading	6.56	2.30	6.77	2.11	.44	.09
Speaking	5.13	1.97	5.38	1.53	.51	.14
Writing	6.33	2.30	6.72	2.11	.71	.17
Listening	6.38	1.40	6.54	1.61	.28	.10
Mean	5.90	1.43	6.27	1.23	.73	.28

TABLE 8
Mean Scores of Learners with and without Extra Lessons in the Non-CLIL Group

	With extra lessons		Without extra lessons		Statistics	
	M	SD	M	SD	*p*	*d*
Reading	7.25	1.72	6.09	1.97	.06	.61
Speaking	6.46	2.64	5.65	2.39	.35	.32
Writing	6.95	2.15	5.24	2.60	.04	.68
Listening	6.50	1.47	5.36	1.77	.08	.63
Mean	5.61	1.50	4.74	1.17	.07	.67

the total number of hours was considered ($p = .012$) and also in two of the activities, specifically, in reading and watching TV ($p = .033$). Post hoc analyses (see Appendix D) confirmed the inferiority of the non-CLIL group in the number of hours spent reading in English when compared to the high-CLIL ($p < .000$) and low-CLIL ($p = .028$)

TABLE 9
Distribution of Hours of EE

		High-CLIL		Low-CLIL		Non-CLIL		KW
Hours of EE	Total hours (All students)	M	SD	M	SD	M	SD	*p*
Music	2.3	2.5	1.98	1.9	2.07	2.5	1.90	.059
Videogames	1.1	1.5	1.97	0.9	1.45	0.9	1.28	.158
Internet	1.1	1.3	1.62	1.0	1.44	1.1	1.58	.383
TV	0.9	1.2	1.62	0.9	1.24	0.7	1.64	.033
Reading	0.7	1.0	1.11	0.9	1.50	0.3	.62	.000
Other activities	0.7	1.0	1.83	0.6	1.32	0.4	.98	.276
Total	7.1	8.71	6.35	6.5	6.59	6.2	4.98	.012

groups, and also in the number of hours watching TV when compared to the high-CLIL group ($p < .000$).

As for the impact of this EE on the learners' performance, when all students were considered, positive correlations appeared in two activities, reading and watching TV, and in both cases, the correlations were positive (Appendix E). Specifically, reading showed a weak correlation with writing ($\rho = .150$; $p = .045$) and a weak correlation with the mean ($\rho = .202$; $p = .007$). Watching TV in English also displayed a weak but positive correlation with speaking ($\rho = .159$; $p = .033$) and with the mean score ($\rho = .173$; $p = .020$). Effect sizes were moderate for reading ($\eta^2 = .064$) and small for watching TV ($\eta^2 = .015$). When we examined the impact of informal EE in each of the three groups independently, each group revealed a different scenario (Appendix F). A moderate correlation between listening to music and speaking was found in the high-CLIL group ($\rho = .287$; $p = .027$); a moderate correlation between reading and the skill of writing was found in the low-CLIL group ($\rho = .321$; $p = .016$); and, finally, the largest number of correlations was found in the non-CLIL group: watching TV positively and moderately correlated with speaking ($\rho = .298$; $p = .044$), writing ($\rho = .302$; $p = .042$), and listening ($\rho = .301$; $p = .042$), and surfing the Internet also associated with improvements in the writing skill ($\rho = .310$; $p = .036$).

DISCUSSION

Linguistic Scores

Our objective was to examine the impact of exposure on the proficiency of children receiving different amounts of English inside and outside school. Our results have revealed several advantages in the high-CLIL group. When compared to non-CLIL learners, high-CLIL students fared better in the mean score and in writing. When compared to the low-CLIL group, they showed an advantage in speaking. This superiority aligns well with earlier CLIL studies, as most authors have reported advantages in CLIL groups when compared to their non-CLIL counterparts (see reviews in Dalton-Puffer et al., 2010; Dalton-Puffer & Bauer-Marschallinger, 2023; Goris et al., 2019; Graham et al., 2018; Muñoz, 2015; Pérez Cañado, 2021). In this respect, our study adds to the literature by demonstrating this advantage in primary school, where research is scant and the results have not been very promising, as authors have mostly reported few (Pérez Cañado, 2018; Stotz and Meuter, 2003; Merino & Lasagabaster, 2018) or no advantages in CLIL learners (Gálvez Gómez, 2021; Housen, 2012;

Pladevall-Ballester & Vallbona, 2016; Soto-Corominas et al., 2023). Also, the high-CLIL group tended to have a more homogeneous score distribution, especially when compared to the non-CLIL group. This might indicate that CLIL programs help to promote equity in terms of proficiency, reducing the likelihood that students will fall far behind and that other factors, such as SES, could have a greater influence (Lorenzo et al., 2021; Pérez Cañado, 2018).

Another relevant finding was the similarity between low-CLIL and non-CLIL learners. This finding needs to be highlighted to avoid the well-documented bias in favor of statistically significant findings (Sutton, 2009) and, also, to reveal the importance of the intensity of CLIL exposure. The low-CLIL group, with a low but sustained amount of exposure to CLIL (2–3 weekly hours during 5 school years, amounting to more than 500 CLIL hours), did not seem to have made any significant progress when compared to the non-CLIL group. This coincides with findings in secondary school (Lasagabaster, 2008; Ruiz de Zarobe, 2008, 2010) and adds a new angle to Muñoz's (2015) suggestion that more than 300 extra hours would be necessary for advantages to emerge in CLIL in primary school. Thus, our data point to the ineffectiveness of low-CLIL programs in primary school and help to refine Muñoz's prediction, as the low-CLIL group has shown no clear significant advantages with 500 extra hours of CLIL lessons accrued over a period of 5 school years (vs non-CLIL). In line with this, our results also highlight the fact that the differences between the three groups are few and small, in spite of the rather large amount of exposure to CLIL (1200 h in the high-CLIL and 500 h in the low-CLIL group). Perhaps, the linguistic tests have not captured all the gains; perhaps, other variables have a greater impact on these young learners than intensity; and, finally, perhaps today's EFL lessons have become more effective in promoting language acquisition in young learners, making the addition of a few extra CLIL lessons less impactful. Additionally, EE has now become one of the many factors influencing EFL learning.

When looking at the skills separately, our findings tally with some studies and clash with others. They coincide with Moreno and de Diezmas (2016) in finding no differences in terms of reading. They also align with Moreno and de Diezmas (2016) and Pladevall-Ballester and Vallbona (2016) regarding listening (no differences in primary school were reported), but clash with Stotz and Meuter (2003). The differences in writing between the high-CLIL and the non-CLIL group coincide with Ruiz de Zarobe (2010) and clash with Moreno and de Diezmas (2016). As for speaking, on the one hand, the advantage found in the high-CLIL group over the low-CLIL group reinforces the positive impact of exposure on this skill and aligns with previous

findings (Moreno & de Diezmas, 2016; Pérez Cañado, 2018). On the other hand, the lack of superiority of the high-CLIL group over the non-CLIL group coincides with Stotz and Meuter (2003). Overall, if anything, our results reinforce the lack of agreement as to which skills benefit more and emphasize how difficult it is to establish intensity-based predictions for young learners. The heterogeneity of the findings also suggests that other elements, escaping our control (whether they are internal or external variables), might be exerting a greater influence than intensity. In fact, the reviewed studies present different characteristics and the control of variables is often lacking. As research is so scarce, more studies with greater control of influencing elements will be needed to nail down the interplay of factors affecting L2 acquisition among young populations in school contexts.

Extramural English

In terms of formal EE, taking private lessons was relatively frequent in the three groups, but especially in the low-CLIL group (around 40%). Perhaps, students in the high-CLIL group felt they did not need extra lessons so much because they already had many hours of English in school. In contrast, the students in the other two groups might have felt that extra hours were more necessary, especially in the low-CLIL group, where the demands of being taught in English could have generated a lower self-concept as English learners in the students (Azpilicueta-Martínez & Lázaro-Ibarrola, 2023). When possible associations between extra lessons and scores were considered, a weak but positive one was found in the non-CLIL group between taking extra lessons and getting higher scores in writing. A possible interpretation is that, in this low-input group, the extra input offered by private lessons helped to make a small difference while this was not the case in contexts where school exposure is already abundant. However, we do not know the reasons why the participants were taking private tuition and, therefore, this is mere speculation. Future studies investigating formal EE should ask participants why they are taking extra lessons. Ideally, information on the quality of these lessons would also be very useful.

As for informal exposure, most students frequently came in contact with English outside school and, in line with previous research, high-CLIL learners devoted more hours to EE. When compared to Swedish students (Olsson & Sylvén, 2015; Sundqvist, 2009; Sylvén & Sundqvist, 2015), our learners devoted similar amounts of time to EE (around 7 h per week), although they mainly listened to music (2.3 h per week) while Swedish learners mainly watched TV (2.3 h per week

vs 0.9 h in our study). Also, several positive associations were found. When all the learners were considered, reading and watching TV appeared to be the main activities that positively impacted their scores, reinforcing previous findings (Lindgren & Muñoz, 2013; Muñoz et al., 2018). When the groups were considered separately, several associations were also found with different activities. The associations were always positive and the non-CLIL group was the one that displayed a greater number of associations. This might be the case because, as we argued in the case of extra lessons, in a low-input context, the extra input offered by EE might have an effect that does not apply when school input is already large. In any case, the small but positive influence of these activities needs to be further explored in future studies. At this point, we do not know if the learners who watch TV and read, for instance, do so because their level is better or if their level has improved thanks to the activities. Probably, as Sylvén and Sundqvist (2015) suggest, the EE and the proficiency level are being mutually reinforcing.

CONCLUSION

This study set out to explore the impact of CLIL intensity and EE on three groups of young English learners in primary school.

The results have revealed proficiency advantages in the high-CLIL group in the mean score (vs non-CLIL), in writing (vs non-CLIL), and in speaking (vs low-CLIL) and, also, a more homogeneous distribution of scores. However, there are no differences in the receptive skills or between non-CLIL and low-CLIL learners. As for EE, most of our students engage in informal activities in English (especially in the high-CLIL group), extra lessons are also quite common, and some positive correlations with students' test scores have been found, particularly in the non-CLIL group.

These findings contribute to our understanding of CLIL intensity in different ways. Firstly, the superiority of the high-CLIL group reinforces the notion that greater exposure via CLIL lessons (+1200 h) generates language benefits among primary schoolers. Secondly, our results question the effectiveness of low-intensity CLIL programs, as we found no advantages in low-CLIL learners (vs non-CLIL learners), and confirm and refine Muñoz's (2015) prediction by setting the threshold of intensity somewhere more than 500 h in primary school. In addition, our findings on EE are consistent with previous research (Lindgren & Muñoz, 2013; Muñoz et al., 2018; Olsson & Sylvén, 2015; Sundqvist, 2009; Sylvén & Sundqvist, 2015). They underscore the value of certain activities over others, suggest that EE may be particularly

relevant in low-input contexts (such as non-CLIL environments), and, most importantly, emphasize the need to further investigate EE and incorporate it as a crucial factor in SLA studies (Sundqvist, 2024).

This study has several limitations that open avenues for future research. The selected tests were traditional English assessments focused on language skills, which may not fully capture the distinctive strengths of CLIL students. For instance, the assessments did not measure interaction, specific vocabulary, or genres developed in CLIL lessons, such as explanations.

Additionally, SES and out-of-school exposure were measured using student self-reports. Even though the children could take the questionnaires home and received assistance from their teachers, their young age might have led to less accurate responses. Future research could benefit from using alternative tools, such as regular interviews with students and their families.

Another limitation, common to most research conducted in instructed learning settings, is the fact that numerous contextual variables cannot be controlled (Dalton-Puffer, Llinares, Lorenzo, & Nikula, 2014; Lasagabaster, 2008). For instance, our students came from different schools (including Basque and Spanish schools), introducing variables that may outweigh the impact of intensity, such as the learners' motivation (Lázaro-Ibarrola & Azpilicueta-Martínez, 2024), the quality of classroom interaction (Llinares & Evnitskaya, 2021), the skills of content teachers, and the overall quality of teaching (Goris et al., 2019).

Also, in future studies, more detailed questionnaires are needed to gain a better understanding of EE activities. For formal EE, it would be important to gather information on the characteristics of extra lessons, such as the number of learners, content, and teaching approach. Regarding informal EE activities, more specific information should be collected. For example, it would be necessary to know how children use the Internet (whether they primarily read or interact online), whether they watch TV with subtitles, if they play interactive video games that require communication, or if they understand lyrics when listening to music.

Finally, future research should include more longitudinal studies. These would help determine whether the small advantages gained during primary education are sustained or if students who begin CLIL in secondary education can catch up with those who started earlier. Older children and adolescents are more efficient school language learners and the advantages obtained in primary education may not hold (Lorenzo, Casal, & Moore, 2010).

All in all, our findings broaden our understanding of CLIL programs in primary school and underline the need to broaden our concept of

intensity by taking into account exposure to the target language both within and outside of the classroom. Ultimately, our study constitutes an effort to comprehend the intricacies of child SLA in instructed settings, to prevent generalizing results from secondary school learners, and to call for more research with this sizable population. The value of this study, with its preliminary but promising findings, probably lies in the fact that it raises more questions than it answers.

ACKNOWLEDGMENTS

The author gratefully acknowledges funding from research grant PID2020-113990GB-I00, provided by the Spanish Ministry of Science, Innovation and Universities. I also express my sincere gratitude to the schools, students, and teachers whose voluntary involvement was crucial to this study.

FUNDING INFORMATION

This work was supported by the Spanish MICIU (Ministerio de Ciencia, Innovacion y Universidades) and AEI (Agencia Estatal de Investigacion) (grant number MCIU/AEI/10.13039/501100011033). Open Access funding provided by Universidad Publica de Navarra.

ETHICS STATEMENT

The project under which this study was conducted obtained institutional approval and the research was carried out in accordance with the institution's approved research protocol. All the tests were supervised and approved by the school teachers and administered with their help. Authorized consents were signed by the parents or legal guardians of the participants prior to data collection. As for the children themselves, the teachers explained the objectives of the research to them and they were allowed to decide whether they wanted to participate or not.

THE AUTHOR

Amparo Lázaro-Ibarrola is a Full Professor of English at the Public University of Navarre (Spain) and leads the Navarre Language Acquisition and Teaching (NAVLAT) research group. Her research focuses on the acquisition of second languages by young learners in instructed settings and aims to make SLA knowledge practical for teaching.

DATA AVAILABILITY STATEMENT

Supporting data (questionnaires and tests) is available at the Iris Database https://www.iris-database.org/.

REFERENCES

Agustín-Llach, M. P., & Canga Alonso, A. (2016). Vocabulary growth in young CLIL and traditional EFL learners: Evidence from research and implications for education. *International Journal of Applied Linguistics*, *26*(2), 211–227. https://doi.org/10.1111/ijal.12090

Azpilicueta-Martínez, R., & Lázaro-Ibarrola, A. (2023). Intensity of CLIL exposure and L2 motivation in primary school: Evidence from Spanish EFL learners in non-CLIL, low-CLIL and high-CLIL programmes. *International Review of Applied Linguistics in Language Teaching. Advance online publication.* https://doi.org/10.1515/iral-2022-0239

Butler, Y. G., & Le, V. N. (2018). A longitudinal investigation of parental social-economic status (SES) and young students' learning of English as a foreign language. *System*, *73*, 4–15. https://doi.org/10.1016/j.system.2017.07.005

Cambridge. (2018). English qualifications. Pre A1 starters, A1 movers and A2 Flyers. In Sample papers for young learners. For exams from 2018, 2. Cambridge Assessment English.

Dalton-Puffer, C., & Bauer-Marschallinger, S. (2023). Chapter 11: L2 proficiency and development in CLIL. In Dario Luis Banegas and Sandra Zappa-Hollman (eds.). *The Routledge handbook of content and language integrated Learning.111-126.* https://doi.org/10.4324/9781003173151-11

Dalton-Puffer, C., Llinares, A., Lorenzo, F., & Nikula, T. (2014). "You can stand under my umbrella": Immersion, CLIL and bilingual education. A response to Cenoz, Genesee and Gorter (2013). *Applied Linguistics*, *35*, 213–218. https://doi.org/10.1093/applin/amu010

Dalton-Puffer, C., Nikula, T., & Smit, U. (2010). Language use and language learning in CLIL: Current findings and contentious issues. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 279–291). Philadelphia/Amsterdam: John Benjamins.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (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. https://doi.org/10.1017/S0272263121000097

Gálvez Gómez, M. M. (2021). The effects of CLIL on FL learning: A longitudinal study. In M. L. Pérez-Cañado (Ed.), *Content and language integrated learning in monolingual settings: New insights from the Spanish context* (pp. 141–165). Switzerland: Springer Nature.

Godfroid, A., & Andringa, S. (2023). Uncovering sampling biases, advancing inclusivity, and rethinking theoretical accounts in second language acquisition: Introduction to the special issue SLA for all? *Language Learning*, *73*(4), 981–1002. https://doi.org/10.1111/lang.12620

Goris, J., Denessen, E., & Verhoeven, L. T. (2019). Effects of content and language integrated learning in Europe: A systematic review of longitudinal

experimental studies. *European Educational Research Journal*, *18*(6), 675–698. https://doi.org/10.1177/1474904119872426
Graham, K., Choi, Y., Davoodi, A., Razmeh, S., & Dixon, L. (2018). Language and content outcomes of CLIL and EMI: A systematic review. *Latin American Journal of Content and Language Integrated Learning*, *11*(1), 19–37. https://doi.org/10.5294/laclil.2018.11.1.2
Heaton, J. B. (1966). *Composition through pictures*. United Kingdom: Longman Group.
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). Bristol: Multilingual Matters.
Jiménez Catalán, R. M., & Ruiz de Zarobe, Y. (2009). The receptive vocabulary of EFL learners in two instructional contexts: CLIL versus non–CLIL instruction. In R. Jiménez Catalán & Y. Ruiz de Zarobe (Eds.), *Content and language integrated learning: Evidence from research in Europe* (pp. 81–92). Bristol, UK: Multilingual Matters.
Kuppens, A. H. (2010). Incidental foreign language acquisition from media exposure. *Learning, Media and Technology*, *35*, 65–85. https://doi.org/10.1080/17439880903561876
Lancaster, N. K. (2018). Extramural exposure and language attainment: The examination of input-related variables in CLIL programs. *Porta Linguarum*, *29*, 91–114. https://doi.org/10.30827/Digibug.54024
Lasagabaster, D. (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal*, *1*(1), 30–41. https://doi.org/10.2174/1874913500801010030
Lázaro-Ibarrola, A., & Azpilicueta-Martínez, R 2024). Motivation towards the foreign language (English) and regional language (Basque) in immersion schools: Does CLIL in the foreign language make a difference? *Language Teaching Research, 28(4), 1682-1700.*https://doi.org/10.1177/13621688211031737
Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, *10*(1), 105–129. https://doi.org/10.1080/14790718.2012.679275
Llinares & Evnitskaya. (2021). Classroom interaction in CLIL programs: Offering opportunities or fostering inequalities? *TESOL Quarterly*, *55*(2), 366–397. https://doi.org/10.1002/tesq.607
Lorenzo, F., Casal, S., & Moore, P. (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), 418–442. https://doi.org/10.1093/applin/amp041
Lorenzo, F., Granados, A., & Rico, N. (2021). Equity in bilingual education: Socioeconomic status and content and language lerning in monolingual southern Europe. *Applied Linguistics*, *42*(3), 393–413. https://doi.org/10.1093/applin/amaa037
Mearns, T., de Graaff, R., & Coyle, D. (2020). Motivation for or from bilingual education? A comparative study of learner views in The Netherlands. *International Journal of Bilingual Education and Bilingualism*, *23*(6), 724–737. https://doi.org/10.1080/13670050.2017.1405906
Merino, J. A., & Lasagabaster, D. (2018). The effect of content and language integrated learning programs' intensity on English proficiency: A longitudinal

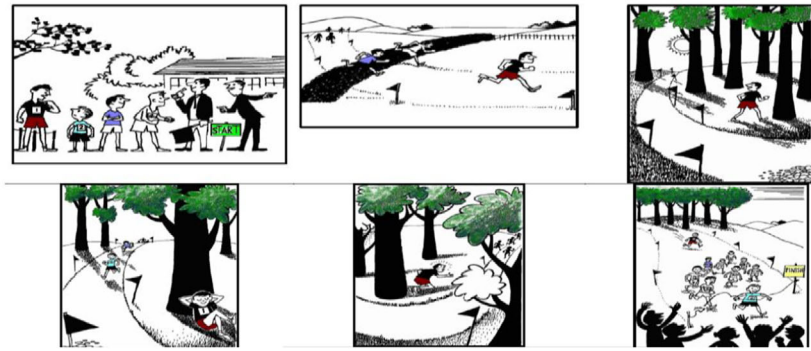
study. *International Journal of Applied Linguistics*, *28*(1), 18–30. https://doi.org/10.1111/ijal.12177
Moreno, N., & de Diezmas, E. (2016). The impact of CLIL on the acquisition of L2 competences and skills in primary education. *International Journal of English Studies*, *16*(2), 81–101. https://doi.org/10.6018/ijes/2016/2/239611
Muñoz, C. (2015). Time and timing in CLIL: A comparative approach to language gains. In M. Juan–Garau & J. Salazar-Noguera (Eds.), *Content–based language learning in multilingual educational environments* (pp. 87–102). Heidelberg, Germany: Springer.
Muñoz, C., Cadierno, T., & Casas, I. (2018). Different starting points for English language learning: A comparative study of Danish and Spanish young learners. *Language Learning*, *68*(4), 1076–1109. https://doi.org/10.1111/lang.12309
Olsson, E., & Sylvén, L. K. (2015). Extramural English and academic vocabulary. A longitudinal study of CLIL and non-CLIL students in Sweden. *Apples-Journal of Applied Language Studies*, *9*(2), 77–103. https://doi.org/10.17011/apples/urn.201512234129
Pérez Cañado, M. L. (2018). CLIL and educational level: A longitudinal study on the impact of CLIL on language outcomes. *Porta Linguarum*, *29*, 51–70. https://doi.org/10.30827/Digibug.54022
Pérez Cañado, M. L. (2021). Inclusion and diversity in bilingual education: A European comparative study. *International Journal of Bilingual Education and Bilingualism*, *26*(9), 1129–1145. https://doi.org/10.1080/13670050.2021.2013770
Peters, E., Noreillie, A., Heylen, K., Bulté, B., & Desmet, P. (2019). The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages. *Language Learning*, *69*(3), 747–782. https://doi.org/10.1111/lang.12351
Pladevall-Ballester, E. (2018). A longitudinal study of primary school EFL learning motivation in CLIL and non–CLIL settings. *Language Teaching Research*, *23*, 765–786. https://doi.org/10.1177/1362168818765877
Pladevall-Ballester, E., & Vallbona, A. (2016). CLIL in minimal input contexts: A longitudinal study of primary school learners' receptive skills. *System*, *58*, 37–48. https://doi.org/10.1016/j.system.2016.02.009
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079
Ruiz de Zarobe, Y. (2008). CLIL and foreign language learning: A longitudinal study in the Basque country. *International CLIL Research Journal*, *1*(1), 60–73.
Ruiz de Zarobe, Y. (2010). Written production and CLIL: An empirical study. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 191–212). Philadelphia/Amsterdam: John Benjamins.
Somers, T., & Llinares, A. (2021). Students' motivation for content and language integrated learning and the role of program intensity. *International Journal of Bilingual Education and Bilingualism*, *24*(6), 839–854. https://doi.org/10.1080/13670050.2018.1517722
Soto-Corominas, A., Roquet, H., & Segura, M. (2023). The effects of CLIL and sources of individual differences on receptive and productive EFL skills at the onset of primary school. *Applied Linguistics*, *45*(2), 364–382. https://doi.org/10.1093/applin/amad031

Stotz, D., & Meuter, T. (2003). Embedded English: Integrating content and language learning in a swiss primary school project. *Le Bulletin Suisse de Linguistique Appliquée, 77*, 83–101.
Sundqvist, P. (2009). Extramural English matters: Out-of-school English and its impact on Swedish ninth graders'oral proficiency and vocabulary. [Unpublished doctoral dissertation]. Karlstad University. Retrieved from https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A275141&dswid=3572
Sundqvist, P. (2024). Extramural English as an individual difference variable in L2 research: Methodology matters. *Annual Review of Applied Linguistics*. https://doi.org/10.1017/S0267190524000072
Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in teaching and learning: From theory and research to practice*. New York: Springer Nature.
Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 435–452). New York: Russell Sage Foundation.
Swain, M. (1981). Time and timing in bilingual education. *Language Learning, 31* (1), 1–15. https://doi.org/10.1111/j.1467-1770.1981.tb01369.x
Sylvén, L. K., & Sundqvist, P. (2015). Extramural English in relation to CLIL: Focus on young language learners in Sweden. In S. Marsh, M. L. Pérez Cañado, & J. Ráez Padilla (Eds.), *CLIL in action: Voices from the classroom* (pp. 47–62). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

APPENDIX A Picture prompts for speaking and writing tests

Prompt for speaking test:

Prompt for written test:



APPENDIX B Statistics: Post hoc tests (Bonferroni)

Mean score (post hoc)	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adjusted Sig.
Non-CLIL-Low-CLIL	5.666	10.367	.547	.585	1.000
Non-CLIL-High-CLIL	23.574	9.685	2.434	.015	.045
Low-CLIL-High-CLIL	17.909	9.125	1.963	.050	.149
Speaking (post hoc)	**Test Statistic**	**Std. Error**	**Std. Test Statistic**	**Sig.**	**Adjusted Sig.**
Low-CLIL-Non-CLIL	−12.828	10.338	−1.241	.215	.644
Low-CLIL-High-CLIL	22.765	9.099	2.502	.012	.037
Non-CLIL-High-CLIL	9.937	9.658	1.029	.304	.911
Writing (post hoc)	**Test Statistic**	**Std. Error**	**Std. Test Statistic**	**Sig.**	**Adjusted Sig.**
Non-CLIL-Low-CLIL	15.318	10.340	1.481	.138	.415
Non-CLIL-High-CLIL	28.534	9.660	2.954	.003	.009
Low-CLIL-High-CLIL	13.216	9.101	1.452	.146	.439

APPENDIX C Statistics: Scores in students with and without extra lessons

High-CLIL group	Reading	Speaking	Writing	Listening	Mean
U de Mann–Whitney	536.500	592.000	563.500	575.000	568.500
Sig. asin. (bilateral)	.484	.941	.692	.790	.735
Low-CLIL group	Reading	Speaking	Writing	Listening	Mean
U de Mann–Whitney	343.000	329.000	335.500	352.000	310.500
Sig. asin. (bilateral)	.602	.446	.515	.711	.286
Non-CLIL group	Reading	Speaking	Writing	Listening	Mean
U de Mann–Whitney	148.000	185.000	141.500	151.500	149.000
Sig. asin. (bilateral)	.069	.350	.048	.081	.073

Non-CLIL group (Spearman's Rho)		Extracurricular lessons
Writing scores	Correlation coefficient	.295*
	Sig. (bilateral)	.046

* The correlation is significant at the 0.01 level (two-tailed).
**The correlation is significant at the 0.05 level (two-tailed).

APPENDIX D Statistics: Extramural English

Reading (post hoc)	Test Statistic	Sig.	Adjusted Sig.
Low-CLIL- High-CLIL	.236	.627	1.000
Non-CLIL- High-CLIL	16.199	.000	.000
Low-CLIL-Non-CLIL	6.741	.009	.028
Watching TV (post hoc)	Test Statistic	Sig.	Adjusted Sig.
Low-CLIL-High-CLIL	.123	.726	1.000
Non-CLIL- High-CLIL	7.543	.006	.018
Low-CLIL-Non-CLIL	5.236	.022	.066

APPENDIX E Associations between EE and scores (all participants)

		EE hours reading	EE hours watching TV	EE hours listening to music	EE hours of Internet	EE hours videogames	EE hours other activities
Reading	ρ	.050	−.003	.088	.012	−.080	.018
	p	.502	.968	.240	.873	.286	.808
Speaking	ρ	.146	.159*	.115	.082	.072	.121
	p	.051	.033	.124	.275	.335	.106
Writing	ρ	.150*	.146*	.048	.098	−.035	−.028
	p	.045	.050	.526	.189	.637	.711
Listening	ρ	.114	.123	.049	−.054	−.008	−.056
	p	.128	.101	.511	.473	.916	.459
Mean	ρ	.202**	.173*	.085	.053	.038	.039
	p	.007	.020	.255	.480	.610	.603

* The correlation is significant at the 0.01 level (two-tailed).
**The correlation is significant at the 0.05 level (two-tailed).

APPENDIX F Associations between EE and scores in each group.

Groups and skills		EE hours reading	EE hours watching TV	EE hours listening to music	EE hours of Internet	EE hours videogames	EE hours other activities
High-CIL group							
Reading	ρ	.033	−.110	.055	−.146	−.203	−.005
	p	.777	.337	.631	.203	.075	.969
Speaking	ρ	.038	.140	.287*	.049	.216	.030
	p	.740	.220	.011	.669	.058	.793
Writing	ρ	−.035	−.040	.077	.016	.056	−.122
	p	.762	.727	.504	.890	.625	.287
Listening	ρ	.009	.062	.136	−.200	−.023	−.188
	p	.936	.592	.235	.078	.844	.100
Mean	ρ	.048	.023	.251*	−.092	.061	−.107
	p	.673	.845	.027	.423	.595	.353
Low-CLIL group							
Reading	ρ	−.005	.007	.040	.071	.070	.101
	p	.970	.957	.771	.605	.609	.461
Speaking	ρ	.262	.104	−.163	−.106	−.065	.112
	p	.052	.445	.229	.436	.632	.413
Writing	ρ	.321*	.178	−.129	−.060	−.183	−.150
	p	.016	.190	.344	.662	.177	.271
Listening	ρ	.095	.034	−.033	−.074	−.059	−.057
	p	.486	.805	.811	.586	.668	.677
Mean	ρ	.117	.093	−.049	.017	.002	.030
	p	.391	.497	.719	.901	.987	.824
Non-CLIL group							
Reading	ρ	.048	.081	.188	.084	−.156	−.166
	p	.752	.591	.212	.581	.300	.270
Speaking	ρ	.167	.298*	.000	.217	−.107	.278
	p	.266	.044	.998	.147	.478	.061
Writing	ρ	.064	.302*	.204	.310*	−.141	.194
	p	.673	.042	.174	.036	.351	.197
Listening	ρ	.258	.301*	.083	.231	.121	.193
	p	.083	.042	.584	.123	.422	.199
Mean	ρ	.197	.310*	.134	.250	−.105	.139
	p	.191	.036	.375	.093	.488	.355

* The correlation is significant at the 0.01 level (two-tailed).
**The correlation is significant at the 0.05 level (two-tailed).