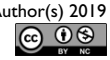
General submission

EERJ European Educational Research Journal

European Educational Research Journal
2019, Vol. 18(6) 675–698



DOI: 10.1177/1474904119872426
journals.sagepub.com/home/eer


Effects of content and language integrated learning in Europe A systematic review of longitudinal experimental studies

JA Goris, **EJPG Denessen**, **LTW Verhoeven**
Behavioural Science Institute, Radboud University Nijmegen, The Netherlands

Abstract

Content and language integrated learning (CLIL), an educational approach in which subject matter and a foreign language – predominantly English – are taught and learnt side by side, has developed into a very popular educational innovation in most European countries. A host of research studies have shown its benefits, and discuss favourable effects especially with respect to L2 gains. However, critical voices have underscored the fact that CLIL attracts or selects mainly high-achieving learners. Hence, the question arises whether it is justified to attribute improved L2 performance mainly to the CLIL intervention, or to favourable learner characteristics. Several reviews of literature were published in the past, but due to a lack of longitudinal findings no conclusive evidence about the added value of CLIL in the process of L2 learning could be produced. The present review aims to fill this void and has undertaken a search of two decades of longitudinal studies into the effects of CLIL on various linguistic skills in the field of English as a foreign language. The findings indicate that robust studies were undertaken in only a limited number of European countries, and that only a few of them were large scale. Yet, the conclusions provide clear indications regarding the contexts in which CLIL leads to significantly better L2 results.

Keywords

Selective content and language integrated learning, longitudinal effects, English as a foreign language gains, literature review, European content and language integrated learning

Introduction

Over the past few decades an increasing number of schools in almost all European countries have adopted innovative educational approaches aimed at better preparing young people for the

Corresponding author:
JA Goris, Behavioural Science Institute, Radboud University Nijmegen, Stelleweg 125, Bergen op Zoom, 4617LV, The Netherlands.
Email: j.goris@pwo.ru.nl

increasingly internationalized world of the 21st century. One innovation in particular – content and language integrated learning (CLIL) – has been met with enthusiasm by teachers, parents and students alike, and has proven particularly successful. The CLIL approach has become a driving force in various types of mainstream education, mostly at the secondary level but also increasingly in primary and vocational schools. Its dual focus aims to develop proficiency in a curriculum subject together with the language through which it is taught (the target language) – nowadays almost invariably English as a foreign language (EFL), including in the case of the learners under discussion in the present article. Achieving this twofold goal requires an integrated approach to instruction and learning practice. In CLIL education the subject teacher is required to adapt didactics in order to make both content and the target language (L2) comprehensible, so that learning takes place in an interconnected way.

The effectiveness of CLIL has quickly become a focus of academic studies, primarily with respect to foreign language learning. Research outcomes in this field are generally positive and indicate higher L2 levels for CLIL tracks compared with conventional language classes (Dalton-Puffer, 2008; Lorenzo et al., 2010). Several studies have been conducted to determine the causes of the improved L2 performance in CLIL environments. Wolff (2007) mentioned the longer period of exposure, better learning conditions owing to authentic lesson materials, the presence of native speakers, extra EFL lessons and the generally richer linguistic content of the CLIL classes. Dalton-Puffer (2007) found the CLIL classroom to provide a real-life situation, in which the foreign language is put to real use which offers possibilities to process information more deeply.

Later studies have shown that CLIL learners are better L2 performers, more motivated, and more linguistically and academically talented: it can be safely stated that CLIL selects or attracts high-achieving students (Bruton, 2011; Küppers and Trautmann, 2013; Paran; 2013). However, this provides little or no information about the effectiveness of the intervention itself. The great majority of research studies in this field have investigated the benefits of CLIL by focusing on only one measure at a certain point, often showing results in favour of CLIL, but a longer-term perspective is missing. There is a lack of studies that follow the development of EFL skills of learners in a CLIL class over a certain period of time and compare the progress made with that of a compatible mainstream, non-CLIL class, so that increased EFL performance can be directly attributed to the CLIL curriculum.

In recent years several reviews of research have been published focusing on the added value of CLIL for various categories of educational practice and motivational outcomes (Dalton-Puffer, 2008, 2011; Pérez-Cañado, 2012; Ruiz de Zarobe, 2011). Even though findings have been mainly positive, the specific contribution of CLIL interventions to the development of EFL skills remains unclear, because only a few longitudinal studies could be discussed. This scarcity of solid longitudinal findings has evoked criticism, and the CLIL literature has pronounced a need for robust outcomes regarding CLIL effects on L2 learning in order to evaluate the merits of the approach. A thorough systematic review of experimental longitudinal research that relates the L2 outcomes directly to CLIL practice is lacking to date. Therefore, the present review evaluates the longitudinal studies published since the introduction of CLIL with regards to the effects on EFL skills of the various CLIL approaches at primary and secondary education levels in various European settings.

CLIL in European contexts

Even though the teaching of content through an L2 is not a new phenomenon or specific to a certain region, CLIL is very much associated with the European context of mainstream compulsory education at the secondary level and, more recently, also in primary and vocational schools. The introduction of the term ‘CLIL’ in Europe took place in 1994, with the approach being described as a dual-focus methodology in which content and language are learned together in an integrated way

(Marsh et al., 2001). Its roots are in the Canadian immersion approach, in which the curriculum is taught in both the L1 and L2 of the state (Westhoff, 1994).

The European region's preparedness for CLIL was triggered by a general dissatisfaction with existing L2 learning practices around the 1980s and 1990s: they were considered ineffective, especially from a communicative point of view. The ongoing European integration and the resulting need for L2 competencies (Marsh, 2002; Oonk, 2004) both in daily life and the workplace (Wolff, 2007) contributed to the appreciation of CLIL, and may have raised awareness among young people that language skills are valuable in an internationalized world. The special Eurobarometer published by the European Commission in 2006 indicated that 68% of the respondents regarded English, the world's lingua franca, as the most important foreign language, and this is reflected in present-day CLIL programmes. Even though the medium of instruction could in principle be any foreign language, the great majority of European CLIL programmes make use of English as a target language.

It is hard to summarize the term CLIL in terms of one single concept. Almost all European countries have introduced CLIL in some form, but there are different backgrounds to its origins and implementation in different nations. As will be described in the results section of this article, each country has its own educational context and different possibilities for implementation, which is displayed in the divergence of CLIL practice. Education is a complex and multi-faceted process, defined by local systems. Theories of second language acquisition (Ellis, 1985; Krashen, 1982) describe L2 development as a complex phenomenon in which progress cannot exclusively be attributed to an educational intervention: acquiring L2 competence is seen as a dynamic process within which a variety of factors interact and whose development is not always linear. The situation is no different for CLIL. Its defining characteristics have been summarized by Coyle et al. (2010) as a synthesis of good teaching practice in which meaningful language and relevant content are integrated, providing scaffolds for progression in both, with the incorporation of cultural awareness and intercultural understanding.

Alongside the positive effects of the CLIL approach, critical voices have also been heard, mainly those of Bruton (2011, 2015) and Paran (2013), who see the CLIL phenomenon as glamorized and place shortcomings in research designs and outcomes under scrutiny. Notably, the failure to match experimental and control groups in terms of language level and aptitude has evoked such criticism. In the specific case of CLIL the potential learners are often high performers, in academic as well as linguistic fields. At the start of secondary education some form of differentiation or streaming in accordance with individual learner characteristics, such as verbal intelligence and academic ability, is common practice in many European countries (Eurydice, 2005). However, for those who wish to study in a CLIL class, additional criteria apply. Secondary schools generally make use of selection criteria for admittance to a CLIL class, involving enhanced motivation and above-average EFL skills, while in some countries prospective learners receive extra EFL training in preparation for joining CLIL streams (Goris et al., 2013). This may lead to significant discrepancies in initial EFL levels between experimental and same-cohort mainstream groups, a phenomenon that seems to be treated as an intrinsic design problem in CLIL research. It has been acknowledged that securing homogeneity can be a difficult task, for example by Lasagabaster (2008), who discusses the fact that some variables are hard to control in a study carried out in an authentic educational setting, often by individuals and with limited means. In the results section of this article we will discuss how the authors of the studies reviewed here have dealt with the issue of matching experimental and control groups.

Previous reviews of CLIL effects on linguistic gains

In recent years, several inventories of outcome-based classroom studies have been presented. No unanimity exists with regard to the linguistic competencies that benefit from CLIL. Dalton-Puffer (2007, 2008, 2009) focuses on L2 gains in divergent European contexts, mostly German-speaking

countries, and reports positive effects for receptive language skills, vocabulary and morphology, as well as creativity, risk-taking, fluency and speaking confidence. Positive but different findings – for example, with regard to writing skills – have been reported in Spain by Ruiz de Zarobe (2011), whose work contradicts previous studies in some respects. The present review aims to compare and contrast research conditions and CLIL contexts, and to present more information about how linguistic gains have developed in the divergent educational contexts of different European countries that have applied a CLIL approach over the last few decades.

A similar goal was pursued in a meta-analysis conducted by Pérez-Cañado (2012), who presented a comprehensive, updated and critical review of the way in which the new educational approach of CLIL has been playing out across Europe. The survey pertained to countries in northern, central, eastern and southern Europe and discussed a broad field of cognitive, educational and affective variables affected by CLIL education, such as motivation, teaching practice, content learning and linguistic merits, mainly concerning various fields of EFL competence. The results unquestionably indicate that CLIL affects L2 language learning outcomes; however, consideration of longitudinal studies is scarce and moreover, no detailed discussion of educational findings is presented. The conclusions (Pérez-Cañado, 2012: 329) state that the last two decades have mainly seen an increasing number of studies of a descriptive nature, focusing on the benefits of bilingual education, whereas solid empirical studies have been sparse. The author also underscores the lack of robust findings resulting from a pre-test–post-test design.

Since Pérez-Cañado's review, years in which CLIL has continued to become a more prominent feature of 21st-century education and place itself in the spotlight in terms of both research into CLIL effects and criticism of its approach. It is generally acknowledged that CLIL is promotive to L2 skills; however, its selectiveness, the ways in which experimental and control groups are compared, and an absence of pre-tests have been criticized as factors biasing research outcomes. The present review aims to evaluate the longitudinal studies into CLIL effects on EFL progress that have been published since the early days of CLIL, present a detailed overview of the various linguistic skills that have been tested and of the findings, and compare variation in terms of learners and educational contexts. We also seek to explore how research outcomes have developed over time, and to discuss how topical CLIL issues such as initial differences and experimental–control group matching have been dealt with by the authors of these existing works.

The present review

As we have discussed above the contexts in which European CLIL have taken shape are very much divergent, yet their goals converge: to provide better L2 learning conditions, either for a select group or for all. Our aim is to come to an evaluation of the contribution of CLIL to competencies in the most used target language, English, of mainstream primary and secondary school learners during the past 20 years of CLIL existence. To be more precise, the authors aim to provide a systematic review of longitudinal experimental research analysing the development of EFL proficiency skills of a certain cohort of learners over time and comparing results of CLIL groups with those of control groups in traditional classes.

Method

Data collection

A systematic search for articles analysing the effects of CLIL education on EFL skills was conducted by the first author together with a university librarian in February 2018. Our main search

principle was that L2 proficiency is an integration of knowledge of words, expressions, insight into the rules of the language and an ability to understand its written texts, which we defined as the research focus of the studies we were looking for. Therefore we combined the keyword 'CLIL' and its earlier term 'bilingual education' with the search words 'vocabulary, grammar, idioms and text comprehension' to find articles in Europe in the databases ERIC and PsycInfo. In total, 235 articles were found in the ERIC database and 172 in that of PsycInfo. This number was considered sufficient as a starting point for further selection. The screening of articles took place in several phases. To start with, the first author classified the titles and abstracts of the articles as relevant or irrelevant. This resulted in 142 studies being relevant for further inspection. In the second screening round a number of inclusion and exclusion criteria were applied. Studies were included if they answered to the conditions of CLIL as discussed in our introduction, contained a measure of one or more EFL skills, were conducted in the course of the past 20 years, written in English, had participants in mainstream primary or secondary education in a European country, made use of a control group and were longitudinal; that is, they had more than one measure in time of the same cohort, with statistical analyses of significant results. They were excluded if the English language used as the medium of instruction in bilingual forms of education was not a foreign language, as for example in English–Welsh programmes in Wales. Full-text articles were retrieved and further inspection by the first author using the same inclusion criteria resulted in 19 articles and two dissertations.

Presentation of the results

To systematically evaluate the retrieved texts, they were first subdivided into two categories: studies conducted at primary or at secondary schools. An overview was made of key characteristics for each study: title, year of publication, country, name of the author or authors, the number of times data were collected, CLIL content subjects, sample size for CLIL and control groups and starting age of the participants – their age at the first research measurement. Longitudinal CLIL effects – that is, findings of a greater increase in EFL scores over time of experimental groups when compared with control groups, provided they were significant – were categorized as pertaining to overall proficiency, which also may include oral skills, vocabulary, grammar, reading and listening skills. After drawing up overviews of the above data in two tables, we concentrated on describing the educational and research context for each of the studies. Our aim was to provide information on the development and implementation of CLIL in the respective countries, selectiveness and classroom practice in the diverse educational settings, the tested linguistic skills and research outcomes, and the practice of matching CLIL and control classes. If possible (i.e. if mentioned in the study under discussion) we also described effect sizes of significant results.

Results

Tables 1 and 2 provide overviews of the results of the selected studies, for primary and secondary education, respectively.

Primary education outcomes

As outlined in Table 1, four primary school studies were evaluated. Starting in the Netherlands the only study (van der Leij et al., 2010) conducted in primary education is an early, small-scale study at a time when primary CLIL was by no means common ground. Tested scales were L2 vocabulary, word reading fluency and orthographic knowledge, the information stored in memory telling us how to represent spoken language in written form.

Table 1. Results for primary education.

Country	Authors	Year	Primary education						Longitudinal CLIL effects per measured scale				
			Publication	Data collecting times	CLIL subjects	N CLIL	N Control	Age at T1	Overall proficiency	Vocabulary	Grammar	Reading fluency	Listening
Netherlands	Van der Leij et al.	2010	Acquiring reading and vocabulary in Dutch and English	2 in 1 year	Reading	23	23	8	_	Significant	_	None	_
Spain	Agustín-Llach and Alonso	2014	Vocabulary growth in young CLIL and traditional EFL learners	3 in 3 years	Natural science	58	49	9–10	_	None	_	_	_
Spain	Agustín-Llach	2015	Age and type of instruction in lexical development	3 in 3 years	Natural science	68	61	9–10	_	None	_	_	_
Spain	Pladevall-Ballester and Vallbona	2016	CLIL in minimal input contexts: A longitudinal study of primary school learners' receptive skills	4 in 2 years	Science Arts & crafts	138	149	9–10	_	_	_	None	Negative

Table 2. Results for secondary education.

Country	Authors	Year	Secondary education						Longitudinal CLIL effects per measured scale				
			Publication	Data collecting	CLIL subjects	N CLIL	N Control	Age at T1	Overall proficiency	Vocabulary or idioms	Grammar	Reading comprehension	Listening
Sweden	Sylvén	2010	Teaching in English or English teaching?	3 in 2 years		99	264	15–16	_	None	_	_	_
Austria	Gierlinger and Wagner	2016	Revisiting CLIL-based vocabulary growth in secondary education	2 in 1 year	Geography History Chemistry	39	48	14	_	None	_	_	_
The Netherlands	Admiraal et al.	2006	Evaluation of bilingual secondary education in the Netherlands	16 in 6 years	History, geography	584	721	12	_	None	_	_	_
The Netherlands	Verspoor et al.	2015	The effects of English bilingual education in the Netherlands	3 in 1 year	50% of the curriculum	Grade 1: 83	Regular 84 Control 49	12	None	None	_	_	_
						Grade 3: 74	Regular 68 Control 41	14					
Netherlands	Goris et al.	2013	Effects of the CLIL approach to EFL teaching: A comparative study	2 in 2 years	50% of curriculum	Netherlands 37	Netherlands 47	12	_	None	Significant	None	_
Germany					history, geography	Germany 50	Germany 47	13	_	None	None	None	_
Italy					modules	Italy: 45	Italy: 37	14	_	Significant	Significant	Significant	_
Germany	Dallinger et al.	2016	The effect of CLIL on students' English and history competences	2 in 1 year	History	483	NonCLIL1: (SS) 354 NonCLIL2: (DS) 444	13	None	_	_	_	Significant
Germany	Rumlich	2017	CLIL theory and empirical reality	2 in 2 years	1 increasing to 3 subjects	503	NonCLIL1: (SS) 473 NonCLIL2: (DS) 182	12–13	None	_	_	_	_
Spain	Pérez-Vidal and Roquet	2015	CLIL in context: Profiling language abilities	2 in 1 year	Science	50	50	12–14	_	Significant	Significant	Significant	None

(Continued)

Table 2. (Continued)

Country	Authors	Year	Secondary education						Longitudinal CLIL effects per measured scale				
			Publication	Data collecting	CLIL subjects	N CLIL	N Control	Age at T1	Overall proficiency	Vocabulary or idioms	Grammar	Reading comprehension	Listening
Spain	Roquet and Pérez-Vidal	2017	Do productive skills improve in CLIL contexts?	2 in 1 year	Science	50	50	12 –14	_	None	None	_	_
Spain	Gené-Gil et al.	2015	Development of EFL writing over 3 years in secondary education	4 in 3 years	Science or social science	30	20	13	_	None	Significant	_	_
Spain	Gené-Gil et al.	2016	A methodology for longitudinal research on EFL written production	4 in 3 years	Science	30	30	13	_	None	None	_	_
Spain	Alonso et al.	2008	Plurilingual education in secondary schools: Analysis of results	2 in 2 years	20–25% of the curriculum	Group 1: 67 Group 3: 44 Baccalaureate:48	Group 1: 20 Group 3: 20 Baccalaureate: 30	12 14 16	Undecided	_	_	_	_
Spain	Ruiz de Zarobe	2008	CLIL and foreign language learning	3 in 3 years	Social science English literature	CLIL1: 24 CLIL2: 36	29	14–15	None	None	None	_	_
Spain	Merino and Lasagabaster	2018	The effect of CLIL programmes' intensity on English proficiency	2 in 1 year	CLIL− : 3.4 CLIL+ : 8.4	CLIL− 208 CLIL+108	77	11–12	Significant	_	_	_	_
Spain	San Isidro and Lasagabaster	2018	The impact of CLIL on pluriliteracy development	3 in 2 years	Social science	20	24	14–15	Significant	_	_	_	_
Spain	Pérez Cañado and Lancaster	2017	The effects of CLIL on oral comprehension and production	3 in 1 year and 6 months	2 subjects	12	12	15–16	Oral skills Significant	_	_	_	_
Spain	Pérez Cañado	2018	CLIL and educational level: A longitudinal study on the impact of CLIL on language outcomes	2 in 1 year 2 in 6 months	Arts, Maths Science PE Technical Social/Natural sciences	1033	991	Primary: 11–12 Secondary: 15–16	Significant	Significant	Significant	Significant	Significant

Note. SS: Same School; DS: Different Schools.

The two groups were comparable on Dutch origin, social-economic background, age and sex, and L1 skills that have shown to affect reading acquisition. Participants were assessed twice within one year halfway through in grades 2 and 3, when they were eight years old. The number of EFL lessons was the same as that of the control class, namely five 20 minute lessons per week in grades 1 and 2. The experimental group received substantial bilingual (Dutch and English-medium) instruction in English reading. The results show favourable effects for reading ability in both languages, as well as significant differences in vocabulary acquisition between the experimental and control group: the authors report an impressive difference in progress between the groups for L2 vocabulary. The effect size was reported to be η^2 p= .12, which may be interpreted as between 'moderate'(.10) and 'large'(.20). Even though this study is not about the L2-medium teaching of content, which is a distinct feature of CLIL, it provides support for its didactics of being exposed to L2-medium instruction.

The majority of studies included in the present review, both in primary and secondary education, were conducted in Spain. Spain is made up of 17 autonomous communities and CLIL implementation may vary greatly from one community to another. CLIL is less of an elitist approach and introduced at broad educational levels, in which a divergence of CLIL policies exists. For example, CLIL is compulsory in all primary schools in Navarre, whereas only a few experimental programmes have been implemented in La Rioja. Moreover, CLIL contents and policy may differ greatly from one school to the next as schools can decide for themselves. Often there are no admittance criteria and admission to CLIL is voluntary. This still implies (self) selectiveness: the average ability and motivation to learn English is presumably higher, a learner characteristic seen in the great majority of present-day non-obligatory CLIL environments.

An example of non-selective CLIL at a broad level is the bilingual project that was launched in 10 Spanish regions in 1996 as the result of an agreement between the Ministry of Education and Science and the British Council. In order to provide EFL education compatible with EU standards an enriched language learning model was implemented in state schools to provide better EFL education from infant school to secondary levels, for learners from all socio-economic backgrounds, who often had no access to foreign language learning. In one of Spain's northern regions a study at primary level was conducted in 2014 (Agustín-Llach and Alonso, 2014) testing receptive vocabulary growth over time. All participants were from the same school, but some years apart. The variable age was controlled for. All 4th, 5th and 6th graders participated in the study, first when the school had only English as a school subject and later when CLIL tuition was introduced. At the start of the study participants were 9 or 10 years old and attended the 4th grade; at the end of the study they were between 11 and 12 and in the 6th grade. The experimental CLIL group had been receiving extra exposure to English in natural science lessons for two hours a week since grade 1; the control group had only received traditional EFL lessons. Data were collected three times by means of 10-minute vocabulary tests during class time in three consecutive school years. CLIL learners showed higher figures for vocabulary sizes, and differences increased with grade until they became significant in the 5th and even more so in the 6th grade. As to longitudinal effects, CLIL learners showed slightly higher growth rates than non-CLIL learners, but no significant statistical effects were found: very similar patterns of lexical development in CLIL and traditional learners were observed. The authors mentioned age as a determining factor in receptive vocabulary knowledge: increasing exposure does not lead to lexical gains in young learners, but more benefit may be expected as they grow older.

A related study by Agustín-Llach (2015) gave similar results for different aspects of vocabulary development of a similar, slightly larger group of CLIL and traditional EFL learners, measured along three years. The study controlled for the variable age. Both groups completed a letter-writing task, which was then scrutinized for L1 influence in the form of borrowings – which the author

mentioned as typical for low level learners and generally an indication of overall lack of lexical knowledge in the L2 – and lexical creations. The results showed that vocabulary related to school and classroom activities and management was frequent in both groups of learners, and both produced roughly similar numbers of words related to the field of science. CLIL learners produced fewer instances of borrowings than traditional learners and these tended to decrease with increasing proficiency: the number of borrowings produced by CLIL learners in 6th grade was significantly lower than that of traditional EFL learners, faintly suggesting CLIL benefits. As in the previous study, the author regarded (young) age as a factor imposing a strong constraint in L2 lexical development, even more so than exposure time in itself.

The last primary school study (Pladevall-Ballester and Vallbona, 2016) was conducted in the Spanish bilingual region of Catalonia in which Catalan, the language of instruction, together with Spanish, are the majority languages, and English is taught as the main foreign language in mainstream education. CLIL was initially the approach adopted to revitalize Catalan after the Spanish dictatorship. All subjects except the Spanish language have since been taught through Catalan. Receptive skills (reading and listening) in English were measured by means of Cambridge language tests at four different times during two academic years and involved pupils aged 9 or 10 in the 5th and 6th grades. In order to guarantee comparability between the two groups, the total amount of exposure up to each testing time was kept the same. Four state-funded private schools participated with two groups, one exposed only to EFL sessions and the other one exposed to EFL sessions and an additional CLIL hour per week. CLIL was new: two of the schools had implemented CLIL in the science subject, and the other two used it in the arts and crafts subject, both as a first time experience. The results showed no significant differences between the groups with regard to reading skills, while the control group significantly outperformed the CLIL group in listening skills. As an important factor underlying this negative CLIL outcome the authors observed that the results of the comparison between the arts and crafts group and its control group at the third measure were significantly higher in favour of the control group, which almost certainly had an effect on the overall results of the sample. Discrepancies in instructional practice between the two CLIL groups regarding their content subjects were observed. The listening competence of the young learners in arts and crafts could have been affected by instructional practices that relied heavily on visual inputs and gestural support compared with the science group. As a result, the arts and crafts learners did not necessarily have to confront the demands of the language so did not have to develop the strategies necessary to improve their listening abilities as much as the science learners did. In addition, the language used in the arts and crafts lessons was mainly based on instructions, and did not require as much cognitive effort on the part of the learners as the language used in the science lessons, which was much more complex and demanding, thus favouring the learners' listening comprehension development. The authors concluded by mentioning that the CLIL implementation was new to the teachers, which probably affected their instructional skills.

Secondary education outcomes

Table 2 presents an overview of studies conducted in secondary education. In the northern part of Europe we found only one study: a Swedish dissertation (Sylvén, 2010). In spite of the popularity of the English language – according to EU figures about 95% of the Scandinavian population have EFL proficiency to some degree – longitudinal CLIL studies are thin on the ground in the Nordic area. The author mentioned the fact that in grade 10, when CLIL starts, the year groups are split up into several different groups for elective reasons and sometimes optional courses, which hampers longitudinal research studies. Selection of students based on proficiency tests is not allowed, but in practice CLIL is self-selective in that it attracts students who are already fairly proficient in English.

The study analysed the testing of four different types of vocabulary in three rounds spread over two years and showed favourable results for the CLIL groups compared with the same-school mainstream control groups, matched for age. A possible bias lay in student motivation to participate in the project, about which the author remarked:

> Whereas the CLIL students were happy enough to show their skills, the control group found it hard to find the motivation. Their general feeling was best described as: 'This is only meant to show how badly we perform in comparison with the CLIL group' (Sylvén, 2010: 49).

Vocabulary growth was a frequent focus of CLIL research as the increased exposure to the L2 is seen as crucial for the acquisition of new words. In Austria, Gierlinger and Wagner (2016) investigated CLIL-based vocabulary growth and conducted a study in non-selective, rather low-achieving mixed-ability classes in lower secondary education. A popular Austrian CLIL form in lower secondary and primary education nowadays is the teaching of content in a number of modules, interspersed with mother tongue teaching. The participating schools offered chemistry, history and geography in around five to seven modules of up to four weeks each, representing 60–80 hours of CLIL class time, spread over the school year. The study was conducted before and after the CLIL intervention within the school year 2010–2011. Participants were from four different schools and formed two CLIL and three control classes, matched for age, in which the CLIL learners were not preselected but formed part of a whole-class of mixed ability policy who participated in their school-wide CLIL enrichment project. The authors found no significant effect on EFL vocabulary scores for the CLIL treatment or for CLIL exposure over time. Moreover, their investigation of CLIL teacher classroom language indicated that teachers mainly used the band of 1000 high-frequency words in the content lessons, which may have restricted more advanced and broader word learning.

In Dutch secondary education three studies (Admiraal et al., 2006; Goris et al., 2013; Verspoor et al., 2015) were conducted at pre-university level with pupils aged 12 at the start, being the start of secondary school and also CLIL. The predominant type of Dutch CLIL is about 50% of the curriculum being taught in English, with the main focus on the content rather than the language. Admittance to a CLIL class is offered mainly at mainstream pre-university secondary education and subject to selection criteria: good overall academic performance, motivation to persevere and an above-average interest to learn the English language. The Admiraal et al. (2006) study was started in 1993 when the CLIL approach was new in mainstream education in the Netherlands. The authors used the abbreviation BE (bilingual education, a term still used to denote CLIL). The study tested receptive vocabulary, reading comprehension and oral proficiency; however, only receptive vocabulary was measured longitudinally with 16 tests spread over six years. The other two scales had one measure only, and both showed higher scores for the BE (CLIL) groups. Vocabulary mean scores at the start of the programme were significantly higher for the CLIL group than for the control group, also after student characteristics – gender, general ability, home language, language contact, motivation to learn English, introduced as covariates – were taken into account. As to score development on the EFL vocabulary test, there was no significant difference between the experimental and control groups; the CLIL learners did not acquire English words faster than the control group. The population and CLIL practice of this early study is not representative for the later, more developed CLIL in the Netherlands. Students from the BE (CLIL) programme differed significantly on student characteristics such as nationality and native language, which may be explained by the fact that three out of the five experimental schools were schools for international education, accommodating more proficient EFL speakers. The other two were mainstream schools populated by almost 100% Dutch native speaker students, offering a mix of a bilingual programme

and regular secondary education. The control groups were partly from these two, and partly randomly sampled from two other, non-CLIL schools.

The second study (Verspoor et al., 2015) conducted in the Netherlands is of a more recent date and fully represents the Dutch CLIL approach as accredited by the European Platform for Education in the Netherlands, merged with Nuffic in 2015,[1] which coordinates and monitors all CLIL schools. The authors tested receptive vocabulary and full linguistic repertoire by means of a productive informal writing task. There were three measurements spread over one year – in October, January and June – at four secondary schools, with participants in grade 1 and grade 3. The authors were of the opinion – and agreed with Bruton – that comparing CLIL classes with only non-CLIL classes would give a distinct bias as in their experience non-CLIL classes have on average a lower scholastic aptitude, less motivation and lower initial EFL proficiency. They included students of the – prestigious and selective – Dutch *gymnasium* as a better match to act as a control class for CLIL and therefore incorporated three streams in the study: CLIL, regular (same school non-CLIL) and *gymnasium* (as control). The results indicated that at the first test in grade 1 the CLIL as well as the control classes had significantly higher scores than the regulars, with no such differences between CLIL and control. However, CLIL learners were significantly more motivated than regular and control peers and gained more: at the second and third tests the CLIL scores were significantly higher than those of both the regular and control groups, while between the latter two there were no significant differences. In the third grade – a different cohort – CLIL did not continue to gain more. The authors concluded that, when all covariates were taken into account, CLIL only outperformed the regulars, not the control classes, so that CLIL maintained the lead rather than gaining more than the other groups. Initial proficiency was found to be a significant contribution to final proficiency. The authors wondered whether similar proficiency results could be accomplished by better or more EFL teaching in regular programmes.

The reason for the initial better performance of the CLIL groups remained unclear. Prospective learners for CLIL streams in the Netherlands receive no special preparation: in principle they have the same EFL primary school lessons, even though practice between schools differs: some start earlier and do more than the obligatory lessons. Both studies tested mainly vocabulary. It is possible that learners with an inclination towards the English language pick up more words from its omnipresence in daily life. Nor was it clear why vocabulary gain seemed to be initial and did not develop in a linear fashion. A possible ceiling effect was mentioned, or the fact that vocabulary tests may have limitations in that they make use of frequency bands, testing the most common words whereas CLIL vocabulary gain is likely to be found in subject-specific or academic use of language.

The third study (Goris et al., 2013) was conducted between 2007 and 2009 and tested more language skills apart from vocabulary. The results in three countries were compared: the Netherlands, Germany and Italy. The experimental groups were representative for national CLIL practice in their region at the time. Tested skills included receptive vocabulary, idioms, grammar and text comprehension. In each country four classes from pre-university secondary schools took part in the study: two CLIL groups and two control groups. There were two measurements: the first one at the beginning of secondary school, also the starting point of CLIL, and the second one two years later, at the end of the second grade. With one exception in the Netherlands – where one control group from a classical *gymnasium* took part – the control groups were from the same schools as the experimental groups. Pupil age, scholastic aptitude, socio-economic status (SES) and demographic background, national majority native language, EFL contact and preparation for CLIL were the same for each experimental-control match per country, even though between-countries differences existed as to age and CLIL preparation. The results indicated that the initial test scores on the whole were highest in the Netherlands and lowest in Italy. Furthermore, the CLIL

groups in all three countries had higher – though not always significantly – initial scores than the control groups on all scales, apart from vocabulary in Italy where the control groups knew slightly more words. The second test round showed higher scores on all scales for the CLIL groups. For gain scores there were differences between countries. In the Netherlands CLIL gain scores were significant for grammar and idioms, in Germany no significant CLIL gains were found, while in Italy such scores were found on all tested scales, which was surprising in view of the modest modular CLIL approach. A possible explanation for the successful outcomes in Italy lay in a Hawthorne effect – the influence resulting from positive feelings caused by taking part in a test – as both the CLIL approach and the EFL presence on the curriculum were very new here and the international university research met with great enthusiasm from both teachers and pupils. The opposite effect could have occurred in Germany, where the strict anonymity required for participation in research possibly contributed to a disinterested attitude – or the fact that the English language was not abundantly popular in society at large.

In Germany two robust studies (2016, 2017) were retrieved. The 2016 study (Dallinger et al., 2016) investigated skill development of 1806 German CLIL and non-CLIL 8th-graders in English and history, controlling for a wide range of student, classroom and teacher characteristics such as student EFL interest. The study made use of two kinds of non-CLIL control groups: cohorts from the same school as the experimental CLIL groups (non-CLIL1) and cohorts from different schools without CLIL programmes (non-CLIL2). There were two measurements: one at the start and one at the end of grade 8. In Germany, parents may decide for or against a CLIL school when registering their child for secondary education – mostly schools preparing for university. This type of school accommodates mainly learners from high-SES homes, in which one or both parents followed academic education. Admittance is decided together with the teacher and school admittance criteria. The German educational system in most states provides extra EFL education in preparation for CLIL, which results in substantial differences in initial EFL proficiency between CLIL and non-CLIL: the former are likely to possess higher prior knowledge in English.

In the study under discussion CLIL classes accommodated more students with immigration backgrounds and higher SES than non-CLIL. The CLIL classes' initial general English skills – as measured by a C-test[2] – and listening comprehension were significantly better than those of the non-CLIL classes. This was also the case at the second test. For the longitudinal pre-test–post-test effects the results of multilevel modelling confirmed that CLIL classrooms showed significantly greater increases in English listening comprehension but not general English skills than non-CLIL classrooms. Controlling for systematic differences by means of multilevel regression analyses greatly diminished the CLIL effect on English skills for both domains, and rendered that on general English skills insignificant. The authors mentioned prior achievement as the strongest predictor of future learning.

Similar findings were discussed by Rumlich (2017). The article summarizes the author's dissertation on CLIL streams at German secondary schools of the *Gymnasium* type in North-Rhine Westphalia, the most populous German state with the largest number of CLIL schools. These schools have the most intense form of CLIL implementation in mainstream German education, and offer continuous forms of CLIL with a duration of at least one school year. At the start one content subject is involved, in grades 8 and 9 two or at most three subjects. In preparation for this high-intensity version students receive additional EFL lessons in grades 5 and 6 before the start of CLIL in grade 7 at ages 12–13. Over 1000 learners were involved in a two-year pre-test–post-test research study into the effects of CLIL on EFL proficiency as measured by C-tests. As in the earlier mentioned German study three groups participated: CLIL classes and two kinds of non-CLIL control groups. They were cohorts from the same school as the experimental CLIL groups (non-CLIL1) who 'might represent a negatively selected group of students with below average EFL proficiency'

(Rumlich, 2017: 115) and therefore inappropriate as controls. The cohorts from different schools without CLIL programmes (non-CLIL2), whose students are unselected and unprepared, were seen as neutral and more appropriate controls groups. The author observed that prospective CLIL and non-CLIL general proficiency in English diverged greatly in favour of the former even before the implementation of CLIL, partly due to the preparatory lessons and partly to selection effects. The findings on EFL development indicated that all groups advanced significantly over time, with little difference between the CLIL and non-CLIL2 controls and a slightly weaker progress for the non-CLIL1 group. It became evident that there was no detectable influence of CLIL on general EFL proficiency. The author concluded with a strong claim for longitudinal evaluations that screen for selection, EFL preparation and class composition effects.

At secondary level Spanish studies also had the lion's share. In Barcelona Pérez-Vidal and Roquet (2015) investigated the effects of a newly introduced CLIL programme which was carefully designed by a team of university experts, in order to identify which areas of L2 competence benefitted the most from CLIL instruction. Two different groups of Catalan/Spanish bilingual learners were analysed longitudinally over one academic year. The productive skill of writing and receptive skills of reading and listening were investigated, as well as lexical-grammatical ability, the skill to use vocabulary and grammar correctly. Written development was analysed quantitatively taking into account syntactic and lexical complexity, accuracy and fluency, and qualitatively for task fulfilment, organization, grammar and vocabulary. The experimental group, who had previous experience with CLIL since they were 10 years old in grade 5, received formal EFL lessons as a school subject, in addition to an English-medium science subject taught with a CLIL approach. The control group received formal instruction of English as a subject only. Data collection started at the end of their first year of secondary education, in grade 7. The CLIL group was not matched for age with the control group as this would have given them the advantage of more EFL exposure, but on the grounds of a similar total number of hours of EFL education. This meant that the control group included learners who were a year older and one grade higher than the CLIL group. The results of two measurements within one school year for receptive skills showed that the CLIL group improved their reading competence significantly more than the control group, but not their listening competence. CLIL students' lexical-grammatical ability also increased significantly more. For the productive skill of writing, there was a significant improvement in favour of the CLIL group in their abilities to write more accurate and syntactically complex texts, and a general improvement in the whole set of qualitative measures (task fulfilment, organization, grammar and vocabulary). The positive results for writing skills appear to be in contrast with Dalton-Puffer's (2008) findings, who classified writing as one of the areas of linguistic competence likely to remain unaffected by CLIL instruction as content teaching is conducted almost completely without writing activities. However, the Spanish authors claimed a transfer of writing skills to the CLIL context from the mainstream EFL lessons, in which writing is often practised.

Even though the results for writing skills development turned out to be positive for the CLIL group, only a part was significant. The same authors – with a reversal of first and second authorship (Roquet and Pérez-Vidal, 2017) – produced a more detailed discussion of the participants' written production outcomes. The study singled out the composition assignment – writing a dialogue based on a picture – that was part of the original testing battery and used the same criteria to evaluate the development of aspects of written abilities. The results of the qualitative and quantitative measures indicated that for syntactic and lexical complexity as well as fluency in writing, the CLIL group did not progress significantly more than the control group; this was only the case in the domain of accuracy. With regard to qualitative results in the field of task fulfilment, organization, grammar and vocabulary, there were no significant differences between the progress of the two groups. Overall, the authors could not confirm the superiority of CLIL for writing skills.

A further study into fine-grained aspects of writing competence was conducted in a different part of Catalonia, on the Balearic Islands, by Gené-Gil et al. (2015). They evaluated complexity, accuracy and fluency (CAF) in a CLIL programme called 'European Sections', introduced in 2004. Schools joined the programme on a voluntary basis and participation of students was optional. Participants were two groups of Spanish/Catalan bilingual secondary students, an experimental CLIL group learning science or social science through the medium of English and a comparable non-CLIL control group with English as a subject only. All participants answered a profile questionnaire enabling the researchers to rule out any important differences in language background and extra-school exposure to the target language. CLIL and non-CLIL students' average age was 13 at the first test at the start of CLIL in 2008. Both CLIL and non-CLIL groups were asked to write a timed composition (25 minutes) in English at every data collection time, four times spread over three school years. The results indicated no significant differences at the first test: CLIL and non-CLIL participants' onset level was equivalent. Over the course of three school years, CLIL students attained significantly improved performance in every domain analysed (written syntactic and lexical complexity, accuracy and fluency), whereas in the case of their non-CLIL counterparts improvement was restricted to the lexical complexity and accuracy domains.

In a later article Gené-Gil et al. (2016) critically examined their methodology. Their major concern was to describe EFL writing in all its complexity and multidimensionality, a difficult task as fine-grained effects of writing skills are difficult to assess, let alone evaluate and compare statistically over time and between groups. They held the pre-test–post-test design with repeated measures to be the only way to control for differences in task complexity, despite the risk of carryover effects. In a study related to the previous one they used micro-analytical measures of diversity in use of vocabulary or lexical complexity and accuracy. Participants were two groups of comparable adolescent learners of EFL: an experimental CLIL group and a non-CLIL comparison group. The former received three hours a week of EFL instruction and had started studying science in English for three hours a week in grade 8 on a voluntary basis. The latter only received EFL instruction. Data about written production were gathered through two communicative writing tasks, designed for this study: a general, interpersonal task and a subject-specific, yet sufficiently general task related to science, the CLIL subject in the participating schools. The results indicated that CLIL learners showed improved writing competence, particularly in micro-analytical measures of diversity in vocabulary use and accuracy, and that they incorporated some specialized lexis into their subject-specific compositions. The authors proposed a methodological framework with a greater focus on evaluation procedures of such details, as most writing activities simply support grammar and vocabulary learning, and to encourage more investigations of EFL writing development in CLIL contexts in which progress often lies in nuances. They confirmed that more exposure to the target language in itself does not necessarily lead to enhanced written competence.

Several Spanish studies were published in the bilingual Basque country. The educational system in the Spanish Basque Autonomous Community (BAC) offers linguistic models with either Spanish or Basque as the language of instruction and the other one a curriculum subject, or with Basque and Spanish-medium teaching side by side. At the start of the millennium foreign language medium teaching, using foreign languages – mostly English – in addition to the two national languages, was introduced. The Department of Education initiated a 'Plurilingual Experience' (PE) in 12 schools in 2003, six of them participating in a case study analysing its effects. The PE programme required that at the compulsory level at least seven hours a week be taught in a foreign language – other than Basque or Spanish. In the post-compulsory or baccalaureate years, when students are 16–18 years old, the number of L2 medium subjects should be at least 20–25% of all lessons. The experimental schools offered the opportunity to study the curriculum subjects in three different languages: Basque, Spanish and English (L3 for the students). Participants had to answer selection criteria of

academic performance and motivation. There were two measures at three levels (grades 1, 3 and baccalaureate): the first one in October 2004, at the start of PE, the second in May 2006, after two years. The control groups were in the same grade as those of the experimental groups but did not take part in the PE. The evaluation looked at English proficiency – listening and reading comprehension, written and oral production and grammatical knowledge using validated Cambridge tests as well as qualitative data concerning opinions from students and teachers.

The report on the results of the PE project 'Trilingual students in secondary education' was published in 2007 (Basque Institute of Educational Evaluation and Research, 2007), and the research team (Alonso et al., 2008) summarized the study. The findings showed that the English-medium experimental group achieved considerably better results than the control group, advantages that increased even more in the course of time. Comparisons and longitudinal growth were calculated and described in percentages – not very common in statistical analyses. To estimate the accumulated gains, which were quite substantial, the grading for each test was distributed on a common scale so that the differences between each phase could be summed. For this, the approximate and provisional results obtained in earlier research on the empirical validation of the CEFR (Common European Frame of Reference[3]) scales were used. The research team stated that this analysis is a theoretical simulation of scores that refer to the hypothetical differential gain that would arise if the data were presented on a common scale, so that results have to be taken with caution. For the present review it means that the question whether the effects were statistically significant remains unanswered. We felt that this landmark study had to be included in our discussion, the more so as it initiated prolific CLIL practice and research within the Basque Country.

Ruiz de Zarobe (2008) conducted a study to investigate if including more English-medium content subjects enhances CLIL effects on L2 proficiency (the more-content-is-better hypothesis) and tested EFL speech production of learners in two different CLIL programmes. Adequacy as to pronunciation, vocabulary, grammar, fluency and production of content were compared with those of a control group enrolled in a traditional EFL programme only. The first CLIL group, CLIL1, entered CLIL at 14, when one content subject, social sciences, was taught through English for three or four hours a week. The second CLIL group, CLIL2, entered a CLIL programme with two English-taught content subjects: social sciences (three to four hours a week) and modern English literature (two hours a week). The control group only received the conventional three hours of EFL per week, the same as the CLIL groups. Learners were in grade 3, aged 14–15 at the first test and in the last year, the pre-university year or baccalaureate, at the last test. By then, the number of participants supplying data had decreased: the control group consisted of seven students, the CLIL2 group of 14 and data of the CLIL1 group could no longer be collected at the time. The results showed that at the first measurement in grade 3 the CLIL groups significantly outperformed the control group in all scales, without important differences between the CLIL types. In the 4th grade CLIL2 once again scored highest, with significant differences in all scales except vocabulary. The participants in the pre-university grade belonged to two groups: CLIL2 and control. The differences between the two were significant for vocabulary and grammar. Although the CLIL groups' scores were higher than those of the control group throughout the grades, there seemed to be no significant increase in proficiency throughout the years. The question if more content is better could be answered affirmatively: a positive relationship was present between the amount of EFL exposure and the linguistic outcomes.

Merino and Lasagabaster (2018) addressed a similar issue: the role played by intensity in CLIL programmes on overall proficiency in English: speaking, reading, listening and writing. Participants were 393 secondary education students from three different autonomous communities: the BAC, an officially bilingual community in Basque and Spanish; La Rioja, a monolingual Spanish community; and Cantabria, also a monolingual community. They were spread over two experimental

groups and one control group. The control group consisted of 77 learners from eight schools with Basque as the means of instruction for all subjects except Spanish. The first experimental group (CLIL–) was made up of 208 CLIL learners from the same eight schools, who had 3.4 CLIL sessions per week. The second experimental group (CLIL+) comprised 108 CLIL learners from five high schools in Cantabria and La Rioja with 8.4 CLIL sessions per week. In two test rounds in the school year 2010–2011, the first one in the final term of grade 7 and the next one at the end of grade 8, Cambridge ESOL tests (Key English Test) were administered. The findings showed that a higher amount of CLIL produced a greater improvement in the L2. At the first test, both CLIL groups showed significantly higher scores than the non-CLIL group. No differences were found between the scores attained by CLIL– and CLIL+. However, at the second test the contrast between CLIL– and CLIL+ had increased and showed that the evolution of CLIL+ students was significantly higher than that of their CLIL– counterparts. The latter made progress in an almost identical degree to the control group. As in the previously discussed study (Ruiz de Zarobe, 2008), the question if more content is better could be answered affirmatively.

A small-scale study into general EFL as measured by proficiency was conducted by San Isidro and Lasagabaster (2018) in a rural multilingual school in Galicia, a Spanish north-western autonomous community with specific linguistic and cultural hallmarks. Two official languages are spoken: Galician and Spanish. The study investigated the impact of CLIL on multiple language learning (L1, L2 and EFL) and content learning. The first measurement of the two-year longitudinal study was at the start of the third grade of secondary education in September 2012, when learners were 14–15 years old. The final test was at the end of the 4th grade. EFL scales had three measures using Cambridge tests and included the four skills: reading, writing, speaking and listening. Participants were two comparable groups without any prior selection, and homogeneous before the implementation of CLIL. The results indicated that both groups improved their competence in English after two years, while the CLIL cohort made significantly greater progress. Another interesting finding was that the CLIL students also outperformed their non-CLIL counterparts in both Spanish and Galician over the two school years, whereas content learning was not negatively affected.

A small-scale study on listening and oral production skills (Pérez-Cañado and Lancaster, 2017) was published on the outcomes of a 4th grade secondary school sample. The pre-tests took place in September 2012; the post-tests were carried out in June 2013; and the second post-tests were completed in January 2014. The CLIL and control groups had initially been matched on the pre-test, when no statistically significant differences were found on listening skills. Over the period of one academic year and six months following the conclusion of the intervention programme, the post-tests were applied. The listening tests were group-administered in one sitting under the same conditions each time. In turn, in the oral production tests, the students were examined in pairs, with individual subtasks lasting up to five minutes. On the post-test statistically significant differences emerged in favour of the CLIL group, for listening skills as well as for oral production skills: all of the tasks in the oral production test (spoken interaction in an interview and in individual speaking) evinced statistically significant differences in favour of the CLIL group, demonstrating that these students are able to communicate more effectively. At the final stage of the investigation, both groups had levelled out on oral comprehension competence: no statistically significant differences were detected for any part of the test or for the test as a whole. For oral production skills, however, the scores of the CLIL students significantly surpassed those of their non-CLIL peers at the second post-tests, with statistically significant differences for the overall test and each of its tasks and skills.

The last study to be discussed here is a large-scale investigation into EFL proficiency (Pérez-Cañado, 2018) involving 1033 CLIL students and 991 EFL learners in 53 public, private and charter schools[4] in three of the monolingual communities in Spain which have the least tradition in

bilingual education: Andalusia, Extremadura and the Canary Islands. Tested skills were grammar, vocabulary, reading, listening and speaking. At the start of the academic year 2014–2015 the experimental and control groups of 11–12-year-olds were matched on a pre-test in terms of SES, English level, verbal intelligence and motivation. Thus, homogeneity between both cohorts was initially guaranteed. At the end of the same academic year, in June 2015, when 828 learners (aged 11–12) were finishing the last grade (grade 6) of primary school and 1196 (aged 15–16) were about to complete the last grade (grade 4) of compulsory secondary education – so all of them were on the switch to the next level – the English language post-tests were administered. Six months later, in December 2015, the delayed post-test was applied to the same participants who were previously in the 4th grade of secondary education and had continued in the 1st grade of the baccalaureate, where CLIL instruction stopped. The results of the English language post-test at the end of primary school showed statistically significant differences on all the linguistic components and skills sampled, invariably in favour of the CLIL group. The effect sizes for vocabulary (Cohen's $d = -0.619$) and the five aspects of speaking (−0.858) can be seen as medium even though 0.08 is close to 'large', 1 standard deviation. After four additional years of participation in CLIL education, at the post-test at the end of the 4th grade, the differences in EFL competence had increased. Statistically significant differences were found in favour of the CLIL cohorts on all the linguistic aspects sampled, at high confidence levels and with large effect sizes. The latter were particularly considerable for use of English, with Cohen's d bigger than 1, so larger than one standard deviation (−1.160), and speaking (−1.230), especially as to lexical range (−1.442) and task fulfilment (−1.482).

Time turned out to be a crucial factor to ascertain the effects of CLIL on foreign language attainment; the longer the students had been benefitting from bilingual education, the greater the differences with their non-bilingual counterparts. The effects pervaded and became even stronger six months later, when the former 4th graders were in the first year of non-compulsory secondary education. Statistically significant differences continued to be discerned in favour of bilingual streams on all the linguistic components and skills sampled, at high confidence levels, and with even larger effect sizes, especially for speaking (Cohen's $d = -2.671$, larger than 2 standard deviations), and except for reading, which had the comparatively lowest effect size (−0.868). At this level, it turned out that productive skills had been more positively affected by CLIL as opposed to receptive. Interesting data were found in the follow-up of the 4th graders on aspects of oral competence that need more time to develop in order to be significantly improved, namely pronunciation and fluency. Also at this point, type of school yielded interesting results: the non-bilingual charter schools were catching up with the public and private bilingual ones, as there were no statistically significant differences between them and both bilingual types of schools on the use of English.

Conclusion and discussion

The present review aimed to evaluate the findings of longitudinal experimental research into the effects of the CLIL approach on EFL proficiency, conducted over the past 20 years. Our findings, as highlighted in Tables 1 and 2, do not provide unequivocal support for the hypothesis that learners in a CLIL class will develop more EFL proficiency over a certain period than their mainstream counterparts; the majority of studies produced null effects. Furthermore, it is striking that studies with a longitudinal perspective have been undertaken in only a limited number of European countries, and mainly during the last five years. The number of participants varies greatly. Nevertheless, even though an overall picture of the longitudinal effects of CLIL in each of Europe's quarters is lacking, all findings, large-scale as well as small-scale, are valuable and helpful in analysing this massive educational innovation, seen by many as best practice for the future.

A variety of EFL competences was investigated and there was a broad diversity of skills analysed. The most frequently tested skill was vocabulary. The three Dutch studies, as well as the Swedish and Austrian ones, tested receptive vocabulary, for which no significant growth was found over time. The Goris et al. (2013) study also tested other skills in the Netherlands and found significant results for idioms and grammar, an indication that CLIL affects certain linguistic skills more than others. The same study did not find significant effects of the CLIL lessons on any linguistic skill in Germany. Several studies tested overall EFL proficiency, for which no significant effects were found in the Netherlands or Germany where spoken skills were not part of the tests.

Productive writing was tested in several Spanish areas. The authors discussed various aspects of writing skills (for the sake of convenience we listed them under 'grammar' and 'vocabulary' in Table 2), some of which indicated the better skills of CLIL learners. Spoken fluency was tested in Spain where five studies addressed this skill as part of overall proficiency in receptive and productive skills. Three of these studies found significant results in favour of CLIL for spoken fluency, the skill commonly believed to be the most favourably affected because of the increased opportunity for authentic communication.

A topical issue in many European educational systems is whether an early start of L2 learning is better than postponing it until a later date, or whether it is wise to include CLIL courses before any teaching of the target language has taken place. In most European countries, CLIL was hardly practised at the elementary level until recently, and as yet, no comprehensive literature discusses early mainstream CLIL results. Apart from one primary school study in the Netherlands, which presented positive effects on vocabulary, mainly as a result of teacher instruction rather than a content subject, three studies were found in two different regions in Spain. No convincing evidence in favour of an early introduction of CLIL emerged, while one study even indicated negative results for listening skills. The authors assumed that (young) age is a constraining factor in L2 development. Studies undertaken with somewhat older primary school learners provided more positive outcomes, as in Pérez-Canado (2018) where a large group of 11–12-year-old CLIL learners already obtained higher EFL gain scores than their mainstream peers after one year.

Another recurrent issue in the CLIL literature is the question of whether more CLIL lessons lead to more positive L2 effects. Two Spanish studies in the present review addressed this question, and both studies provided affirmative answers. It seems best to consider these results as context-specific, for across countries they do not always materialize. The results in the Netherlands or Sweden, where a large percentage of the curriculum is offered in the target language, were not strikingly better than those in Germany, for example, which practised a moderate approach by including two or three content subjects. Several authors provided indications that CLIL results grow over time; learners profit more after a longer period of learning in a CLIL track.

It is difficult to interpret our findings without discussing the practice of matching experimental and control groups and the initial CLIL head start, issues criticized by Bruton (2011). The most reliable practice from a statistical point of view would be to match pupils, ideally a random sample, as in Admiraal et al. (2006) who included a mix of same-school and random-sampled participants from non-CLIL schools in the control group, on crucial variables such as age, SES, intellectual and linguistic capacity, as well as initial EFL levels. Notably, this last feature presented problems and was not always observed in the studies under review, the main reason is that in many countries, pupils receive extra EFL preparation if they aspire to CLIL, as in Germany and Italy. The higher time of EFL instruction for the experimental groups in comparison with the mainstream groups was not always controlled, which is a recurring issue in CLIL research. One of the reasons may be that previous lesson time is difficult to measure in exact numbers as pupils come from a variety of primary schools or pre-CLIL years that do not always have similar EFL curricula. In Sweden, CLIL is self-selective and introduced for those interested in upper

secondary education, which implies that learners are already fairly proficient. This inevitably leads to an EFL head start when compared with mainstream counterparts. At the same time, the studies in the Netherlands showed that such head starts are also possible without special preparation. Therefore, the results of these studies must be interpreted with these limitations in mind. The CLIL initial head start was even found to be significant in several studies, which makes it difficult to explain the progress these pupils make in the following years; does it result from CLIL or the better L2 skills? The use of a control group with lesser EFL skills is thought to present bias in the research findings. Suitable options to match experimental and control groups were found in the Netherlands as there are two types of pre-university education, the *VWO*, the mainstream type of preparing those with sufficient academic talent for university studies, and the classical *gymnasium*, a more prestigious school type for those that excel in comparison with CLIL candidates. The study by Verspoor et al. (2015) involved groups from both types, and their study made clear that the difference in EFL progress between CLIL and non-CLIL *gymnasium* learners was not significant, whereas the difference was clear between CLIL and the 'regular' *VWO* learners from the mainstream departments of the participating CLIL schools. The former appeared to be an appropriate match for the CLIL learners, in contrast with the so-called same-school control groups, which the authors described as having, on average, a low scholastic aptitude, low motivation and low initial EFL proficiency.

In Germany, where schools preparing for university do not have a similar divide as in the Netherlands, the authors (Dallinger et al., 2016; Rumlich, 2017) expressed similar views, and Rumlich (2017: 115) described same-school control groups as 'a negatively selected group of students with below-average EFL-proficiency'. Cohorts from different schools without CLIL programmes were included as more appropriate controls in both German studies. They too, provided evidence for the same-school experimental-control group mismatch, as did student discouragement to participate in a school research project mentioned by Sylvén (2010).

In this respect, the Austrian study presented a more appropriate research context: the CLIL students were average learners and participated in their school-wide CLIL enrichment project. They were not preselected but formed part of a whole class of mixed ability policy. Suitable research contexts were also present in Spain, where experimental mainstream matches were found without initial differences in EFL skills, even though the problem was acknowledged here too, in the case of a control group that was one year older than the test group, which is not ideal either.

What our review has not made clear is the specific impact of various CLIL target language content subjects and if such effects exist apart from the classroom language used by all CLIL teachers. It is difficult to imagine that a language-rich subject, such as history, influences the target language to the same degree as a subject that depends on complex cognitive explanations, such as mathematics. A few authors elaborated on this issue even though Pladevall-Ballester and Vallbona (2016) explained that more complex language favoured the learners' listening comprehension development. For the rest, our study produced no indications that the target language of some subjects' content affects EFL skills more than or differently from others.

Another issue not dealt with in the present review is the impact of content teacher skills, both from a didactic and linguistic point of view. Great varieties exist in the levels of teacher L2 skills. As an example we mention that, in Germany, content teachers are qualified to teach both the L2 and the subject, whereas, in the Netherlands, they are only qualified to teach their subject, while a B2 level for the target language is required for CLIL teaching. Also, classroom practice diverges across European countries. Sometimes native speaker teachers and language assistants are employed, while the role of the EFL teachers in the CLIL process also diverges. We feel that the impact of teacher skills is beyond the scope of this review, as it is a complex and many-faceted phenomenon that needs further study.

Our findings produced interesting contradictions between Spain on the one hand and the rest of Europe on the other. Whereas the latter were found to produce mainly null effects, in Spain, significant effects were more frequent. We must look at the history of CLIL to contribute to possible explanations. In most countries, CLIL developed from the bottom up, inspired by parental or educational demands. It was introduced almost without exception as a selective option at pre-university level, intended to prepare high SES students for international careers and studies. In Spain, CLIL had its origins in the opposite direction as it developed from the top down as a joint initiative of educational authorities. Its ideology was to provide better EFL learning opportunities for all, as, from the early years, the quality of teaching of English as a foreign language was minimal and not all Spanish children had access to EFL training. Spain, just like other southern-European countries, such as Italy, used to be and still is a society with low EFL proficiency. Italy was included in the Goris et al. (2013) study, which found several significant effects that were difficult to explain from the analyses of the results. According to the Special Eurobarometer on Europeans and their languages, issued by the European Commission in 2006 and updated in 2012, Spain and Italy represented the countries where the smallest percentage of citizens could hold a conversation in a language apart from their national L1. This may account for the positive effects of a massive L2 learning innovation. In Spain there was much more room for improvement than in Sweden or the Netherlands, which the Eurobarometer listed as among the top eight countries in which 9 out of 10 inhabitants could speak at least two languages, and in which English was spoken quite well by the great majority of people.

The focus of the present review was to investigate if CLIL has met its promise of providing a better EFL learning approach. The answer is by no means negative, but the degree to which it is positive varies. High EFL-proficiency countries with elitist and highly selective CLIL, such as the Netherlands and Germany, have gained little on the testing scales. In Spain, however, a low EFL-proficiency country, the CLIL approach was planted in fertile soil. Spain was desperately in need of improved EFL teaching in an increasingly internationalized market place, but at the same time there was ample experience in the teaching of content through two languages in Spanish bilingual regions. In the Basque Country, for instance, educational practice entails teaching content through Spanish and Basque, two completely unrelated national languages, which is claimed to be conducive to further L3 acquisition, in the case of CLIL of the English language.

The final thought of the present review may be summarized by saying that the research published in the past two decades about the benefits of CLIL has presented many valuable and robust findings, predominantly during the last five years and conceivably in response to criticism and reviews. Our most conclusive finding is that CLIL is profiled best in the divergent contexts of Spain, which sets positive precedence for other low EFL-proficiency countries in the EU that are still at less advanced stages of EFL skills. In this respect, more comprehensive studies into the quality of mainstream EFL education in these countries would be a welcome contribution to the research field. As we have seen in Spain, schools received support in introducing and practising CLIL. At the same time, little is known about improvements or developments in mainstream EFL teaching, so that comparing CLIL and mainstream performance may be biased in favour of the former.

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

JA Goris https://orcid.org/0000-0001-7431-0421

Notes

1. https://www.nuffic.nl/en/nuffic-glossary/
2. by Lucy Katona and Zoltan Dornyei from: Forum English Teaching
   The C-test is an integrative testing instrument that measures overall language competence, very much like the cloze test. It consists of four to six short, preferably authentic, texts in the target language, to which 'the rule of two' has been applied: the second half of every second word has been deleted, beginning with the second word of the second sentence; the first and last sentences are left intact. If a word has an odd number of letters, the 'bigger' part is omitted; for example, proud becomes pr-. One-letter words, such as I, are ignored in the counting. The students' task is to restore the missing parts. In a typical C-test there are 100 gaps; that is, missing parts. Only entirely correct restorations are accepted.
3. The Common European Framework of Reference for Languages (CEF or CEFR) was put together by the Council of Europe as a way of standardizing the levels of language exams in different regions. It is very widely used internationally and all important exams are mapped to the CEFR.
4. Charter schools are state-financed schools, most of which have a religious orientation.

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.

Agustín-Llach MP (2015) Age and type of instruction (CLIC vs. traditional EFL) in lexical development. *International Journal of English Studies* 16(1):75–96.

Agustín-Llach MP and Alonso AC (2014) Vocabulary growth in young CLIL and traditional EFL learners: Evidence from research and implications for education. *International Journal of Applied Linguistics* 2(2): 1–13.

Alonso E, Grisaleña J and Campo A (2008) Plurilingual education in secondary schools: Analysis of results. *International CLIL Research Journal* 1(1): 3.

Basque Institute of Educational Evaluation and Research (2007) *Trilingual Students in Secondary School: A New Reality*. Bilbao: ISEI-IVEI.

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

Bruton A (2015) CLIL: Detail matters in the whole picture. More than a reply to J Hüttner and U Smit (2014). *System* 53: 119–121.

Coyle D, Hood P and Marsh D (2010) *CLIL: Content and Language Integrated Learning*. Cambridge: Cambridge University Press.

Dallinger S, Jonkmann K, Hollm J, et al. (2016) The effect of content and language integrated learning on students' English and history competences: Killing two birds with one stone? *Learning and Instruction* 41: 23–31.

Dalton-Puffer C (2007) *Discourse in Content and Language Integrated Learning (CLIL) Classrooms*. Amsterdam: John Benjamins Publishing Company.

Dalton-Puffer C (2008) Outcomes and processes in content and language integrated learning (CLIL): Current research from Europe. In: Delanoy W and Volkmann L (eds) *Future Perspectives for English Language Teaching*. Heidelberg: Carl Winter, pp.139–157.

Dalton-Puffer C (2009) Communicative competence and the CLIL lesson. In: Ruiz de Zarobe Y and Jiménez Catalán RM (eds) *Content and Language Integrated Learning: Evidence from Research in Europe*. Bristol: Multilingual Matters, pp.197–214.

Dalton-Puffer C (2011) Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics* 31: 182–204.

Ellis R (1985) *Understanding Second Language Acquisition*. Oxford: Oxford University Press.

European Commission (2006) Europeans and their languages. *Special Eurobarometer* 243. Brussels: European Commission.

European Commission (2012) Europeans and their languages. *Special Eurobarometer* 386. Brussels: European Commission.

Eurydice (2005) Summary Sheets on Education Systems in Europe. http://www.indire.it/lucabas/lkmw_file/eurydice/NATIONAL_SHEET_EN_2007_PDF.pdf

Gené-Gil M, Juan-Garau M and Salazar-Noguera J (2015) Development of EFL writing over three years in secondary education: CLIL and non-CLIL settings. *Language Learning Journal* 43(3): 1–18.

Gené-Gil M, Juan-Garau M and Salazar-Noguera J (2016) A methodology for longitudinal research on EFL written production: Capturing writing multidimensionality by combining qualitative and quantitative procedures. *Journal of Research Design and Linguistics and Communication Science* 3(1): 23–48.

Gierlinger EM and Wagner TA (2016) The more the merrier: Revisiting CLIL-based vocabulary growth in secondary education. *LACLIL* 9(1): 37–63.

Goris J, Denessen E and Verhoeven L (2013) Effects of the content and language integrated learning approach to EFL teaching. *Written Language & Literacy* 16(2): 186–207.

Katona L and Dörnyei Z (2004) What the C-test is. Forum English Teaching.

Krashen SD (1982) *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon.

Küppers A and Trautmann M (2013) It is not CLIL that is a success: CLIL students are! Some critical remarks on the current CLIL boom. In: Breidbach S and Viebrock B (eds) *Content and Language Integrated Learning (CLIL) in Europe: Research Perspectives on Policy and Practice*. Frankfurt am Main: Peter Lang, pp. 285–296.

Lasagabaster D (2008) Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal* 1: 31–42.

Lorenzo F, Casal S and Moore P (2010) The effects of content and language integrated learning in European education: Key findings from the Andalusian sections evaluation project. *Applied Linguistics* 31(3): 418–442.

Marsh D (2002) *CLIL/EMILE. The European Dimension: Actions, Trends and Foresight Potential*. Jyväskylä: UNICOM, Continuing Education Centre.

Marsh D, Maljers A and Hartiala AK (2001) *Profiling European CLIL Classrooms: Languages Open Doors*. Finland: University of Jyväskylä and The Netherlands: European Platform for Dutch Education.

Merino JA and Lasagabaster D (2018) The effect of content and language integrated learning programmes' intensity on English proficiency: A longitudinal study. *International Journal of Applied Linguistics* 28(1): 18–30.

Oonk GH (2004) *European Integration as a Source of Innovation in Education*. Alkmaar/Den Haag: Europees Platform.

Paran A (2013) Content and language integrated learning: Panacea or policy borrowing myth? *Applied Linguistics Review* 4(2): 317–342.

Pérez-Cañado ML (2012) CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism* 15(3): 315–341.

Pérez-Cañado ML (2018) CLIL and educational level: A longitudinal study on the impact of CLIL on language outcomes. *Porta Linguarum* 29: 51–70.

Pérez-Cañado ML and Lancaster NK (2017) The effects of CLIL on oral comprehension and production: A longitudinal case study. *Language, Culture, and Curriculum* 30(3): 300–316.

Pérez-Vidal C and Roquet H (2015) CLIL in context: Profiling language abilities. In: Juan-Garau M and Salazar-Noguera J (eds) *Content-based Language Learning in Multilingual Educational Environments*. Berlin: Springer International Publishing Agency, pp.237–255.

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

Roquet H and Pérez-Vidal C (2017) Do productive skills improve in content and language integrated learning contexts? The case of writing. *Applied Linguistics* 38(4): 489–511.

Ruiz de Zarobe Y (2008) CLIL and foreign language learning: A longitudinal study in the Basque Country. *International CLIL Research Journal* 1(1): 5.

Ruiz de Zarobe Y (2011) Which language competencies benefit from CLIL? An insight into applied linguistics research. In Ruiz de Zarobe Y, Sierra JM and Gallardo del Puerto F (eds) *Content and Foreign Language Integrated Learning: Contributions to Multilingualism in European Contexts*. Bern: Peter Lang, pp. 129–153.

Rumlich D (2017) CLIL theory and empirical reality: Two sides of the same coin? *Journal of Immersion and Content-Based Language Education* 5(1): 110–134.

San Isidro X and Lasagabaster D (2018) The impact of CLIL on pluriliteracy development and content learning in a rural multilingual setting: A longitudinal study. *Language Teaching Research*. Epub ahead of print 22 January 2018. DOI: 10.1177/1362168817754103.

Sylvén LK (2010) *Teaching in English or English teaching? On the effects of content and language integrated learning on Swedish learners' incidental vocabulary acquisition*. PhD Thesis, University of Gothenburg, Sweden.

Van der Leij A, Bekebrede J and Kotterink M (2010) Acquiring reading and vocabulary in Dutch and English: The effect of concurrent instruction. *Reading & Writing* 23(3–4): 415–434.

Verspoor M, De Bot K and Xu X (2015) The effects of English bilingual education in the Netherlands. *Journal of Immersion and Content-Based Language Education* 3(1): 4–27.

Westhoff G (1994) *Tweetalig onderwijs in de praktijk. Verslag van een studiereis*. Report, Utrecht: WCC.

Wolff D (2007) CLIL: Bridging the gap between school and working life. In: Marsh D and Wolff D (eds) *Diverse Contexts – Converging Goals: CLIL in Europe*. Frankfurt am Main: Peter Lang, pp.15–25.

Author biographies

JA Goris was born in Bergen op Zoom, the Netherlands. She trained to be a teacher of English as a Foreign Language and worked in secondary and adult education. She obtained a master's degree in english language and literature at Radboud University Nijmegen in 2001 and taught english literature to an international classroom in the U.K. In the past decade she conducted research into content and language integrated learning in secondary pre-university education in various european countries. The results are reported in her dissertation which was published in 2019. At present she is active in the field of integrating academic content and language in english - taught programmes in higher education and trains professionals in various disciplines, mainly health care and L2 pedagogy.

EJPG Denessen is professor in Socio-Cultural Diversity and Education at Leiden University and an associate professor in the Department of Education at Radboud University, Nijmegen. He is specialized in research on educational inequalities with a focus on teachers and parents as relevant actors in this respect.

LTW Verhoeven is professor in communication, language and literacy in the Behavioural Science Institute at Radboud University Nijmegen. The focus of his research is on language, literacy and science learning in typically and atypically developing children in culturally and linguistic diverse environments. He completed an MA Pyschology and an MA Special Education at Radboud University and received his Phd (honours degree) at the University of Tilburg. He did postdoc studies at the University of California at Berkeley and at the UC at Santa Barbara.