Porta Linguarum 29, enero 2018 51-70

CLIL and Educational Level: A Longitudinal Study on the Impact of CLIL on Language Outcomes

María Luisa Pérez Cañado
University of Jaén

Received: 17 June 2017 / Accepted: 3 October 2017
ISSN: 1697-7467

ABSTRACT: This article reports on a longitudinal study carried out with 1,033 CLIL students and 991 EFL learners in 53 public, private, and charter schools across 12 Spanish provinces into the effects of CLIL on foreign language achievement (grammar, vocabulary, reading, listening, and speaking). The evolution of the bilingual and non-bilingual strands, which were matched on a pre-test in terms of English level, verbal intelligence and motivation, from Primary Education to Compulsory Secondary Education to Baccalaureate is traced through the administration of post- and delayed post-tests. In addition to these intergroup comparisons, intragroup development is also examined to determine the evolution of both the CLIL and the non-CLIL students across educational levels in terms of the linguistic components and skills. Finally, discriminant analyses are performed with all the intervening variables of the study (motivation, verbal intelligence, extramural exposure to English, setting, and socioeconomic status) in order to determine whether CLIL is truly responsible for the differences ascertained or whether other variables account for a greater proportion of the variance.
Keywords: CLIL, longitudinal, language attainment

AICLE y nivel educativo: Un estudio longitudinal sobre los efectos del AICLE en el aprendizaje lingüístico

RESUMEN: Este artículo presenta los resultados de un estudio longitudinal llevado a cabo con 1.033 alumnos AICLE y 991 alumnos de inglés como lengua extranjera en 53 centros públicos, privados y concertados de 12 provincias españolas sobre los efectos del AICLE en el aprendizaje de la lengua extranjera (gramática, vocabulario, comprensión lectora, comprensión oral y producción oral). La evolución del aprendizaje lingüístico de ambos grupos, cuya homogeneidad en nivel de inglés, motivación e inteligencia verbal se había garantizado previamente, se examina de Educación Primaria a Educación Secundaria Obligatoria y a Bachillerato a través de la aplicación de post-tests y pruebas de seguimiento. Además de estas comparaciones intergrupales, también se realizan análisis intragrupales para determinar la evolución lingüística del alumnado bilingüe y no bilingüe a través de los precitados niveles educativos. Por último, se realizan análisis discriminantes con todas las variables intervinientes (motivación, inteligencia verbal, exposición extramural al inglés, ámbito y nivel socioeconómico) para determinar si el AICLE es realmente responsable de las diferencias observadas o si estas se pueden adscribir a otras variables.
Palabras clave: AICLE, longitudinal, aprendizaje lingüístico

1. Introduction

For the past decade, European countries have been stepping up linguistic measures in order to meet the so-called "mother tongue + 2 objective": the mandate established by the Commission of the European Communities (1995) that all European citizens should be proficient in their mother tongue and at least two other foreign languages. One of the most commonly embraced solutions to "transcend the perceived weakness of traditional FL (foreign language) teaching" (Dalton-Puffer, 2011: 185) and thereby meet this ambitious objective has been the introduction of CLIL (Content and Language Integrated Learning)[1] across the continent. Indeed, as Pladevall-Ballester and Vallbona (2016: 37) underscore, "The implementation of CLIL programmes has become commonplace in most European educational systems".

As a consequence of this increased implementation of CLIL, the body of research tapping into its effects has also grown considerably, causing CLIL to become an extremely "prolific phenomenon" (Jäppinen, 2005: 149) (cf. Dalton-Puffer, 2011 and Pérez Cañado, 2012 for an overview of quantitative and qualitative research into the effects of CLIL). However, the overwhelming majority of studies conducted are cross-sectional and lack a longitudinal perspective. This has caused numerous authors, particularly in the past half a decade, to call for increased prominence to be given to longitudinal investigations within the CLIL research agenda (Lasagabaster & Ruiz de Zarobe, 2010; Ruiz de Zarobe, 2011; Bruton, 2011a; Pérez Cañado & Ráez Padilla, 2015; Pérez Cañado, 2016). As Piesche, Jonkmann, Fiege, and Kebler (2016: 109) have put it, "longitudinal studies with pre-, post-, and follow-up assessments are still rare". Furthermore, those which exist focus exclusively on a single educational stage (be it Primary or Secondary Education) and do not examine the effects of CLIL across educational levels.

This is precisely the niche which the present investigation seeks to address. It will report on the results of a longitudinal study into the impact of CLIL on foreign language outcomes across educational levels (Primary, Compulsory Secondary, and non-compulsory Secondary Education), which supersedes many of the lacunae presented by prior investigations. Indeed, it works with one of the largest cohorts in the studies hereto conducted (2,024 students in three monolingual communities of Spain: Andalusia, Extremadura, and the Canary Islands); guarantees the homogeneity of CLIL experimental and non-CLIL control groups; focuses on three different educational levels (Primary and Secondary Education and Baccalaureate); factors in intervening variables pertaining to type of school (public, charter[2], and private); and carries out discriminant analyses to determine which variables are truly responsible for the differences ascertained.

After framing the topic against the backdrop of prior investigations, the article goes on to describe the research design of the study and reports on across-group comparisons of bilingual and non-bilingual streams in terms of English as a Foreign Language (EFL) achievement (grammar, vocabulary, reading, listening, and speaking). The evolution of the

[1] CLIL is defined as "a dual-focussed education approach in which an additional language is used for the learning and teaching of both content and language" (Marsh and Langé, 2000: 2). The emphasis on both teaching and content points to the very hallmark of CLIL: it involves a "two for one" approach (Lyster, 2007: 2), where subject matter teaching is used at least some of the time as a means of increased meaningful exposure to the target language.

[2] Charter schools are state-financed schools, most of which have a religious orientation.

bilingual and non-bilingual strands, which were matched on a pre-test in terms of English level, verbal intelligence and motivation, from Primary Education to Compulsory Secondary Education (CSE) to Baccalaureate is traced through the administration of post- and delayed post-tests. In addition to these intergroup comparisons, intragroup development is also examined to determine the evolution of both the CLIL and the non-CLIL students across educational levels in terms of the linguistic components and skills. Finally, discriminant analyses are performed with all the intervening variables of the study (motivation, verbal intelligence, extramural exposure to English, setting, and socioeconomic status) in order to determine whether CLIL is truly responsible for the differences ascertained or whether other variables account for a greater proportion of the variance.

2. A critical reading of prior research

Despite the substantial number of publications which the increased interest in CLIL has spawned, it is surprising to ascertain that, to date, only a handful have a longitudinal focus. Those which *can* be identified have four main foci, which are precisely the ones around which Wolff (2005) considers CLIL investigations should be articulated: the effects of CLIL on the foreign language, the L1, subject content competence, and motivational aspects. A roughly equal number has focused on Primary or Secondary Education and those which have centered on foreign language (FL) competence have considered both receptive and productive, and oral and written skills, albeit generally not concomitantly. Across-group comparisons have predominated, since very few of the studies in question have also factored in within-cohort development.

The bulk of the longitudinal investigations conducted have revolved around the effects of CLIL on FL competence. An initial landmark study was carried out a decade ago by Admiraal, Westhoff and de Bot (2006) in The Netherlands. They worked with 1,305 Secondary Education students who had received four years of CLIL instruction through English in five Dutch schools and found statistically significant differences in favor of the CLIL experimental group on oral proficiency and reading comprehension, but no differences for receptive vocabulary.

Also in Northern Europe and around the same date, Serra (2007) conducted a longitudinal study in Switzerland, albeit with Primary Education students. She centered on L2 oral production and oral and written comprehension in Italian or Romansch as a second language with German-speaking pupils from grades 1 to 6 and found that the experimental and control groups performed equally well on these aspects of L2 learning.

A similar focus on oral production runs through Ruiz de Zarobe's (2008) longitudinal study with Basque CLIL and non-CLIL groups in the third and fourth year of Compulsory Secondary Education and again in the second year of post-compulsory education. Speech production was assessed in terms of pronunciation, vocabulary, grammar, fluency, and content, with statistically significant differences being detected in favor of the CLIL groups (CLIL and CLIL with extra English literature classes). Both increased exposure and the CLIL program were found to positively impact oral competence skills.

Rallo Fabra and Jacob (2015) also worked with Secondary Education level in a Spanish bilingual community (in this case, the Balearic Islands). However, their results, focused ex-

clusively on fluency and pronunciation within oral production, are not as positive as Ruiz de Zarobe's (2008). Over the course of two years, no statistically significant differences emerged between the CLIL and non-CLIL branches on either fluency or pronunciation, casting doubt upon what can be considered sufficient time for CLIL to have an impact.

Again in a Spanish bilingual community –Catalonia- and once more over the course of two years, Pladevall-Ballester and Vallbona (2016) examined the effects of CLIL on the receptive skills (reading and listening) of Primary school learners in 5th and 6th grade. Intergroup comparisons yielded that, when the number of hours of exposure to the FL (English) was kept constant, non-CLIL learners outstripped their CLIL counterparts on the listening skill, while no significant differences emerged for reading competence. These outcomes were complemented with intragroup analyses, which revealed that significant progress was made in both contexts (CLIL and non-CLIL). The authors consider that more time and exposure to CLIL are perhaps necessary for the positive effects of these programs to be felt.

The final study of this nature which can be found in the specialized literature is also from Spain, albeit conducted in a monolingual community (Andalusia) (Pérez Cañado & Lancaster, 2017). It unfolded over the course of a year and a half with students in 4th grade of CSE and followed them until Baccalaureate. The homogeneity of the CLIL and non-CLIL groups was guaranteed in an initial pre-test and significant differences were found on both oral production and comprehension skills one year later, at the end of CSE, in favor of the CLIL stream. However, in the long run, when these same students were in Baccalaureate, similarly to Pladevall-Ballester and Vallbona (2016), it was productive, as opposed to receptive, skills which were more positively affected by CLIL. The outcomes also provided interesting data on what aspects of oral competence are particularly amenable to being taught through CLIL (e.g., more cognitively complex listening activities) and which need to de developed over a longer time span in order to be significantly improved (e.g., pronunciation and fluency).

Thus, this overview of prior research into the longitudinal effects of CLIL on FL learning allows us to derive two overarching conclusions. The first is that longitudinal investigations have countered some of the recurrent outcomes of most cross-sectional studies, most conspicuously, these fact that it is productive, as opposed to receptive, skills which are most positively affected by CLIL in the long run. This attests to the need to conduct further longitudinal studies, as the long-term effect of CLIL could be very different from the short-term results yielded by cross-sectional research. And the second take-away is that the research carried out thus far presents potentially serious flaws which could compromise the validity of its outcomes. These lacunae are acknowledged by the authors themselves and affect several fronts. To begin with, none of the studies summarized above (except that by Pérez Cañado & Lancaster, 2017) has guaranteed the homogeneity of the samples, which have, furthermore, been numerically limited in most cases. In this sense, Surmont, Struys, Van Den Noort, and Van De Craen (2016: 324) acknowledge the "creaming effect" of selection procedures and Dallinger, Jonkmann, Hollm, and Fiege (2016: 29) explicitly underscore that "selection processes in CLIL-programmes resulted in substantial differences between CLIL- and non-CLIL-classrooms". In addition, none of the studies factor in and control for intervening variables. Finally, no multivariate analyses are carried out which would allow the outcomes to be attributed to CLIL instructional practices: "multivariate analyses (such as factor or discriminant analyses) should be performed in order to determine which variables (…) are truly responsible for the better mathematical results of the CLIL group compared to the non-CLIL group" (Surmont, Struys, Van Den Noort & Van De Craen, 2016: 331).

As Paran (2013: 331) underscores, "we simply do not have enough evidence" and further research into the longitudinal effects of CLIL on FL outcomes is thus fully warranted. This is precisely the remit of the present study, which strives to provide updated empirical evidence on the issue by superseding the main limitations of prior investigations into the topic. It is to its description that we now turn.

3. The study

The present study is framed within a broader research project (cf. Acknowledgements) which has carried out a large-scale evaluation of CLIL programs in three of the monolingual communities in Spain which have the least tradition in bilingual education (Andalusia, Extremadura, and the Canary Islands). Quantitatively, it has studied the effects of CLIL on the English language competence (grammar, vocabulary, and the four skills), Spanish language competence, and content knowledge of Natural Science subjects taught through the foreign language of Primary (6th grade) and Secondary (4th grade) Education students. It has also determined whether such effects pervade one year after CLIL instruction is discontinued, when these same Compulsory Secondary Education students are in the first grade of Baccalaureate. In turn, from a qualitative standpoint, it has probed students', teachers', and parents' satisfaction with all the curricular and organizational aspects of CLIL schemes and carried out a detailed SWOT analysis of the way in which they are functioning, employing questionnaires, semi-structured individual and focus group interviews, and direct behavior observation. This study is inserted within the quantitative side of the investigation and focuses specifically on the effects of CLIL on English as a foreign language competence through the following research questions.

3.1. Research questions

RQ1: Do CLIL programs implemented with Primary and Secondary school students (experimental group) develop superior linguistic competence (grammar, vocabulary, reading, listening, and speaking) to that promoted by EFL programs with students from the same level (control group)? Phrased more simply, is there a linguistic competence differential between CLIL and EFL groups at Primary and Secondary school level in Andalusia, Extremadura, and the Canary Islands?

RQ2: Do the possible differential effects exerted by CLIL programs on English language competence pervade in the first grade of Baccalaureate (six months after the CLIL program is discontinued) or do they gradually peter out?

RQ3: Does the CLIL (experimental) group's linguistic competence significantly improve from the fourth year of CSE to the first year of Baccalaureate?

RQ4: Does the non-CLIL (control) group's linguistic competence significantly improve from the fourth year of CSE to the first year of Baccalaureate?

RQ5: If there is a competence differential between the treatment and comparison groups, is it truly ascribable to language learning based on academic content processing?

3.2. Research design

This quantitative part of the broader study is an instance of applied, primary, quasi-experimental research, with a pre-test/post-test control group design. It meets the four necessary requirements for studies to be methodologically acceptable which Cummins (1999: 27) stipulated for research focusing on the linguistic assessment of content/immersion learners:

1. Studies must compare students in bilingual programs to a control group of similar students.
2. The design must ensure that initial differences between treatment and control groups are controlled statistically.
3. Results must be based on standardized test scores.
4. Differences between the scores of treatment and control groups must be determined by means of appropriate statistical tests.

3.3. Sample

The study has worked with a sample of 2,024 students in 53 public, private, and charter schools in the 12 provinces of three monolingual autonomous communities in Spain: Andalusia, Extremadura, and the Canary Islands. 828 students are finishing 6th grade of Primary Education (ages 11-12) and 1,196 are about to complete 4th grade of Compulsory Secondary Education (ages 15-16). The majority of the cohort (78.3%) studies at public schools where CLIL branches and monolingual EFL streams co-exist. In turn, 17% of the pupils are enrolled in charter non-bilingual schools and the smallest percentage (4.7%) are private bilingual school students. 64% of the schools are located in urban areas, while the remaining 36% are rural. Practically equal percentages are part of CLIL streams (49%) and traditional EFL branches (51%) and there is a perfect balance in terms of gender (1,012 are male students and 1,011 are female).

As Fernández Fontecha (2009) underscores, Spain encompasses a diversity of models practically tantamount to the number of regions where it is applied, given the decentralization of our educational system, which transfers educational powers to each autonomous community. This circumstance, together with large amount of participating schools in the sample, precludes a single blueprint of CLIL implementation across the board. However, certain common features can be distilled vis-à-vis CLIL provision. Bilingual schools must teach from a minimum of 50% of the curriculum of two to four content subjects through CLIL in the first foreign language (which can be English, French, or German). Bilingual branches or sections can co-exist with regular, mainstream groups who only receive input in the target language in FL classes. The CLIL stream must receive daily exposure to the first foreign language in both Primary and Secondary Education. CLIL teaching generally takes place 5 hours per week, compounded with FL instruction (3-4 hours a week), L1 classes (3-5 hours per week), and L3 classes from the second cycle of Primary Education (2-3 hours a week), with a view to developing plurilingual intercultural competence (Jáimez Muñoz, 2007). Depending on the available teachers' profile, each school can determine the subjects taught through the first FL, although at least one must belong to the area of Natural and Social Sciences. The most common ones being implemented via CLIL include Science, Art, and Physical Education at Primary level, and Social and Natural Sciences, Mathematics,

Physical Education, and Technology in Secondary Education.

It is paramount to highlight that the homogeneity of CLIL and non-CLIL learners has been guaranteed from the outset of the study. The level of self-selection in bilingual groups has been a common concern running through the specialized literature. It has often been claimed that bilingual classes normally comprise the more motivated, intelligent, and linguistically proficient students (Bruton, 2011a, 2011b, 2013). In order to ensure that we were working with homogeneous and, thereby, truly comparable groups, the entire first year of the study was devoted to matching students within schools in terms of verbal intelligence, motivation, and level of English to guarantee the homogeneity of the treatment and comparison groups. To this end, initial motivation and verbal intelligence tests were administered and English grades were collected from nearly double the final number of schools who participated in the sample (90) and the CLIL and non-CLIL groups' results were compared. Those schools which evinced the greatest homogeneity were selected to make up the final sample. It was interesting to note that, in the majority of cases, the monolingual and bilingual cohorts evinced no statistically significant differences on the three aspects considered. However, in those cases where differences were ascertained on one or several of the aspects sampled, the so-called outliers (the students with the highest or lowest scores) were eliminated from the sample until no statistically significant differences between the groups emerged. Thus, homogeneity has been guaranteed in our sample, which comprises students with the same verbal intelligence, motivation, and level of English for the sake of comparability.

3.4. Variables

Three types of variables have been taken into account: dependent, independent, and moderating ones.

- The dependent variable is the students' English language (FL) competence (grammar, vocabulary, reading, listening, and speaking).
- In turn, the independent variable corresponds to the CLIL programs implemented in the different types of schools.
- Finally, as moderating variables, the following have been considered:
  - Verbal intelligence
  - Motivation
  - Socioeconomic status (SES)
  - Type of school (public – private – semi-private)
  - Setting (urban – rural)
  - Exposure to English outside school.

3.5. Instruments

Four instruments have been employed for information-gathering: verbal intelligence, motivation, and English language tests. In addition, an initial questionnaire was administered to the students, comprising personal data and information on their parents' age and educational level, which was taken as a proxy for socioeconomic status (SES). All three tests are previously validated and tried-and-tested instruments in the field of psychology or language teaching research.

- The verbal intelligence test was part of the EFAI (*Evaluación Factorial de las Aptitudes Intelectuales*) battery (Santamaría, Arribas, Pereña & Seisdedos, 2014). It has two different versions, adapted to 6th grade of Primary Education and 4th grade of Compulsory Secondary Education. The former version comprises 26 items, while the latter involves 23. In both cases, the students had to choose from four multiple choice options involving analogies, antonyms, or odd-one-out and had five minutes to complete as many items as possible.
- In turn, to measure motivation, Pelechano's (1994) *MA test* was used. This test comprises 35 items and isolates four motivational factors of achievement and anxiety: (i) vain desire to work and self-esteem (containing 10 items); (ii) anxiety in the face of exams (with a negative-inhibitory content and made up of 9 elements); (iii) lack of interest in studying (comprising 9 items); and (iv) realistic personal self-demand (composed of 7 elements).
- The language tests (one for 6th grade of Primary Education and another one for 4th grade of CSE and 1st grade of Baccalaureate) were specifically designed and validated for the study (cf. Madrid, Bueno, & Ráez, in press, for the results on their internal validity and reliability). They comprised use of English, vocabulary, reading, writing[3], and speaking sections with a total score of 100 points. A rubric was designed and validated for the assessment of speaking performance, comprising five main criteria: grammatical accuracy, lexical range, fluency and interaction, pronunciation and task fulfilment (cf. Pérez Cañado & Lancaster, 2017).

3.6. Procedure

The study has spread out over the course of the past four academic years. To begin with, the *Delegación de Educación* and the provincial coordination of bilingual programs was contacted in all three communities in order to request a list of the public schools with English bilingual school programs who possessed the features we targeted in our study (two classes: one mainstream EFL and one CLIL), of private bilingual schools, and of charter schools without CLIL groups in each province. Roughly double the amount of schools who would finally partake in the project were initially selected (90 in all). They were contacted to introduce the project, explain its procedure and benefits, and receive their signed consent to participate in it. The verbal intelligence and motivation tests were applied in each of the schools over the course of an hour at the outset of the academic year 2014-2015, after exactly ten years of CLIL implementation in the autonomous communities in question. Information was also collected on the sociocultural level of the students, their English grades, and their extramural exposure to the language. The tests were corrected and analyzed by a psychologist hired for that purpose and the existence of statistically significant differences across groups was determined. The schools which evinced homogeneity in terms of the variables considered were selected as the final cohort for the study. Finally, at the end of the academic year 2014-2015, when the students were finishing both Primary and Secondary Education, the English language tests were administered over the course of two hours each (one for the written part of the exam and one for the speaking section). Six months later, in December

[3] The results corresponding to the writing skill are not included in this article since they are still under analysis.

2015, the delayed post-test was applied to the same students who were previously in 4th grade of CSE and who were now in 1st grade of Baccalaureate, again over the course of two hours. A single rater was hired for their correction to ensure rater reliability. The analysis of results ensued at the beginning of the academic year 2015-2016.

3.7. Data analysis: Statistical methodology

The data have been analyzed statistically using the SPSS program, in its 21.0 version. In order to guarantee the homogeneity and comparability of the sample, participants have been matched for verbal intelligence, motivation, and English level by calculating the statistical significance of the differences between the experimental (CLIL) and control groups (mainstream EFL) through a one-way repeated measures analysis of variance (ANOVA) and paired samples *t* tests. To address research questions (RQs) 1 through 4, ANOVA and paired samples *t* tests have again been employed to determine the existence of statistically significant differences between and within groups. To calculate effect sizes, Cohen's *d* has been employed using Gpower 3.1. Finally, to respond to RQ 5, successive discriminant analyses have been performed to determine which variable(s) are responsible for the differences between the experimental and control groups.

4. Results and discussion

In order to address the five research questions, results pertaining to intergroup comparisons will be presented initially (RQs 1 and 2), followed by those affecting intragroup analysis (RQs 3 and 4). Finally, the results of the successive discriminant analyses will be rendered in order to determine which variables explain the potential differences found between the experimental and control groups (RQ 5).

4.1. Across-cohort comparison

No statistically significant differences were detected between the treatment and comparison groups at the outset of the academic year in terms of English level (operationalized through the students' grades on this subject), motivation, and verbal intelligence (measured via the corresponding tests). Thus, homogeneity between both cohorts was initially guaranteed. At the end of that academic year (June 2015), the English language post-tests were administered to 6th-grade of Primary Education and 4th-grade of CSE students. For Primary Education, statistically significant differences emerge on all the linguistic components and skills sampled, invariably in favor of the bilingual group. Effect sizes, however, are low for listening, reading, and use of English (cf. Table 1). In fact, if these results are qualified in terms of type of school, no statistically significant differences are detected on listening between public CLIL and non-CLIL branches (p=0.361; d=-0.075), or on use of English (p=0.175; d=-0.120) and listening (p=0.310; d=0.091) between public bilingual and charter non-bilingual strands. Medium effect sizes can be discerned for the remaining aspects sampled (vocabulary and the five subaspects of speaking). It appears that differences between the experimental and control groups are particularly marked for the productive speaking skill at

this point. Thus, at the end of Primary Education, CLIL students already outstrip their EFL counterparts on all the linguistic aspects sampled. Our results consequently run counter to those studies which did not find differences between bilingual and non-bilingual groups at this educational stage (Serra, 2007; Pladevall-Ballester & Vallbona, 2016). They fall in line, however, with other investigations which found that, in the long term, CLIL improved productive more than receptive skills (Admiraal *et al.*, 2006; Pérez Cañado & Lancaster, 2017).

Table 1. FL results for Primary Education

Linguistic aspect	Group	Mean	Standard deviation	Cohen's d	p value
Use of English	Non-CLIL	10.45	6.11	-0.462	<0.001
	CLIL	13.30	6.25		
Vocabulary	Non-CLIL	7.65	3.93	-0.619	<0.001
	CLIL	11.02	6.97		
Listening	Non-CLIL	11.30	2.61	-0.233	<0.001
	CLIL	11.89	2.40		
Reading	Non-CLIL	4.80	3.50	-0.525	<0.001
	CLIL	6.75	3.98		
Speaking (Total)	Non-CLIL	5.43	2.27	-0.858	<0.001
	CLIL	7.42	2.37		
Grammatical accuracy	Non-CLIL	1.01	0.52	-0.727	<0.001
	CLIL	1.42	0.61		
Lexical range	Non-CLIL	1.04	0.53	-0.750	<0.001
	CLIL	1.42	0.49		
Fluency and interaction	Non-CLIL	1.06	0.51	-0.752	<0.001
	CLIL	1.45	0.52		
Pronunciation	Non-CLIL	1.28	0.44	-0.884	<0.001
	CLIL	1.67	0.42		
Task fulfilment	Non-CLIL	1.02	0.43	-0.941	<0.001
	CLIL	1.45	0.47		

What happens at the end of the next main educational stage, namely, Compulsory Secondary Education? After four additional years of participation in CLIL programs, the differences in FL competence are further reinforced, and statistically significant differences invariably emerge in favor of the CLIL cohorts on absolutely all the linguistic aspects sampled, at extremely high confidence levels and with large effect sizes. The latter are particularly considerable for use of English and speaking, especially lexical range and task fulfillment (cf. Table 2). If type of school is again factored in, now public and private bilingual classes outstrip their non-bilingual public and charter peers across the board (cf. Madrid & Barrios, in this volume for a more fine-grained analysis of type of school as an intervening variable). Thus, it clearly transpires that time is a crucial factor to ascertain the effects of CLIL on foreign language attainment; the longer the students have been benefitting from bilingual education, the greater the differences with their non-bilingual counterparts. The impact of CLIL is thus particularly felt in the long term, something which has also been highlighted by authors such as Rallo Fabra and Jacob (2015) or Pladevall-Ballester and Vallbona (2016).

Table 2. FL results for Compulsory Secondary Education

Linguistic aspect	***Group***	***Mean***	***Standard deviation***	***Cohen's d***	***p value***
Use of English	Non-CLIL	19.59	11.02	-1.160	<0.001
	CLIL	31.19	8.99		
Vocabulary	Non-CLIL	7.53	3.71	-0.940	<0.001
	CLIL	10.71	3.06		
Listening	Non-CLIL	3.65	1.74	-0.873	<0.001
	CLIL	5.05	1.46		
Reading	Non-CLIL	2.73	1.82	-0.755	<0.001
	CLIL	4.01	1.57		
Speaking (Total)	Non-CLIL	6.28	2.32	-1.230	<0.001
	CLIL	8.83	1.74		
Grammatical accuracy	Non-CLIL	1.21	0.52	-1.218	<0.001
	CLIL	1.75	0.32		
Lexical range	Non-CLIL	1.21	0.53	-1.442	<0.001
	CLIL	1.83	0.27		
Fluency and interaction	Non-CLIL	1.27	0.55	-1.209	<0.001
	CLIL	1.82	0.30		
Pronunciation	Non-CLIL	1.33	0.41	-1.157	<0.001
	CLIL	1.76	0.29		
Task fulfilment	Non-CLIL	1,250	0,4620	-1,482	<0,001
	CLIL	1,829	0,2844		

Do these effects pervade six months later, when CLIL instruction has been discontinued for the bilingual groups and they are in the first year of non-compulsory Secondary Education? According to our outcomes, they not only pervade, but become even stronger. Indeed, statistically significant differences continue to be discerned in favor of bilingual streams on all the linguistic components and skills sampled, at extremely high confidence levels, and with even larger effect sizes. This is especially the case of speaking (especially, again, lexical range and task fulfillment) and now all the skills, except reading, which has the comparatively lowest (albeit still notable) effect size. Fluency and pronunciation also have large effect sizes, which accords with Pérez Cañado and Lancaster's (2017) finding that these subskills require a longer time span in order to be significantly improved. This outcome is, however, not in harmony with Rallo Fabra and Jacob (2015), who did not detect differences on these aspects between the experimental and control groups, perhaps, as the authors themselves claim, because there was insufficient time in their study for CLIL to have an impact (cf. Table 3). However, at this point, type of school yields interesting results: the non-bilingual charter schools now appear to be catching up with the public and private bilingual ones, as there are no statistically significant differences between them and both these bilingual schools on use of English (p=.536 for the public bilingual and p=.451 for the private bilingual), vocabulary (p=.536 for the public bilingual and p=.0.95 for the private bilingual), listening (p=.575 for the public bilingual and p=.312 for the private bilingual), and reading (p=.199 for the public bilingual and p=.892 for the private bilingual). Thus, the broader take-away here is that the effects of CLIL pervade but are mitigated if these programs are discontinued, so that their maintenance in non-compulsory stages of Secondary Education and even in Tertiary Education should be encouraged in order to maintain the language competence differential.

Table 3. FL results for Baccalaureate

Linguistic aspect	Group	Mean	Standard deviation	Cohen's d	p value
Use of English	Non-CLIL	19.94	9.30	-1.292	<0.001
	CLIL	31.96	9.30		
Vocabulary	Non-CLIL	7.84	2.99	-1.157	<0.001
	CLIL	11.33	3.02		
Listening	Non-CLIL	3,44	1.83	-1.102	<0.001
	CLIL	5.37	1.71		
Reading	Non-CLIL	2.77	1.76	-0.868	<0.001
	CLIL	4.20	1.5		
Speaking (Total)	Non-CLIL	5.900	1.99	-2.671	<0.001
	CLIL	9.378	0.88		

Table 3. FL results for Baccalaureate (Continuation)

Linguistic aspect	Group	Mean	Standard deviation	Cohen's d	p value
Grammatical accuracy	Non-CLIL	1.10	0.47	-2.204	<0.001
	CLIL	1.83	0.25		
Lexical range	Non-CLIL	1.12	0.45	-2.626	<0.001
	CLIL	1.91	0.21		
Fluency and interaction	Non-CLIL	1.20	0.49	-2.130	<0.001
	CLIL	1.89	0.22		
Pronunciation	Non-CLIL	1.27	0.34	-2.018	<0.001
	CLIL	1.82	0.24		
Task fulfilment	Non-CLIL	1.20	0.41	-2.395	<0.001
	CLIL	1.89	0.22		

4.2. Within-cohort comparison

These across-group comparisons are complemented with intragroup analyses in order to determine whether the experimental and control groups' linguistic competence significantly improves from the end of CSE to the first year of Baccalaureate. Our outcomes indicate that both groups have significantly improved in the overall language test, something which accords with Padevall-Ballester and Vallbona's (2016) results (cf. Figure 1). However, effect sizes are low for both groups, perhaps because only six months elapsed between the post- and delayed post-testing phases. No significant headway is made by either of the groups on reading, and the non-bilingual cohort does not advance on listening either, while the same occurs for use of English in the bilingual stream's case (cf. Table 4). This falls in line with Lancaster's (in press) findings, where the CLIL and EFL groups did not significantly improve on listening. Thus, receptive skills once again come across as those where the least long-term progress is made, as has occurred in the intergroup comparisons[4]. If qualified by type of school, interesting outcomes emerge: public bilingual and non-bilingual strands significantly improve between both testing phases, whereas the charter non-CLIL and private CLIL schools do not significantly ameliorate overall (cf. Table 5).

[4] The outcomes for speaking and writing are not presented, as the results corresponding to the delayed post-test for these two skills are still in the process of being corrected and analyzed.

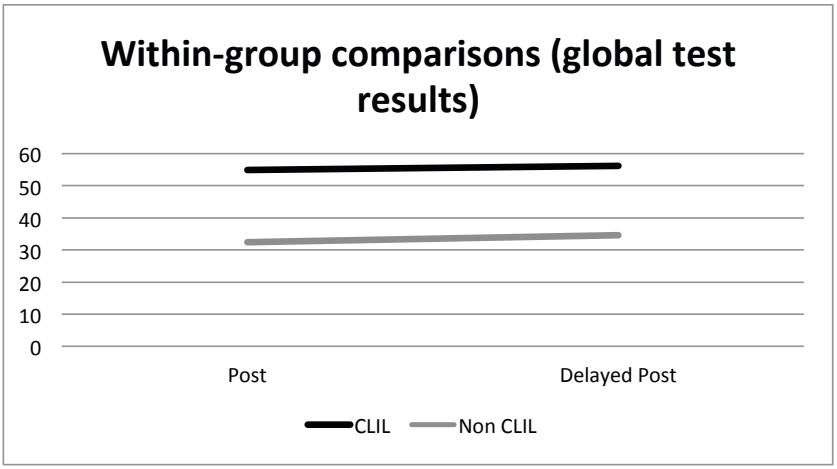

Figure 1. Within-group comparisons on the global test results

Table 4. Evolution of the control and experimental groups from the post- to the delayed post-tests

Group	***Linguistic spect***	***Test***	***Mean***	***Standard deviation***	***Cohen's d***	***p value***
CLIL	Use of English	Post	34.02	7.38	-0.038	0.302
		Delayed post	34.30	7.38		
	Vocabulary	Post	11.29	2.83	-0.194	0.001
		Delayed post	11.83	2.71		
	Listening	Post	5.27	1.42	-0.202	0.007
		Delayed post	5.58	1.63		
	Reading	Post	4.33	1.53	-0.053	0.431
		Delayed post	4.41	1.48		
	Total	Post	54.91	11.04	-0.110	0.002
		Delayed post	56.12	10.87		
Non-CLIL	Use of English	Post	19.09	9.49	-0.138	0.027
		Delayed post	20.36	8.92		
	Vocabulary	Post	7.38	2.88	-0.217	0.013
		Delayed post	8.01	2.92		
	Listening	Post	3.32	1.65	0.011	0.886
		Delayed post	3.30	1.90		
	Reading	Post	2.64	1.60	-0.183	0.090
		Delayed post	2.95	1.78		
	Total	Post	32.41	13.23	-0.166	0.007
		Delayed post	34.60	13.18		

Table 5. Evolution of the different types of schools from the post- to the delayed post-tests

School	*Linguistic aspect*	*Test*	*Mean*	*Standard deviation*	*Cohen's d*	*p value*
Public bilingual	Total	Post	54.12	11.25	-0.112	0.004
		Delayed post	55.38	11.23		
Public non- bilingual	Total	Post	30.90	11.54	-0.205	0.004
		Delayed post	33.32	12.03		
Private bilingual	Total	Post	59.97	7.96	-0.122	0.367
		Delayed post	60.86	6.62		
Charter non-biligual	Total	Post	62.60	7.82	0.293	0.178
		Delayed post	60.20	8.52		

4.3. Explaining the language competence differential

Finally, in order to determine which variables best explain the differences discerned between bilingual and non-bilingual groups on FL competence, successive discriminant analyses have been performed. With this statistical technique, we have strived to assess the discriminating potential of the different variables (independent and moderating) with which we have carried out our investigation in the bilingual and non-bilingual groups. Since our objective is to isolate those variables which best discriminate between both groups, we have performed successive discriminant analyses in which we have selected the variables which display the greatest significance in the tests of equality of group means.

These analyses have allowed us to ascertain that the differences detected in linguistic competence between the bilingual and non-bilingual groups can be ascribed to the independent variable (the CLIL program), especially in the long term. The higher the educational level, the greater the weight which this variable has in explaining the differences between the experimental and control groups. This thus confirms that CLIL programs have a more powerful effect on language attainment particularly in the long run. Indeed, at the end of Primary Education, the CLIL program does not discriminate much between the groups. SES, rural-urban setting, and motivation do not have much weight either. At the end of CSE, however, the independent variable has greater significance in explaining the differences between the groups. SES, verbal intelligence, and motivation now also carry greater weight. Finally, in Baccalaureate, it is patent that the bilingual program is the variable with the greatest weight in explaining the differences between the treatment and comparison groups (cf. Tables 6 and 7).

Table 6. Test of equality of group means

PRIMARY EDUCATION	Wilks' Lambda	F	df1	df2	Sig.
Use of English	0.948	7.201	1	132	0.008
Vocabulary	0.912	12.679	1	132	0.001
Reading	0.904	13.996	1	132	0.000
Grammatical accuracy	0.882	17.605	1	132	0.000
Lexica range	0.876	18.700	1	132	0.000
Fluency and interaction	0.875	18.807	1	132	0.000
Pronunciation	0.835	26.037	1	132	0.000
Task fulfilment	0.818	29.348	1	132	0.000
COMPULSORY SECONDARY EDUCATION	**Wilks' Lambda**	**F**	**df1**	**df2**	**Sig.**
SES	0.934	15.813	1	224	0.000
Verbal intelligence	0.968	7.393	1	224	0.007
Lack of interest	0.963	8.497	1	224	0.004
Use of English	0.754	72.889	1	224	0.000
Vocabulary	0.835	44.292	1	224	0.000
Listening	0.879	30.740	1	224	0.000
Reading	0.830	45.980	1	224	0.000
Grammatical accuracy	0.729	83.171	1	224	0.000
Lexical range	0.656	117.512	1	224	0.000
Fluency and interaction	0.738	79.351	1	224	0.000
Pronunciation	0.749	75.173	1	224	0.000
Task fulfilment	0.652	119.581	1	224	0.000
BACCALAUREATE	**Wilks' Lambda**	**F**	**df1**	**df2**	**Sig.**
Will	0.919	5.657	1	64	0.020
Lack of interest	0.912	6.206	1	64	0.015
Use of English	0.626	38.240	1	64	0.000
Vocabulary	0.734	23.215	1	64	0.000
Listening	0.778	18.230	1	64	0.000
Reading	0.877	8.996	1	64	0.004
Grammatical accuracy	0.478	69.886	1	64	0.000
Lexica range	0.401	95.414	1	64	0.000
Fluency and interaction	0.510	61.610	1	64	0.000
Pronunciation	0.539	54.663	1	64	0.000
Task fulfilment	0.436	82.766	1	64	0.000

Table 7. Summary of canonical discriminant functions

PRIMARY EDUCATION				
Function	Eigenvalue	% of variance	Cumulative %	Canonical correlation
1	0.332	100.0	100.0	0.499
Test of function	Wilks' Lambda	Chi-square	df	Sig.
1	0.751	36.652	8	0.000
COMPULSORY SECONDARY EDUCATION				
Function	Eigenvalue	% of variance	Cumulative %	Canonical correlation
1	0.664	100,.0	100.0	0.632
Test of function	Wilks' Lambda	Chi-square	df	Sig.
1	0.601	111.043	12	0.000
BACCALAUREATE				
Function	Eigenvalue	% of variance	Cumulative %	Canonical correlation
1	1.867	100.0	100.0	0.807
Test of function	Wilks' Lambda	Chi-square	df	Sig.
1	0.349	61.625	11	0.000

5. Conclusion

The present study has allowed us to provide updated empirical evidence on the effects of CLIL programs on the foreign language competence of students across three different educational levels: Primary Education, Compulsory Secondary Education, and Baccalaureate. It has strived to overcome the main lacunae presented by prior investigations into the topic in terms of sample size, homogeneity, variables, or statistical analysis.

Vis-à-vis the first RQ, our outcomes allow us to firmly state that there is indeed a linguistic competence differential between CLIL and EFL groups, in favor of the former, already at the end of Primary Education (albeit less so for the receptive skills of reading and listening), but even more markedly so at the end of CSE. Indeed, confidence levels of statistical significance and effect sizes are considerably greater in this second educational

stage, where bilingual students invariably outstrip their non-bilingual counterparts on absolutely all the linguistic aspects sampled.

In line with RQ2, these differential effects of CLIL programs on English language competence are sustained in the first year of Baccalaureate, six months after the CLIL program has been discontinued for the bilingual group, as the differences between the experimental and control groups are even more pronounced than at the end of CSE. However, non-bilingual charter school students now present no statistically significant differences with the bilingual private and public school learners, something which points to the desirability of maintaining CLIL programs for the language competence differential to be sustained.

The third and fourth RQs have allowed us to observe that both CLIL and non-CLIL strands significantly improve their overall linguistic performance from the post- to the delayed post-tests, although with low effect sizes, something understandable since only six months have elapsed between both testing phases. No improvement is documented, however, for the receptive skills in either of the groups. Finally, only public schools significantly ameliorate their linguistic attainment from CSE to Baccalaureate, as opposed to charter and private centers, where no significant overall improvement is ascertained.

Finally, as regards the fifth and final RQ, the successive discriminant analyses performed have allowed us to ascertain that CLIL programs are the variable which best explains the differences detected, especially as we advance in educational level.

Thus, an important implication accruing from these findings is that time is needed for the full effect of CLIL to be felt on foreign language attainment, something in line with Hughes' (2010) assertion that these types of programs require approximately 20 years to come to fruition. It is productive (especially speaking and, within it, fluency and task fulfillment), as opposed to receptive (particularly reading and listening) which are especially impacted by bilingual education approaches, although absolutely all the linguistic components and skills are positively affected by the development of CLIL programs, especially in the long term.

It is furthermore CLIL –and not any other co-variate- which is responsible for the linguistic competence differential, so its continued implementation would undoubtedly be recommendable, according to our results. Further longitudinal investigations would also be desirable into the effects of CLIL on language competence, L1 development, and content subject mastery in order to determine the exact amount of time required for a success-prone implementation of these types of programs. It is empirical data such as those provided by this study which will allow us to determine whether, when, how, and under what conditions CLIL is truly effective and to ensure that we keep its implementation on track.

6. References

Admiraal, W., Westhoff, G. and de Bot, K. (2006). "Evaluation of bilingual secondary education in The Netherlands: Students' language proficiency", in *English Educational Research and Evaluation,* 12, 1: 75-93.

Bruton, A. (2011a). "Are the differences between CLIL and non-CLIL groups in Andalusia due to CLIL? A reply to Lorenzo, Casal and Moore (2010)", in *Applied Linguistics,* 2011: 1–7.

Bruton, A. (2011b). "Is CLIL so beneficial, or just selective? Re-evaluating some of the research", in *System*, 39, 523–532.

Bruton, A. (2013). "CLIL: Some of the reasons why... and why not", in *System* 41: 587–597.
Commission of the European Communities. (1995). *White paper on education and training. Teaching and learning. Towards the Learning Society*. Available from: http://ec.europa.eu/white-papers/index_en.htm#block_13, accessed 10 May, 2017.
Cummins, J. (1999). "Alternative paradigms in bilingual education research: Does theory have a place?", in *Educational Researcher,* 28, 7: 26–32.
Dallinger, S., Jonkmann, K., Hollm, J. and Fiege, C. (2016). "The effect of content and language integrated learning on students' English and history competences: Killing two birds with one stone?", in *Learning and Instruction*, 41: 23-31.
Dalton-Puffer, C. (2011). "Content-and-Language Integrated Learning: From practice to principles?", in *Annual Review of Applied Linguistics*, 31: 182–204.
Fernández Fontecha, A. (2009). "Spanish CLIL: Research and official actions", in Y. Ruiz de Zarobe and R. M. Jiménez Catalán (eds.), *Content and Language Integrated Learning. Evidence from research in Europe*. Bristol: Multilingual Matters, 3-21.
Hughes, S. (2010). "The effectiveness of bilingual education: A case study". Paper presented at the *25th GRETA Convention: Celebrating 25 Years of Teacher Inspiration.* Granada: University of Granada.
Jáimez Muñoz, S. (2007). "Glossary related to the Plurilingualism Promotion Plan: A language policy for Andalusia", in *GRETA. Revista para Profesores de Inglés,* 15, 1and 2: 67-79.
Jäppinen, A. K. (2005). "Thinking and content learning of Mathematics and Science as cognitional development in Content and Language Integrated Learning (CLIL): Teaching through a foreign language in Finland", in *Language and Education*, 19, 2: 147-168.
Lasagabaster, D. and Ruiz de Zarobe, Y. (2010). "Ways forward in CLIL: Provision issues and future planning", in D. Lasagabaster and Y. Ruiz de Zarobe (eds.), *CLIL in Spain: Implementation, results and teacher training*. Newcastle upon Tyne: Cambridge Scholars Publishing, 278–295.
Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam: John Benjamins Publishing Company.
Madrid, D. and Barrios, E. (2018). "CLIL across diverse educational settings: Comparing programmes and school types", in *Porta Linguarum* 29, 29-50.
Madrid, D., Bueno, A., and Ráez, J. (In press). "Investigating the effects of CLIL on language attainment: Instrument design and validation", in M. L. Pérez Cañado (ed.), *Content and Language lntegrated Learning in monolingual settings: New lnsights from the Spanish context*. Amsterdam: Springer.
Marsh, D. and Langé, G. (eds.). (2000). *Using languages to learn and learning to use languages*. Finland: University of Jyväskylä.
Paran, A. (2013). "Content and language integrated learning: Panacea or policy borrowing myth?", in *Applied Linguistics Review*, 4, 2: 317–342.
Pelechano, V. (1994). "Prueba MA", in *Análisis y Modificación de la Conducta*, 20: 71-72.
Pérez Cañado, M. L. (2012). "CLIL research in Europe: Past, present, and future", in *International Journal of Bilingual Education and Bilingualism,* 15, 3: 315–341.
Pérez Cañado, M. L. (2016). "Stopping the "pendulum effect" in CLIL research: Finding the balance between Pollyanna and Scrooge", in *Applied Linguistics Review,* DOI: 10.1515/applirev-2016-2001.
Pérez Cañado, M. L. and Lancaster, N. K. (2017). "The effects of CLIL on oral comprehension and production: A longitudinal case study" in *Language, Culture, and Curriculum* 30, 3: 300-316.

Pérez Cañado, M.L. and Ráez Padilla, J. (2015). "Introduction and overview", in D. Marsh, M .L. Pérez Cañado and J. Ráez Padilla (eds.), *CLIL in action: Voices from the classroom*. Newcastle upon Tyne: Cambridge Scholars Publishing, 1-12.

Piesche, N., Jonkmann, K., Fiege, C. and Kebler, J. (2016). "CLIL for all? A randomised controlled field experiment with sixth-grade students on the effects of content and language integrated science learning", in *Learning and Instruction*, 44: 108-116.

Pladevall-Ballester, E. and Vallbona, A. (2016). "CLIL in minimal input contexts: A longitudinal study of primary school learners' receptive skills", in *System*, 58: 37-48.

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

Ruiz de Zarobe, Y. (2008). "CLIL and foreign language learning: A longitudinal study in the Basque Country", in *International CLIL Research Journal* 1, 1: 60-73.

Ruiz de Zarobe, Y. (2011). "Which language competencies benefit from CLIL? An insight into Applied Linguistics research", in Y. Ruiz de Zarobe, J. M. Sierra, and F. Gallardo del Puerto (eds*.), Content and foreign language integrated learning. Contributions to multilingualism in European contexts*. Frankfurt-am-Main: Peter Lang, 129-153.

Santamaría, P. Arribas, D., Pereña, J. and Seisdedos, N. (2016). *EFAI. Evaluación Factorial de las Aptitudes Intelectuales.* Madrid: TEA Ediciones.

Serra, C. (2007). "Assessing CLIL at primary school: A longitudinal study", in *International Journal of Bilingual Education and Bilingualism*, 10, 5: 582-602.

Surmont, J. Struys, E., Van Den Noort, M. and Van De Craen, P. (2016). "The effects of CLIL on mathematical content learning: A longitudinal study", in *Studies in Second Language Learning and Teaching*, 6, 2: 319-337.

Wolff, D. (2005). "Approaching CLIL", in D. Marsh (coord.), *The CLIL quality matrix. Central workshop report*. Available from: http://www.ecml.at/mtp2/CLILmatrix/pdf/wsrepD3E2005_6.pdf, accessed 10 May, 2017.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness, under Grant FFI2012-32221, and by the Junta de Andalucía, under Grant P12-HUM-23480. We would also like to thank the school management and the students who participated in the study.