DEPOSIT AND CONSULTATION OF DISSERTATION

One copy of your dissertation will be deposited in ORA (Oxford University Research Archive), where it is intended to be freely available online. In order to facilitate this, you are requested to complete and sign the form below.

Please use block capitals

Surname RICHTER
First names (in full) KAI WALKER
Faculty board EDUCATION
Degree name and pathway MSC IN APPLIED LINGUISTICS AND SECOND LANGUAGE ACQUISITION
Title of dissertation EDUCATIONAL OUTCOMES IN MULTILINGUAL CLIL SCHOOL SETTING: A SYSTEMATIC REVIEW N.B. The title stated here must be precisely the same as that stated on the title page of the thesis submitted. **A candidate wishing to amend the title previously approved by the faculty must apply to the faculty board for permission to do so.**
Supervisor DR VICTORIA MURPHY
Subject keywords *Enter your own keywords or phrases to describe your work. This information helps us describe your work in ORA* SYSTEMATIC REVIEW, CLIL, MULTILINGUAL, IMMERSION
Research methods used *This information helps us describe your work on SOLO for future students e.g. quantitative, interviews, vocabulary test, systematic review, etc.* SYSTEMATIC REVIEW, MIXED METHODS

Declaration by the candidate as author of the dissertation

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I understand that the Department requires that I shall deposit one copy of my dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals, and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [www.bodleian.ox.ac.uk/ora/deposit-in-ora/deposit-licence] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]: Kai Richter
Date: 19/10/21

Educational outcomes in multilingual CLIL school settings: A systematic review

Kai Richter
Department of Education, University of Oxford
St. Anne's College
Supervisor: Professor Victoria Murphy

Dissertation submitted in partial fulfilment of the requirements for the degree of
Master of Science in Applied Linguistics and Second Language Acquisition

August, 2021

ABSTRACT

This systematic review aims to assess the spread and effectiveness of multilingual school programs that integrate the teaching of academic content knowledge and language. With multilingualism as a growing reality across the globe, such educational models represent a promising option for supporting the development and vitality of majority and minority languages alike. As such, there is a critical need for an evaluation of the outcomes of multilingual programs.

A total of 19 studies were identified by the search, with the majority of studies coming from secondary schools in Spain. Mirroring the results from bilingual school programs, the included studies reported positive outcomes for attitudes and target language development, without any associated costs for the L1 or academic content knowledge.

Though a risk of bias assessment revealed that the group of studies possesses a moderate strength of evidence as a whole, the findings have severely limited generalizability, given the dearth of studies conducted outside of Spanish secondary schools. Nonetheless, analogous patterns between bilingual and multilingual programs and initial findings from the limited studies conducted outside of Spain provide tentative indications that multilingual school programs may be a fruitful avenue forward for developing competence in several languages at no expense to academic performance.

Multilingualism is a trend rapidly developing worldwide, and educational models and research must keep pace with this reality; future research must build on the current body of work and assess the effectiveness of multilingual programs in a broader variety of global contexts.

Acknowledgments

My heartfelt thanks to everyone who has helped me through this arduous process in such an unusual year. To my friends and family, thank you for always being there for me, and answering stressed calls or texts despite many hours of time difference.

Thank you to the incredibly knowledgeable librarians in the Department of Education who helped me devise my search, and to Anastasia and Tianyi for putting in time and effort to be additional reviewers for this project. Thank you to Will for your help with Irish translations.

Professor Murphy, thank you for always pushing me to think critically and strive for excellence, while simultaneously supporting me when I was feeling overwhelmed. Your knowledge and encouragement have been invaluable and inspirational.

To my Oxford friends and ALSLA classmates (not mutually exclusive), thank you for being there for me this year. I will fondly remember many a walk throughout various lockdowns, weekend trips, and even the occasional party. May we have many more suppers together in the future. Thank you to Rosanna for being a pillar of emotional and intellectual support, and for finishing my pancakes when the weights of academia did not allow me to do so myself.

To Robin Cottage, thanks for making a house into a home and putting up with my crazed linguistics rants and much more. To Aitor, thank you for always listening and supporting me at every turn, this dissertation would not be the same without your patience, kindness, and constant support.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BAC	Basque Autonomous Community
CLIL	Content and language integrated learning
CS	Communication strategy
EFL	English as a foreign language
EPPI Centre	The Evidence for Policy and Practice Information and Co-ordinating Centre
IDESR	International Database of Education Systematic Reviews
LOI	Language of instruction
L1/L2/L3	First/second/third language
MMAT	Mixed Methods Appraisal Tool
MOI	Medium of instruction
PICO	Participants, intervention, control, outcomes
PRISMA	Preferred Reporting Items of Systematic reviews and Meta-Analyses
PROSPERO	International Prospective Register of Systematic Reviews
RCT	Randomized control trial
RoB	Risk of bias
SES	Socio-economic status

CHAPTER 1. Introduction

1.1 Background

The world today is more interconnected on a global scale than ever before, and this has led to a dramatic spread of and contact among various languages. As such, multilingualism is the new norm in a wide range of global contexts (Bale, 2010). Even in settings that are often perceived as monolingual, such as the UK, a closer consideration reveals a rich diversity of language and culture. In 2021, there were reported to be 1.6 million school children (19.2%) in the UK with a first language other than English (UK Department for Education, 2021). As such, it is becoming increasingly important for all citizens in our global society to develop linguistic proficiency in more than one language. While the coexistence of multiple languages is a commonplace reality in many parts of the world, such as most regions of Africa and Asia, robust models of multilingual education (as defined in 2.3.2) are unfortunately less ubiquitous (Benson, 2021). Indeed, the continued growth of English as a lingua franca and pressure from this and other colonial or high-prestige languages endanger numerous minority languages (García, 2014). Multilingual education has the potential to address the mounting need for competence in global languages of business and communication, while simultaneously upholding the linguistic vitality and esteem of non-majority languages.

The integration of academic content and language learning in classroom settings has flourished worldwide as an effective form of multilingual education. These programs developed out of a desire to foster bi- or multilingualism in learners from a young age (Marsh, 2000). The growth of these programs is due in large part to evidence of bilingual schools' prevailing success in cultivating high levels of linguistic proficiency in a target language, with no associated academic costs (Genesee, 1987). Various labels have been ascribed to such programs—such as Content and Language Integrated Learning (CLIL), immersion, and Content-Based Instruction—and these models play out at different levels ranging from nursery school to higher education (Ruiz de Zarobe & Jiménez Catalán, 2009). These disparate terms overemphasize what are generally trivial differences in comparison to the underlying commonalities and frustrate attempts to share findings across similar contexts. This dissertation employs the term CLIL as an umbrella for programs that integrate academic content and language teaching (see 2.1.1).

1.2 Rationale

Multilingualism grows as a prevailing reality across the globe, however most educational programs do not currently reflect this phenomenon. CLIL as a form of language education offers a promising avenue forward. Nonetheless, despite largely positive empirical evidence reported on CLIL programs, the field as a whole suffers from methodological weaknesses in research, a geographic bias in reporting, and crucially, a dearth of studies investigating educational outcomes in multilingual schools. This dissertation adopts a systematic review approach to assess the totality of research conducted on the effectiveness of multilingual CLIL programs on a global scale and to evaluate the strength of these findings.

1.3 Outline

In Chapter 2 of this dissertation, the definition of the constructs of CLIL and a multilingual school are articulated, and the unique contribution of this systematic review is elaborated. Subsequently, Chapter 3 outlines the methodology and design of the review, while Chapter 4 illustrates patterns of findings from the identified studies. Chapter 5 delves into potential causes and implications of the findings and offers recommendations for the future of research on multilingual CLIL.

CHAPTER 2. Literature review

2.1 Content and Language Integrated Learning

2.1.1 A typology of CLIL

At its core, CLIL as a model of education has been defined as the dynamic fusion of content-specific and additional language teaching, or learning a subject through a language other than the L1 (Coyle et al., 2010). In practice, this can take shape in one of many formats across educational institutions, ranging from teaching 90% of subjects in a target language to teaching one non-language subject in the target language several times a week (Dalton-Puffer et al., 2010). The first modern iteration of bilingual education was developed in Canada in the mid-1960s, with the advent of French immersion programs geared towards majority English learners (Murphy, 2014). In response to the widespread success of French immersion (see 2.2.1), similar educational models have proliferated worldwide, with particular growth in the past several decades (Maljers et al., 2007).

Given the diversity of models arising depending on local contexts, Lasagabaster and Sierra (2010) argue that different regional programs should be viewed as distinct entities. In particular, they contend that CLIL, as has uniquely been developed in Europe, is different from North American immersion models with respect to the language of instruction (LOI), teaching staff, starting age, teaching materials, learning objectives, and inclusion of immigrant students. Cenoz et al. (2014) dissent, questioning the claim that CLIL and immersion are disparate conceptualizations and citing counterexamples to argue against the supposed criteria that differentiate programs in North American and European schools. For instance, Lasagabaster and Sierra (2010) claim that immersion programs all begin when a learner is still in primary school, whereas CLIL students in Europe start much later. Cenoz et al. (2014) highlight, in contrast, that late immersion models exist in North America, as do early CLIL programs across Europe. Cenoz (2015) goes further to reject the categorical distinction between CLIL and immersion, claiming that this arbitrary division prevents the sharing of vital research between the two systems. While there are features of CLIL in Europe that differ from North American immersion, or other bilingual models worldwide, these dissimilarities are generally trivial when compared to the similarities (Cenoz, 2015). To draw categorical divisions between such comparable programs hinders the ability to share findings from research in different contexts across the globe, and so it is counter-

productive to argue that CLIL and immersion are completely distinct forms of bilingual education.

In an effort to synthesize relevant findings from a range of contexts, this dissertation will consider a broad variety of educational programs, including both CLIL and immersion. The general strengths and weaknesses of these varied programs are discussed in section 2.2. While it is true that modern versions of bilingual education developed out of the tradition of Canadian immersion (Dalton-Puffer et al., 2010), this dissertation opts to use CLIL as an umbrella term for all these programs, with immersion as an important subset of the whole. Because immersion programs, those that offer at least 50% instruction in the target language, typically result in stronger benefits for students than less intensive bilingual programs (see 2.2.1), it is useful to be able to identify immersion as a sub-class within the wider CLIL spectrum of programs. Different investigators employ diverse terminology to refer to these programs, and this review chooses to use CLIL for its widespread adoption across multiple global contexts (Banegas et al., 2020).

2.1.2 Key concerns in the evaluation of CLIL programs

Before summarizing the findings from various programs, it is important to first highlight several central concerns in the field of CLIL research that have implications for optimal research methodology. Bruton (2011) argues that in most contexts CLIL is an elitist model comprised of only the most talented students. In many cases, especially in Europe, students must apply for entry into CLIL programs, with selection based on general academic performance or admissions tests (Pérez Cañado, 2020). In other cases, any student is welcome to opt into such a bilingual program, though researchers point to frequent examples where students enjoy greater levels of parental support and come from a higher social class than their non-CLIL peers (Bruton, 2011). Investigating the self-selecting nature of CLIL programs in a Dutch secondary school, Verspoor et al. (2015) found that CLIL sections of students performed significantly better in target language testing than the non-CLIL sections within the same school; however, there were no significant differences between the CLIL group and students from a second, selective school that did not offer CLIL classes. The authors contend that the findings comparing CLIL students with other high-aptitude non-CLIL students indicate that CLIL pedagogy itself was not responsible for the positive outcomes within the bilingual school, but rather that the separation of students into CLIL and non-CLIL sections was

a form of streaming that merely selected high-performing students. Thus, positive findings from CLIL programs may reflect existing differences related to factors such as aptitude, motivation, or family socio-economic status (SES) rather than inherent superiority of CLIL pedagogy. Consequently, in the last decade there has been a call for more longitudinal studies that account for potential baseline differences between groups.

In addition to selection factors, Graham et al. (2018) suggest that the L2 benefits demonstrated in empirical research on CLIL could stem from increased exposure to the language, as CLIL students often receive many more hours of instruction in the target language. However, even if the foreign language performance advantage reported for CLIL students stems from nothing more than additional hours spent in the language, this still represents a merit of CLIL systems. That is to say, the effectiveness of CLIL lies in its ability to increase a learner's exposure to a foreign language without decreasing instructional time for other subjects or drastically altering the school day (Coyle et al., 2010).

Thus, due to the potential baseline differences that are common in CLIL models, in assessing the literature it is especially important to control for potential confounding factors either by employing a longitudinal design or by carefully matching intervention and control groups (Admiraal et al., 2006).

2.2 Educational outcomes in bilingual CLIL programs

2.2.1 Target language development

Many studies investigating CLIL outcomes focus on target language development, and overall, CLIL has presented very positive results in this area. Murphy (2014) emphasizes that immersion programs for majority language learners, models that consist of at least 50% instruction in the L2, are some of the most successful forms of primary bilingual language education. The large-scale Canadian French immersion programs are well-known for their success in developing L2 language and literacy skills (Lambert & Tucker, 1972). Genesee (1987), for example, synthesized results from various studies conducted across Canada in total French immersion programs in primary schools and reported that immersion students generally did not present significant differences from native French-speaking controls in their comprehensive skills of reading and listening. However, the immersion learners still scored significantly lower in areas of writing and speaking and tended to present more

grammar and pronunciation errors in these productive domains. Despite these shortcomings, immersion students still regularly outperformed groups that only received limited hours of traditional French language instruction across all language competencies, and at no academic cost (see 2.2.2, 2.2.3). The strengths of this work include the comparison of longitudinal results from over ten studies and the provision of both native-French and non-immersion controls to assess significant between-group differences.

In contrast to immersion models that involve at least 50% instruction in the target language, some programs offer less exposure, typically a couple of additional hours per week beyond the traditional target language classes. Several systematic reviews have reported findings of less-intensive CLIL models and allow for an analysis of the effectiveness of these programs. First, in a review of 25 studies conducted in CLIL schools, Graham et al. (2018) found evidence of equal or superior L2 performance for CLIL students over non-CLIL students. In particular, the findings indicated that CLIL pedagogy may benefit receptive skills (mainly vocabulary) in primary learners, though in secondary school contexts no differences were found in this area. Regarding productive skills, with six studies reporting measures of writing and four reporting measures of speaking, there was no evidence of significant differences between CLIL and non-CLIL sections in overall proficiency, although CLIL students did tend to produce more natural writing and less-accented speech. A major weakness of Graham et al.'s (2018) review is the absence of any reporting of risk of bias or effect sizes of individual studies. Therefore, it is difficult to assess the strength of the evidence provided. Moreover, while the investigators set no geographic parameters in their search, 13 of the 16 studies reporting measures of L2 outcomes in primary or secondary schools were conducted in Spain, and no studies concerned schools outside of Europe. Whether an indication of bias in the search method or a genuine reflection of the limits of existing research, the findings from this study are weakened by the narrow geographic field, and so it is unclear whether CLIL programs in other parts of the world would produce similar results.

Goris et al. (2019) conducted a similar systematic review, though only including longitudinal studies from Europe in the analysis. A total of 21 studies were found, with roughly half of the works presenting generally no differences in L2 proficiency development between CLIL and non-CLIL groups. The other half of studies found

significant benefits for CLIL groups in the varied areas of overall proficiency, vocabulary, grammar, reading comprehension, and listening. Only one study, Pladevall-Ballester and Vallbona (2016), presented a significant disadvantage for CLIL groups, and this was in a Spanish primary school. As with Graham et al.'s (2018) review, Goris et al.'s (2019) search resulted in a high proportion of studies from Spain, 13 out of 21. Interestingly, results from Spain were more likely to indicate significant positive effects of CLIL than studies conducted in other European countries, such as Germany, the Netherlands, or Sweden. The authors contended that Spain is fertile ground for CLIL programs, given that its citizens present on average lower English proficiency than many northern European countries (European Commission, 2012). The findings of Goris et al.'s (2019) review are especially convincing since the investigators specifically targeted studies with longitudinal designs that controlled for potential confounding factors. In addition, the authors provided in-depth descriptions of each included study and reported effect sizes. While they did not include an assessment of the risk of bias of each study, the overall review has more strengths than weaknesses.

Banegas et al. (2020) provided another review of CLIL programs, focusing on educational institutions in Latin America and assessing 64 publications across 11 countries. The authors concluded that there was a dearth of empirical studies conducted in Latin American CLIL contexts, and that the limited body of work did not suggest significant L2 effects for CLIL students. Despite the impressive number of studies included in this review, only 20 of the works actually concerned schools offering the L2 as a medium of instruction (MOI), while the remainder of the investigated schools merely implemented task-based learning in English as a foreign language (EFL) classes. Moreover, the results of individual studies were not described in detail in this review, nor were effect sizes included. Given the small proportion of included studies that would qualify as CLIL as described in 2.1 and the poor reporting of results, it is difficult to glean noteworthy findings from this study regarding the effectiveness of CLIL in Latin America.

In sum, CLIL programs with a large number of instructional hours (immersion) typically present strong L2 outcomes, with hardly any studies reporting negative effects of CLIL on target language development. Immersion students often outperform peers in non-immersion sections across all language skills, but still lag behind native learners in productive grammar and pronunciation skills, in particular. In contrast, the results

from CLIL programs that have fewer instructional hours are more varied, with CLIL students only outperforming non-CLIL students in certain areas. A large body of research has attested to benefits of CLIL in the general productive skills of speaking and writing, though the findings for listening are more neutral (Merino & Lasagabaster, 2018). Furthermore, it may be the case that CLIL programs have more benefit in places where the target language is not as strongly developed, such as Spain. Finally, very few studies report any detrimental effects or disadvantages for CLIL students relative to non-CLIL learners regarding target language outcomes (Admiraal et al., 2006).

2.2.2 L1 development

Upon implementing a CLIL program, a common concern among various stakeholders is that additional time in a target language will negatively impact L1 development (Merino & Lasagabaster, 2018). As a young learner's first language is not fully matured, some parents and educators question whether less time in the L1 would potentially be detrimental to typical L1 development in language and literacy skills. Nonetheless, this concern has proved largely unfounded, as CLIL students regularly display L1 proficiency levels on par with their non-CLIL peers (González Gándara, 2015). Most notably, even total immersion programs where students receive only 10% L1 instruction have shown this pattern of parity in L1 performance among groups after several years' participation in the program; a finding that has been widely replicated across North American immersion programs (Genesee, 1987). For instance, in Padilla et al.'s (2013) study of a Chinese immersion primary school in the US, while the mean scores for immersion students on English standardized testing were lower than their non-immersion peers in the first year of the program, significant differences in scores were not found five years later. Padilla et al.'s (2013) assessment of English language skills involved a state-wide standardized test and a non-immersion control group within the same school. Immersion section students were admitted based on a lottery system, and so the non-immersion students in the same school were more likely to be matched on factors such as motivation and parental support. As such, the findings regarding L1 proficiency from this study strongly support the notion that CLIL methodology does not have negative effects on L1 skills.

2.2.3 Content knowledge development

A question with less consistent results has been that of the impact of learning a subject through an L2 on the subsequent development on content knowledge in that area.

Graham et al.'s (2018) systematic review found that two studies reported positive benefits of CLIL when students were studying math in the L2, while a further two studies found that CLIL methodology resulted in worse outcomes when students learned a science subject through the L2. In one of the latter two studies, Fung and Yip (2014) found in a secondary school in Hong Kong that students with weaker L2 proficiency fared better with regard to acquisition of physics knowledge when instructed in the L1, while high-achieving students performed better when taught through the L2. They concluded that gifted students received motivational benefits from CLIL instruction, but that instruction in the L2 had a detrimental effect for students identified as low achieving. Lo and Fung (2020) contend that the language of assessment itself poses a great challenge to learners with low L2 proficiency, who may grasp the nonlinguistic content but struggle to express complex thoughts in another language. Therefore, worse results in subject knowledge testing may not reflect inferior understanding.

Meyerhöffer and Dreesmann (2019) investigated whether CLIL pedagogy impacted biology content knowledge gains for German secondary students and concluded that teaching the subject in L2 English had no negative impact on subject-specific learning. One of the greatest strengths of this study was the inclusion of three groups for comparison: one mainstream control group and two CLIL treatment groups, pre-selected gifted students and non-selected mainstream students. The lack of significant differences among all groups suggests that the CLIL model was effective in developing students' science knowledge independent of student selection factors.

At the primary level, Fernández-Sanjurjo et al. (2019) investigated science outcomes in 709 CLIL and non-CLIL students in Spain and found that the CLIL groups performed worse on tests. However, a major weakness in the study is that all students were tested in their L1, despite the CLIL group having learned the material in their L2. Students regularly perform worse when there is not a match between the LOI and the language of testing (Murphy, 2014), and so this represents a large source of bias against the CLIL group in this study. In contrast to the above results, Hughes and Madrid (2020) found in their study of Spanish students' science knowledge that non-CLIL groups slightly outperformed the CLIL groups in primary school, but that this effect was not present for the secondary school cohort, where the CLIL groups excelled. They posited that learners in a CLIL system may experience a slight delay in acquiring academic

language, but that this disparity lessens over time. This study offered a robust sample size of 232 students and carefully considered various confounders such as verbal reasoning skills, motivation, SES, English exposure outside of school, and urban vs. rural context. As such, the evidence provided by Hughes and Madrid (2020) is very compelling.

Overall, the evidence regarding nonlinguistic content knowledge outcomes for CLIL students is mixed, though it appears that deficits may be largest for young learners, but that these may diminish after several years.

2.2.4 Attitudes

A final outcome to consider in CLIL programs is the impact of this form of pedagogy on students' attitudes towards non-native languages and to multilingualism itself. Marsh (2000) claims that CLIL nurtures confidence in language skills and opens doors to wider interest in other languages and cultures. In a study of 788 Spanish or Chinese immersion students in the US, Lindholm-Leary (2016) reported that students presented very positive attitudes towards the language they studied and a near unanimous agreement (96%) that their control of the target language was better than their non-immersion peers'. While these responses represent subjective beliefs, and are not necessarily reflective of actual proficiency differences, it is still notable that students in these dual-language programs held such strong convictions. In Finland, Merisuo-Storm (2007) found in a study of 145 primary students that the pupils in CLIL sections had significantly more positive views towards learning a foreign language than their non-CLIL peers. The inclusion of a comparison group in this study helps to reinforce the findings of Lindholm-Leary's (2016) work. While CLIL exists in a variety of contexts, the overall picture regarding language attitudes is that integration of language and content has positive effects in this domain.

2.3 Multilingualism in schools

Much of the research surrounding CLIL has centered on bilingual education, that is schools that focus on the development of two languages (Dalton-Puffer et al., 2010). Less research has targeted CLIL programs geared towards the successful development of three or more languages in their students. The outcomes of multilingual education are of critical importance now, given the growing interconnectedness of our globalizing world; the ability to speak more than one language is not only a pressing need for many, but also the lived reality of much of the population worldwide (Genesee, 2008).

Therefore, the effectiveness of programs that seek to develop multilingualism must be evaluated, as the relevance of such models continues to grow across an abundance of geopolitical contexts. The following discussion first establishes a working definition for a multilingual CLIL school and then emphasizes this review's focus on multilingual program design.

2.3.1 Additive multilingualism in educational programs

Language education programs take a variety of forms worldwide, and different models can have divergent outcomes. One major distinction in the design of a program that teaches more than one language is the focus on so-called subtractive or additive bi- or multilingualism (Baker & Wright, 2017). In a subtractive bilingual program, the teaching of an L2 effectively functions to supplant or demote the status of the L1 (Lambert, 1981). This practice is exemplified by transitional or submersion models of education. In a transitional bilingual program, a student's L1 is used initially to support learning, but is slowly reduced and eventually replaced by exclusive L2 instruction (Skutnabb-Kangas & Heugh, 2013). Critically, the L1 is not assigned equal value and is only used as a scaffold until the learner transitions to a mainstream L2 classroom. As such, the program does not strive for bilingual proficiency. A submersion program involves instruction exclusively in the L2, regardless of a learner's proficiency in said language. These forms of subtractive bilingual programs are common for many immigrants or minority language speakers who are pressured to assimilate to a majority culture and learn a majority language (Crawford, 2008). In contrast, in an additive bilingual or multilingual context, all languages are viewed as valuable and the learning of one need not take away from the position of another (Cenoz & Valencia, 1994). Additive programs teach in more than one language and aim at improving literacy and overall proficiency in all the LOIs; as such, an additive view towards languages is seen by many experts to be a key component of CLIL education (Cenoz, 2015).

Additive bilingual programs have been widely demonstrated to be the most successful for development in both languages (Genesee, 2006). Thomas and Collier (2002) conducted a five-year study on the educational outcomes of English language learners in the US. The researchers collected data from five regional sites, comprising eight different bilingual program structures and a total of 210,054 students ranging from early primary to end of secondary school. The results indicated a decisive benefit of immersion programs, as these programs exhibited the best results for the L1, the L2,

and all academic content subjects. The impressive sample size, selection of varied contexts in rural and urban areas, and the longitudinal design indicate that these findings are highly robust. In contrast, subtractive models not only hinder L1 growth, but also deliver weaker results in L2 progression when compared to additive bilingual programs (Menken & Kleyn, 2010). Even though transitional programs can present positive results in the short term, a longitudinal analysis reveals that students in these programs suffer when the L1 is removed (Thomas & Collier, 2002).

2.3.2 Defining the multilingual school

Though a multilingual school must aim to support proficiency and literacy in more than two languages, not all of these languages need be MOIs. In many immersion programs, the majority language is used 10% of the time, only in language arts lessons (Baker & Wright, 2017). An important distinction here is the difference between majority and non-majority language. A majority language is the language of the wider community, and typically the language of education (Murphy, 2014). The categorization of majority and non-majority languages is not always straightforward, though an overview of what could qualify as a non-majority language helps to elucidate this distinction. Bale (2010) uses the term heritage language to encompass a wide variety of non-majority languages, including minority, regional, indigenous, colonial, or immigrant languages, among many others. Murphy (2014) outlines various conceptualizations of heritage languages and identifies their salient characteristics, noting that a heritage language may be a minority language when compared to the majority language in society, even if it is not ethnolinguistically a minority language on the global scale. Instances of this type of heritage language include Spanish in the US, Turkish in the Netherlands, and Urdu in the UK. Murphy (2014) continues to identify another form of heritage language, a regional minority language that is not spoken by the wider population and is not the language of power, governance, or formal education, such as Cree in Canada or Māori in New Zealand. Another type of non-majority language is the foreign language, a language that is neither indigenous nor the language of a significant minority or majority population. While a language, such as Chinese in Australia, may be a heritage language for some, for others it would be a foreign language.

This differentiation between majority and non-majority languages is important, as research has shown that limited instruction in a majority language (within an additive program) is sufficient to foster proficiency in the language (Cenoz, 1998; Etxeberria,

1999; Genesee, 2006). A relevant example comes from the Basque Autonomous Community (BAC) in Spain, where Basque (a regional minority language) and Spanish (the majority language) coexist in society. Students in Basque immersion schools have been shown to develop high levels of Basque proficiency at no cost to their Spanish skills when compared to non-immersion peers (Sierra & Olaziregi, 1989). The converse is not true, and students who receive limited instruction in Basque do not achieve the same linguistic proficiency in this language as those students in Basque immersion, even if Basque is the language spoken at home. This review's definition of a multilingual school is therefore: a school that promotes additive multilingualism by teaching more than two languages, where any non-majority languages are used as MOIs and a majority language can optionally be used as an MOI or taught as a subject.

2.3.3 Multilingual pedagogy: design vs. implementation

Some theorists critique the conceptualization of bilingual programs as additive or subtractive, arguing that this dichotomous view is too linear to capture the dynamism of bilingualism (García, 2009). To formulate a more flexible characterization of bilingualism, García and Wei (2014) recommend the concept of translanguage, or translanguaging, to describe the practices of bilinguals. According to the theory of translanguage, there are no clear boundaries between the languages of bilinguals, and so classroom practices in bilingual education need not strictly enforce separation between LOIs, but rather utilize the full linguistic tools available to students. Thus, translanguaging as a pedagogical tool emphasizes the systematic and intentional use of more than one language in the classroom (Williams, 2001). In the past decade, several lines of research have explored the impact of translanguaging in the classroom (e.g. Cenoz, 2017; Duarte, 2020).

The theory of translanguage does well in drawing attention to the complex reality of many bilinguals and emphasizing that in practice most bilinguals regularly mix their languages. However, language mixing in the classroom as a pedagogical tool is not mutually exclusive with establishing separate LOIs in a school (San Isidro, 2017). A full discussion of translanguaging as a teaching practice is beyond the scope of this study and more empirical research in this area is sorely needed. Instead of stressing the ways in which teachers implement CLIL instruction in their classrooms, this review investigates overall program design at the school level.

2.4 The spread of multilingual schools

Having established the criteria for a multilingual CLIL school, this section reviews the contexts in which multilingual education frequently develops. Other systematic reviews have focused exclusively on bilingual schools and reported few studies conducted outside of North America and Europe, and so this dissertation aims to broaden the review of research and address the critical needs of multilingual populations around the world. The purposes of this section are twofold: 1) to highlight the prevalence of multilingual regions worldwide and therefore the importance of assessing the effectiveness of multilingual programs; and 2) to review geographic contexts with multilingual education practices where one might expect to find empirical research studies.

Multilingual education proliferates in geographic areas that are already bi- or multilingual. In trilingual areas, educational programs are geared toward the whole population, as students regularly experience multilingualism outside of school as well (Ytsma, 2001). An example of this on the national scale is Luxembourg, a multilingual country where Luxembourgish, German, French, and English are commonly taught in school and used in the wider society (Baetens Beardsmore & Lebrun, 1991). In the Asian context, trilingual education exists in parts of China, such as with Tibetan, Mongolian, and Mandarin in the Qinghai province (Dai & Cheng, 2007). Furthermore, India is a multilingual country where the national language of Hindi coexists with the colonial language of English and a wide variety of regional minority languages (Mohanty, 2006). While the simultaneous use of three LOIs in India is rare, in 2011 there were a total of 31 languages used as MOIs across the country, revealing the multilingual diversity of the nation (Meganathan, 2011). The situation is similar in Sri Lanka (Wedikkarage, 2018) and Hong Kong (Li & Tong, 2020), where local languages—Sinhala and Tamil in Sri Lanka and Cantonese and Mandarin in Hong Kong—are used alongside English in society and in schools. With students immersed in such diverse communities, it is imperative that educational institutions keep pace with the multilingual lived reality of their pupils. Unfortunately, many countries' educational policies have lagged behind the dynamic multilingualism of their citizens; this pattern is especially prevalent in parts of Africa and the Asia-Pacific region where a multitude of minority languages abound (Benson, 2021; Heugh et al., 2012; Shameem, 2007).

In bilingual regions, schools that had a history of teaching two local languages are now often incorporating a third, global language as an LOI; due to the international spread of English as a lingua franca, this is frequently the additional language (Hélot & de Mejía, 2008). Trilingual education with two local languages and one foreign language has been cited in a broad range of contexts, for instance in Kazakhstan (Kalizhanova et al., 2021), Paraguay (Ministerio de Educación y Ciencias, 2018), and Estonia (L'nyavskiy-Ekelund & Siiner, 2017), among others. In the European Union, monolingualism is far from the norm, with 54% of citizens able to speak at least two languages and 25% of the population able to speak at least three (European Commission, 2012). Given this linguistic context, it is not surprising that seven countries (Austria, Estonia, Latvia, Luxembourg, the Netherlands, Spain, and Sweden) offered a trilingual CLIL provision as early as 2006 (Eurydice, 2006).

In contrast to the aforementioned environments, multilingual education is not common in geographic zones where one language is predominant (Ytsma, 2001). In particular, majority English-speaking countries such as the UK and the US do not have many multilingual programs. While often perceived as monolingual, countries where one majority language predominates often have very multilingual or multicultural populations, though national educational programs rarely cater to the variety of languages present in the population (Baker & Wright, 2017). One instance of a multilingual program in a country with English as the majority language is a trilingual school offering English, Spanish, and Yaqui (an indigenous Mexican language) in the southwest United States (Trujillo, 1997). The example of this school illustrates that even countries with one official or majority language can still have multilingual communities, with programs developed according to their linguistic needs.

The choice of MOI in schools is currently at the forefront of language-in-education policy discussions worldwide, and so the educational outcomes of multilingual programs have significant implications for the potential future growth of CLIL (Lightfoot et al., 2021).

2.5 Gaps in the literature

The importance of multilingual education cannot be understated. In our rapidly globalizing society where bi- or multilingualism is the growing reality for most of the world, providing adequate education to foster linguistic proficiency in more than one language is vital (Clyne et al., 2004). In addition, the spread of English threatens other

languages worldwide, especially in multilingual communities that may feel obligated to choose this foreign language over other heritage languages as an MOI in schools (Bale, 2010). CLIL as an educational model has the potential to promote multilingual development, without sacrificing minority languages. Nonetheless, despite the prevalence of multilingual regions worldwide and many schools that employ multiple LOIs, most research on CLIL has been conducted in North America and Europe (Banegas et al., 2020); it is unclear to what extent empirical studies have evaluated educational programs in other settings, or whether findings as to the effectiveness of CLIL are consistent across a broad range of geopolitical contexts.

As reviewed in 2.2, bilingual CLIL programs have largely demonstrated positive outcomes in L2 development and attitudes towards other languages and cultures, while presenting no significant deficits for the development of L1 skills or academic content knowledge. However, critics question the reliability of such findings, cautioning against the elitist or self-selecting nature of CLIL programs and highlighting the need for more rigorous methodologies that control for differences between CLIL and non-CLIL cohorts (Bruton, 2011). Moreover, other researchers suggest that the benefits of CLIL stem purely from the amount of exposure to the foreign language, pointing to the strong findings from total immersion programs as evidence of a time-on-task advantage (Graham et al., 2018). If the quantity of exposure were genuinely the crux of CLIL programs' success, this would have major implications for multilingual models, in which educators must balance instructional time in more than two languages. Given the fixed number of hours in a school day, additional exposure to a given language concomitantly reduces exposure to the other languages of the school. Whether or not CLIL programs still produce positive results in a multilingual environment has yet to be satisfyingly addressed by research (Merino & Lasagabaster, 2018).

Finally, no systematic review to date has assessed the extent of empirical research on multilingual schools or synthesized the findings from such work to evaluate the effectiveness of CLIL methodology in multilingual environments (see Chapter 3). This review aims to contribute to the research on multilingual educational programs by assessing the scope of research conducted in a broad range of settings and appraising the findings of such work.

2.6 Research questions

In order to address the gaps in the literature detailed in 2.5, this systematic review proposes the following research questions:

1. What is the extent and nature of empirical research investigating educational outcomes in multilingual CLIL school settings (as per the specified inclusion/exclusion criteria)?
    a. What is the balance of research in primary vs. secondary educational contexts?
    b. What is the geographic profile of the literature on multilingual schools?
2. What is the effectiveness of CLIL in multilingual schools?
    a. What are the language outcomes—L1, L2, or L3?
    b. What are the academic content knowledge outcomes?
    c. What are the language attitudes or other stakeholder perspectives?

CHAPTER 3. Methodology

The research described in this dissertation employed a systematic review design, which included an in-depth search of the relevant literature as a means of answering the outlined research questions. A systematic review follows rigorous methodology to determine the extent of literature in a given field and to synthesize the sometimes-varied results of said work, providing a more complete picture than any one study can do alone (Gough et al., 2017). The traditional literature review, while informative, is subject to the bias of the author, as he or she is free to pick and choose what research is deemed to be most relevant. With a clear protocol established before conducting the search, systematic reviews objectively capture a wider body of work, insofar as the review is well designed to identify relevant research (Boland et al., 2014). Moreover, due to the transparent nature of the systematic review, it is open to replication whether to corroborate findings, include newly published research or adjust where shortcomings are found (Petticrew & Roberts, 2006).

Prior to conducting this review, a search for other systematic reviews on this topic was performed in the International Database of Education Systematic Reviews (IDESR, n.d.), the PROSPERO database (The University of York, n.d.), the EPPI Centre Database of Educational Research (EPPI-Centre, n.d.) and the Campbell Collaboration Online Library (Campbell Collaboration, n.d.). The search terms of 'CLIL,' 'immersion,' or 'third language' resulted in a total of 213 reviews that had been published or were in progress, however no previous systematic reviews matched the aims of the current study. Of the reviews found, 195 were not related to educational programs or language study, and the remaining 18 did not focus on CLIL in a multilingual educational context. Thus, it is reasonable to assume that the current research does not reproduce work conducted elsewhere, whether published or in progress.

In order to avoid replication of work and to establish transparency and accountability of the review design, the protocol was registered with IDESR on May 30th, 2021 (see Appendix A for protocol document). The initial protocol was completed before conducting the final search, however due to time management issues it was submitted after the search had been conducted, but before data extraction or analysis of the results. Any alterations to the protocol document between initial drafting and submission were minimal and consisted mainly of proofreading and formatting changes.

3.1 Eligibility criteria

The following criteria (Table 1) were established to identify whether a study would be included in this review. The items highlighted for inclusion and exclusion were the bibliographic information, date of publication, participant background, educational intervention, and measures of outcomes. In contrast, the publication status, study design and language of publication were not used as eligibility criteria, as elaborated further in section 3.2. Table 1 details complete inclusion and exclusion criteria, as well as a rationale for each point.

Table 1. Eligibility criteria

ITEM	INCLUSION CRITERION	RATIONALE
Bibliographic information	**Include 1:** Studies with a full reference or sufficient information.	Without sufficient bibliographic information, retrieval of works is infeasible.
	Exclude 1: Studies with insufficient bibliographic information.	
Date of publication	**Include 2:** Studies published on or after January 1, 1994.	The term CLIL was first used in the European context in 1994 (Nikula, 2017). While other forms of content and language integrated learning existed before this time, the European Union's renewed focus on language education increased research efforts in the field, particularly in multilingual contexts. This study aims to assess modern research in this area.
	Exclude 2: Studies published before January 1, 1994.	
Participants	**Include 3:** Studies on typically developing foreign language learners. Include studies even if no explicit reference is made to learning ability if reasonable assumption can be made that participants are comprised mainly of typically developing individuals.	This review seeks to assess effectiveness of CLIL methodology as applies to typically developing school populations. The findings for non-typically developing individuals may not hold for a larger population, and so these latter results should not be extrapolated, nor will they be included in this review.
	Exclude 3: Studies that exclusively target non-typically developing learners or learners with Developmental Language Disorder.	

	Include 4: Studies conducted in primary or secondary schools (students aged 5-18).	The language learning capacity and the educational goals for very young (under 5) or adult (over 18) learners are quite different from those learners in primary or secondary education. This study focuses on the outcomes of CLIL programs in these middle year programs, where learners are well suited to learn another language.
	Exclude 4: Studies in early years, university, or adult educational contexts.	
Intervention	**Include 5:** Studies involving schools where three languages are used as LOIs for 2+ hours a week beyond traditional language arts classes. Alternatively, one of three languages may only be offered as a language arts class where it is a majority language.	The focus of this study is on additive multilingual CLIL programs, where schools actively promote language proficiency and literacy in more than two languages (see 2.3.1). Schools offering only two LOIs may be included with the understanding that a national majority language can be sufficiently supported in language arts classes.
	Exclude 5: Studies on schools that support language proficiency and literacy in fewer than three languages for all students.	
	Include 6: Studies where learners received regular educational intervention (CLIL program) for at least one year.	This study aims to highlight the long-term impact on educational outcomes for students who partake in CLIL educational models. Requiring at least a year of exposure allows for the effects of the pedagogical form to take shape.
	Exclude 6: Studies where participants only engage in short-term CLIL projects or units.	
Outcome	**Include 7:** Primary empirical research studies reporting any measure of CLIL program effectiveness, including but not limited to language outcomes, content knowledge outcomes, or stakeholder attitudes. Include studies reporting either quantitative or qualitative outcomes.	A synthesis of empirical findings in this field of literature is impossible without the reporting and evaluation of concrete data.
	Exclude 7: Systematic reviews and studies that provide narrative evaluation of an educational program but do not provide empirical measures of program outcomes.	

3.2 Items not included as eligibility criteria

Three important items were considered and ultimately rejected as eligibility criteria: publication status, study design, and language of publication. The decision to not initially exclude on these grounds offers the benefit of widening the field of research and presenting a more complete assessment of the literature, though at the same time it may affect the quality of the final pool of research. Nonetheless, it was concluded that the benefit of this broader search outweighed any potential shortcomings.

Due to the inherent risk of a vote counting approach, where a methodologically weak study is given equal weight to that of a stronger study (Petticrew & Roberts, 2006), this review employed Slavin's (1986) best-evidence approach. This method of factoring in the variability in the weight of evidence in individual studies (see 3.8) mitigates against weaknesses or biases introduced by the wider inclusion criteria described below.

3.2.1 Publication status

While the inclusion of gray literature opens the search to research that has not met the same rigorous standards as peer-reviewed published literature, this review sought to limit any effect of publication bias and to ascertain the full extent of work conducted in this field.

3.2.2 Study design

This review recognizes that not all methodological designs yield equally robust results. For instance, within quantitative research, a randomized control trial (RCT) offers greater internal validity than a cross-sectional survey (Petticrew & Roberts, 2006). However, RCTs are rare and often difficult to implement in the social sciences, and in educational research in particular. Critically, with reference to CLIL, the results from studies that omit a non-CLIL control group or that are cross-sectional in nature and do not account for initial selection bias into CLIL programs must be considered with caution (Goris et al., 2019). This caveat that certain research designs may provide misleading results notwithstanding, a significant body of the research into CLIL does employ such methodologies. Thus, this review aimed to capture the totality of research available by not excluding on the basis of study design, though it still considers differences in design and addresses related weaknesses as part of the discussion.

3.2.3 Language of publication

Moher et al. (2003) propose that a systematic review should adopt as inclusive a search as possible with respect to language of publication. The degree of language bias often varies depending on the discipline in question (Morrison et al., 2012). With the understanding that research into CLIL programs will, by definition, take place in multilingual environments and that research may be published in a variety of languages, this study made every effort to assess relevant literature regardless of language of publication. Where a study was written in a language not spoken by the author (i.e. not English, Spanish, Chinese, or Italian), abstracts were initially screened using Google Translate, which provided sufficient detail to assess the inclusion criteria.

For full analysis of papers and data extraction, there was only one paper (Ní Dhiorbháin, 2020) published in a language the author did not speak (Irish). The initial review protocol stated that a proficient speaker of the required language would be consulted for translation. While a full translation was not possible due to logistical constraints, a proficient speaker of Irish was consulted, and they confirmed that the Google translation of Ní Dhiorbháin's (2020) abstract was highly accurate. Therefore, Google Translate was subsequently used for the full analysis of this study.

3.3 Information sources

The list of databases consulted for this systematic review can be found in Table 2, which includes prominent databases in the fields of education, linguistics, psychology, as well as multidisciplinary sources to cover wider fields of the social sciences. Furthermore, searches were conducted on OpenGrey and ProQuest Dissertations & Theses Global to encompass any gray literature not present in the other databases. All databases were accessed electronically via subscription from the University of Oxford's Bodleian Library.

Table 2. List of source databases

DISCIPLINE	DATABASE
Education	ProQuest Education Collection (including ERIC)
Linguistics	ProQuest Linguistics Collection (including LLBA)
Psychology	PsychInfo
Multidisciplinary	Web of Science Scopus
Gray literature	OpenGrey ProQuest Dissertations & Theses Global

The initial search was followed by both forward and backward citation searches, whereby any study that met all the inclusion criteria was reviewed further to find other relevant matches. In the backward citation search, an eligible study's references were reviewed and compared against the selection criteria. For the forward citation search, all studies found on Web of Science to have cited an eligible paper were similarly reviewed.

3.4 Search terms

An optimal systematic review balances sensitivity and specificity, casting a wide enough net that eligible studies do not escape the search, meanwhile ensuring a narrow enough focus that investigators need not sift through a large proportion of irrelevant research (Brunton et al., 2017). The search terms in a systematic review are of paramount importance in establishing this equilibrium.

For this study, an experienced librarian at the University of Oxford's Department of Education was consulted to formulate the initial search. The two elements of mode of instruction (CLIL) and multilingualism were identified as the crux that should underpin the search. Following pilot scoping searches in the Proquest Education Collection, it was found that many results targeted university or adult participants. To account for this, a field specifying target participants was added, narrowing down the focus of results. Several similar labels were included within each of these three categories to encompass the variability of terminology. Other systematic reviews on CLIL outcomes (e.g. Graham et al., 2018) have employed limited terminology, which may have led to the geographic skew of results from European contexts. In an effort to counter this geographic bias of results, a broader range of search terms were included in the section of mode of instruction. Different terms within each category were connected with the operator 'OR' and each field was joined with 'AND.' The final search terms are listed in Table 3.

These search terms produced a manageable number of results. Moreover, a pre-determined exemplar article that met all inclusion criteria, Merino and Lasagabaster (2018), was correctly identified by this final pilot search. As such, it was concluded that an appropriate balance between sensitivity and specificity had been reached.

The first two search fields were applied to abstracts only, after piloting revealed a large number of articles being flagged exclusively due to journal title, such as *Multilingual*

Matters. The field of participant was searched in all fields except full text, where possible.

Table 3. Boolean search terms

Mode of instruction	AND	Multilingualism	AND	Participants
immersion OR CLIL OR "content and language integrated learning" OR CBI OR "content based instruction" OR CBLT OR "content based language teaching" OR EMI OR "English Medium Instruction" OR "language of instruction" OR "medium of instruction"		plurilingual* OR multilingual* OR trilingual* OR L3 OR "third language"		primary OR secondary OR "high school" OR elementary OR adolescent* OR child*

3.5 Data management

Upon completing the final search on May 4th, 2021, all relevant information was uploaded to Rayyan, a software program for systematic reviews where multiple collaborators can compare abstracts and other information against eligibility criteria (Ouzzani et al., 2016). There, duplicates were eliminated and initial screening was conducted. Bibliographic information (title, author, publication date, journal) and full texts of all works not eliminated in the initial screen were uploaded and organized on Mendeley reference manager. Detailed notes documenting the search process and the origin of each study meeting the eligibility criteria were kept in a physical research journal and a Microsoft Excel document, respectively. Data extracted from final reports marked for inclusion was likewise recorded in a Microsoft Excel document.

All research was conducted on a MacBook Air running Mac OS Big Sur Version 11.2.1.

3.6 Selection process

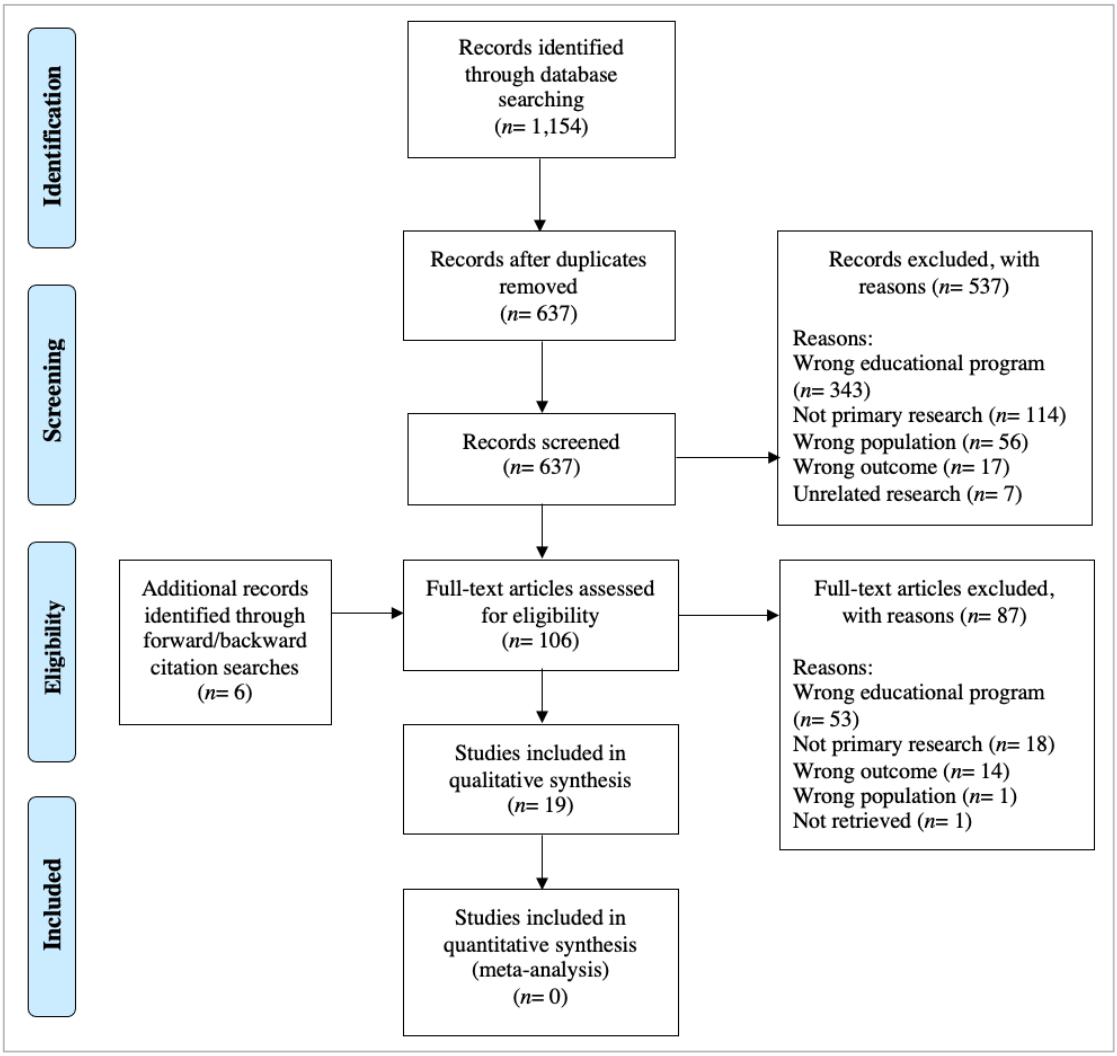
A flow chart overview of the screening process is outlined in Figure 1 below.

3.6.1 Initial screening

The initial search produced a total of 1,154 results, which was subsequently narrowed to 637 after eliminating duplicates. The remaining titles and abstracts were screened against the eligibility criteria and 100 were marked for full-text screening. To be considered for full-text screening, a given abstract must not explicitly violate any of the exclusion criteria, meriting a deeper investigation. In the instance that a given work did not meet a criterion, the screening stopped and the criterion in question was noted. In

total, 537 results did not meet the eligibility criteria and were excluded for the following reasons: Wrong educational program (343), not primary research (114), wrong population (56), wrong outcome (17), unrelated research (7).

Figure 1. Flow diagram of screening process (adapted from Moher et al., 2009)



3.6.2 Full-text screening

Following the initial screening, full reports were obtained from the relevant databases. Where complete texts were not available or where the full report failed to unambiguously satisfy all the inclusion criteria, the author was contacted for further information. Only studies that clearly met all the inclusion criteria were selected for the review, resulting in 13 texts overall. The 87 studies excluded were eliminated for the following reasons: Wrong educational program (*n*= 53), not primary research (*n*= 18), wrong outcome (*n*= 14), wrong population (*n*= 1), not retrieved (*n*= 1). One full text,

Ytsma (1997), could not be retrieved, as the online journal publishing the study could not be located and the author was deceased.

The resulting list was then used for backward and forward citation, and so on, until no further studies were identified. This resulted in an additional 6 works. The final 19 texts reported a mix of outcomes and varied sample sizes and ages tested, and the body of work was deemed too heterogeneous for meta-analysis to be appropriate.

3.6.3 Ensuring screening reliability

To ensure reliability of the main author's screening, a second reader who is well-versed in applied linguistics was recruited and informed of the aims and criteria of this study. The second screener reviewed 10% of abstracts (64). As per the protocol, the second reader also intended to conduct a portion of full text screening, however this was not possible due to time constraints. Both reviewers were blind to the other's decisions at time of screening. Where disagreements occurred, the two screeners conferred to reach a conclusion. A Kappa value of 0.7 is generally considered acceptable in the field of education (Frey, 2018) and was used in this study as a threshold. The interrater reliability for the title and abstract screening process was 0.78. As this figure met the desired threshold, it was concluded that the inclusion and exclusion criteria were reasonably applied and that the initial selection process had a low risk of systematic bias.

3.7 Data extraction

Before executing the complete search, a data extraction form was specifically designed for this study (Appendix B). This document was modeled after the Cochrane good practice data extraction form, chosen in particular for its flexibility and suitability for a wide range of study designs (Cochrane Effective Practice and Organisation of Care, 2017). The data collected encompasses the essential categories of participants, intervention, comparison, and outcomes (PICO; Petticrew & Roberts, 2006), and was reorganized for the purposes of this study. Fields were added to reflect the dynamic language background of the participants' community and to indicate the extent and duration of exposure to CLIL pedagogy.

Upon confirmation that a given study met all eligibility criteria, data was extracted and recorded in an Excel document. Where there were multiple outcomes to be reported,

the relevant section of the data extraction form was copied and pasted to include the additional results.

3.8 Risk of bias of individual studies

Methodological weaknesses in a given study can introduce bias, or a systematic deviation in the reliability of the results (Boland et al., 2014). While it is infeasible to eliminate bias entirely, there are appropriate steps that researchers should take to both limit overall biases in their work and to clearly identify any remaining bias through transparent reporting. Certain study designs, such as RCTs, possess a high degree of internal validity and so are less susceptible to several main methodological biases (selection bias, response bias, attrition bias, and observer bias) than an observational study, for example (Petticrew & Roberts, 2006). Though high internal validity does indicate some degree of methodological quality, it is also necessary to consider the external validity of a study, or the generalizability of findings. An RCT, while highly internally valid, boasts little external validity if conducted with a limited scope, either in number of participants or representativeness of the overall population. Therefore, in conducting a systematic review, one must consider the overall bias introduced by each individual study and weigh the results accordingly.

This review follows Slavin's (1986) best-evidence synthesis method, which does not discriminate against any research design in study selection, but rather assesses methodological rigor and potential bias in its evaluation of the overall work. In other words, studies are not weighed equally in a vote counting method; instead, each study is critically appraised and assigned a value of methodological quality to reflect its internal and external validity (Boland et al., 2014). Given the comprehensible paucity of RCTs in educational research, the more inclusive best-evidence synthesis method avoids the risk of prematurely concluding that no research exists in a field, without sacrificing review quality.

One of the most significant factors when assessing bias is to have tools sensitive to the structure of each research design (Dixon-Woods et al., 2005). Liabo et al. (2017) emphasize the difficulty of assessing within one review disparate methodologies that elicit quantitative and/or qualitative results, and they recommend either two tools tailored to the different designs or one instrument that can flexibly handle mixed methods. As this review actively sought a variety of research designs, it employed an instrument specifically adapted to the needs of mixed methods research, the Mixed

Methods Appraisal Tool (MMAT; Pluye & Hong, 2014). Crowe and Sheppard (2011) found that the MMAT was the only critical appraisal tool that was designed for a systematic mixed studies review.

The MMAT, updated in 2018 with the feedback of over fifty experts to improve content validity, provides five different categories for assessing research methodology: qualitative, quantitative RCT, quantitative non-randomized, quantitative descriptive, and mixed methods (Hong et al., 2019). The tool contains five criteria across each of the five categories, which are marked 'yes,' 'no,' or 'can't tell,' and an additional comment box to justify ratings. A study is evaluated in each relevant category: only once if the study reports only quantitative or qualitative findings and three times if the study reports both types of findings. A template of the MMAT and a user guide can be found in Appendix C. The final search revealed 18 studies reporting quantitative data, none of which were RCTs. Moreover, all of these studies were observational studies where the investigators did not assign intervention, but rather inspected results from intact classes. Two studies (Herraiz Martínez & Sánchez Hernández, 2019; Wang & Kirkpatrick, 2013) were classified as quantitative descriptive studies as they did not have control groups, however otherwise they were analogous to the other studies. Therefore, a post-hoc decision was made to include only three core sections of the MMAT relevant to the studies included—quantitative non-randomized, qualitative, and mixed methods—as reported in 4.3 (see Table 6).

The creators of the MMAT advise against assigning each study a global rating of strength, as this risks obscuring individual merits and weaknesses of each study (Hong et al., 2019). Nonetheless, a global assessment of risk of bias for each study facilitates the comparison among studies; indeed, an overall strength rating is a common component of many risk of bias tools (Li et al., 2017). This study, in addition to reporting the individual ratings for each study in question, assigned a global weight of evidence score of weak, moderate, or strong based on the risk of bias. Basing from the guidelines of other tools such as the Effective Public Health Practice Project Quality Assessment Tool (Armijo-Olivo et al., 2012), a strong rating was given where a study had zero criteria rated as 'no' and at least four 'yes,' a moderate strength was indicated by no more than one 'no' and at least three 'yes,' and weak ratings were given to studies with two or more 'no' answers or fewer than three 'yes.' For mixed methods designs, the global strength of evidence was marked as the weakest of the three ratings.

Given the subjective nature of assessing risk of bias and overall weight of evidence, it is best practice to include more than one reviewer in this process (Petticrew & Roberts, 2006). For this work, the second reviewer conducted a blind second screening of three studies. This equaled to 16% of total studies and comprised of two studies with purely quantitative research designs and one study with a mixed methods design. This sample was considered representative of the overall group of studies (15 quantitative, 1 qualitative, 3 mixed methods). As with the title and abstract review, an overall agreement of 0.7 was required. There was perfect agreement (1.0) regarding the three global strength of evidence scores, although interrater reliability was lower when considering individual criteria. For the two studies reporting quantitative findings, a value of 0.70 was obtained. However, the agreement for the mixed methods study was lower than desired at 0.47. The two reviewers discussed the disparity and reached an agreement, finding that differences stemmed mainly from the appraisal of the qualitative section and what degree of detail was required to justify having met a given criterion. While the overall Kappa value (0.6) was below the required threshold, time constraints did not allow for further screening by the second reviewer, and the primary reviewer finished the remainder of screening considering the feedback from the second reviewer.

3.9 Data synthesis

Meta-analysis of results is often inappropriate in the social sciences, where study interventions or outcomes can be highly heterogeneous (Petticrew & Roberts, 2006). Given the diversity of outcomes within this group and small number of studies reporting comparable measures, a narrative synthesis of study quality and findings is employed. Thomas et al. (2004) contend that the inclusion of qualitative information helps to triangulate the findings of quantitative data, and as such the resulting dialogue between these two outcomes in a narrative synthesis is highly informative.

In the proceeding chapters, this review follows the three-step narrative synthesis structure outlined by Petticrew and Roberts (2006): (i) studies are grouped into logical categories based on their reported outcomes; (ii) the findings and quality within each set is analyzed; and finally (iii) the findings among all groups are synthesized.

CHAPTER 4. Results

This chapter reports the general characteristics and findings of the final group of studies retrieved through the search process. A full list of the 19 studies and their complete references can be found in Appendix D. The first section of this chapter presents an overview of the general characteristics of the research, 4.2 reports on the findings and offers a short narrative summary of each study, 4.3 evaluates the risk of bias of each individual study and the cumulative confidence across studies as a whole, and 4.4 provides a summary of findings with reference to the research questions of this review.

4.1 General characteristics

Table 4 presents an overview of the general characteristics of the 19 studies included in this review. The following sub-sections review this information and illustrate patterns in the data regarding publication details, geographic and instructional context, study design, and reported outcomes.

Table 4. General characteristics of included studies

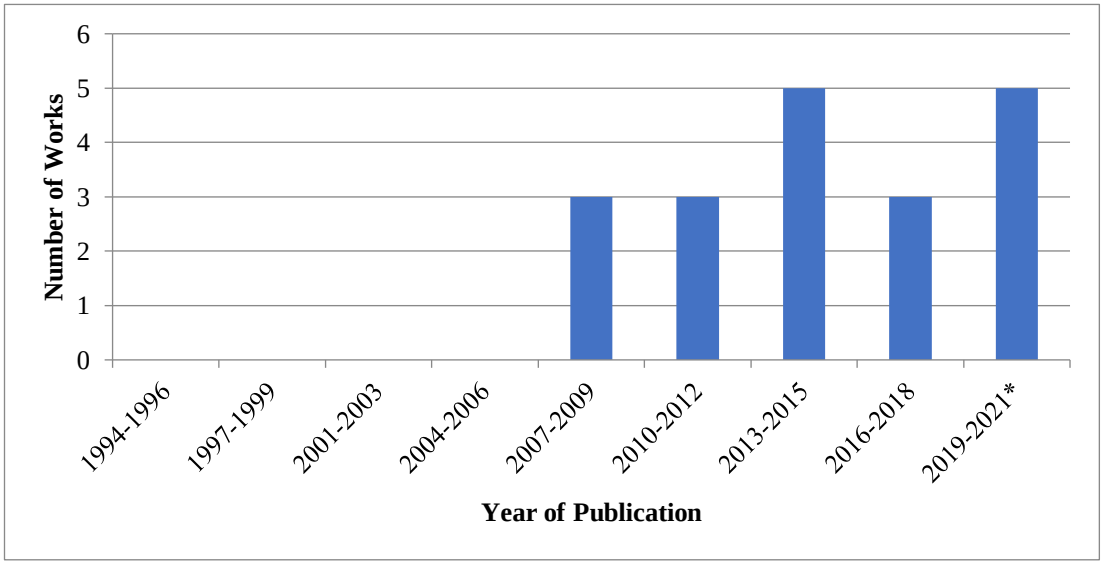
Study	Publication status	Study location	School languages	Student population age	Data type	Study duration	Non-CLIL control (matching)	Sample size	General outcomes	Specific outcome measures
1. Amengual-Pizarro & Prieto-Arranz (2015)	Book chapter	Spain	Catalan, English, Spanish	Secondary	Quantitative	Longitudinal 2 years	Yes (age)	151	Attitudes	Views towards multilingualism
2. García Mayo & Villarreal Olaizola (2011)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Longitudinal 3 years	Yes (age)	78	L3 skills	Morphosyntax
3. Grisaleña et al. (2009)	Journal article	Spain	Basque, English, Spanish	Secondary	Mixed methods	Longitudinal 2 years	Yes (age)	229	Attitudes, L3 skills	Program satisfaction; four skills
4. Gutiérrez Mangado & Martínez-Adrián (2018)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Cross-sectional	Yes (age)	35	L3 skills	Morphosyntax, pragmatics
5. Herraiz Martinez & Sánchez Hernández (2019)	Journal article	Spain	Catalan, English, Spanish	Secondary	Quantitative	Cross-sectional	No	19	L3 skills	Pragmatics
6. Lasagabaster (2009)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Cross-sectional	Yes (age)	277	Attitudes	Views towards multilingualism
7. Lázaro Ibarrola (2011)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Longitudinal 2 years	Yes (age)	26	L3 skills	General proficiency, Morphosyntax
8. López-Deflory & Juan-Garau (2017)	Journal article	Spain	Catalan, English, Spanish	Secondary	Mixed methods	Cross-sectional	Yes (none)	73	Attitudes	Views towards multilingualism
9. Martínez Adrián & Gutiérrez Mangado (2015)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Cross-sectional	Yes (age, initial exposure)	44	L3 skills	Morphosyntax

10. Merino & Lasagabaster (2018)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Longitudinal 1 year	Yes (age)	285	L1 skills, L3 skills	Reading, writing; four skills
11. Ní Dhiorbháin (2020)	Journal article	Ireland	English, German/ French, Irish	Primary	Qualitative	Cross-sectional	N/A	6	Attitudes	Program satisfaction
12. Ollo Jiménez & Martínez-Adrián (2019)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Cross-sectional	Yes (age, hours of exposure)	78	L3 skills	General proficiency, L1-based CSs
13. Pérez-Vidal & Roquet (2015)	Journal article	Spain	Catalan, English, Spanish	Secondary	Quantitative	Longitudinal 1 year	Yes (hours of exposure)	100	L3 skills	Reading, writing, listening, morphosyntax
14. Prieto-Arranz et al. (2015)	Book chapter	Spain	Catalan, English, Spanish	Secondary	Quantitative	Longitudinal 3 years	Yes (age)	87	L3 skills	Reading, listening
15. Ruiz de Zarobe (2008)	Journal article	Spain	Basque, English, Spanish	Secondary	Quantitative	Longitudinal 3 years	Yes (age)	21	L3 skills	Speaking
16. San Isidro (2010)	Book chapter	Spain	English, Galician, Spanish	Secondary	Quantitative	Cross-sectional	Yes (age)	287	L3 skills	Four skills
17. San Isidro & Lasagabaster (2019)	Journal article	Spain	English, Galician, Spanish	Secondary	Quantitative	Longitudinal 2 years	Yes (age)	44	Content, L1 skills, L3 skills	Content knowledge; four skills; four skills
18. San Isidro & Lasagabaster (2020)	Journal article	Spain	English, Galician, Spanish	Secondary	Quantitative	Longitudinal 2 years	Yes (age)	88	Attitudes	Views towards multilingualism
19. Wang & Kirkpatrick (2013)	Journal article	Hong Kong	Cantonese, English, Mandarin	Primary	Mixed methods	Cross-sectional	No	144	Attitudes	Program satisfaction

4.1.1 Publication details

The individual years of publication can be found in the study column in Table 4, while Figure 2 illustrates the trends over time, reporting the number of studies meeting the search criteria published in three-year bands from January 1st, 1994- May 4th, 2021. No studies fitting the search criteria were published in the 13-year range between 1994 and 2007. The oldest study[15] was published in 2008, while the most recent studies[11,18] were published in 2020. Since the year 2008 there has been a small but steady output of work, with 19 studies published in the past 14 years, or a little over one study per year on average.

Figure 2. Number of included studies by publication year



As seen in Table 4, the majority—16 out of 19, or 84%—of included studies are peer-reviewed journal articles. In contrast, the remaining three studies[1,14,16] are book chapters, with the former two chapters coming from the same book (Juan-Garau & Salazar-Noguera, 2015). Thirteen of the studies were retrieved via the database search, and an additional six studies[2,3,7,9,15,16] were retrieved through forward or backward citation searching.

4.1.2 Geographic and instructional context

Geographic region

Table 4 demonstrates that most of the published work found by this systematic review was conducted in Spain, accounting for 17 out of 19 of the included studies. The other two studies regard schools in Hong Kong[19] and Ireland[11]. Furthermore, 17 of 19 studies were published in English, while one was published in Spanish[3] and another in Irish[11].

Figure 3 provides a further breakdown of geographic areas, highlighting different LOIs across the research by splitting the 17 Spanish studies based on region: Basque-speaking regions (the BAC and Navarre), Catalan-speaking regions (Catalonia, the Valencian Community, and the Balearic Islands), and Galicia (where Galician is spoken). Overall, the most studies (47%) were conducted in Basque-speaking regions and an additional 26% and 16% of studies focused on schools in Catalan-speaking regions and Galicia, respectively. None of the 19 studies investigated learning outcomes in more than one geographic region.

Figure 3. Geographic region

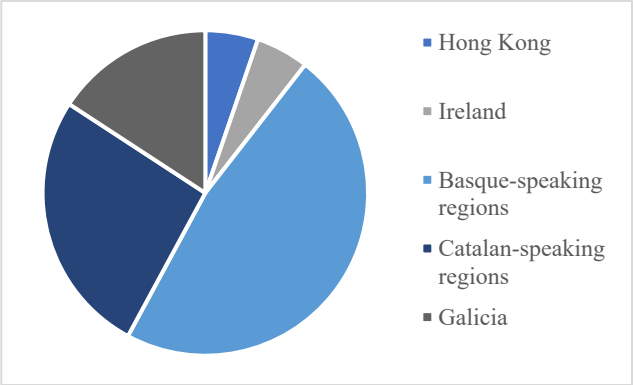
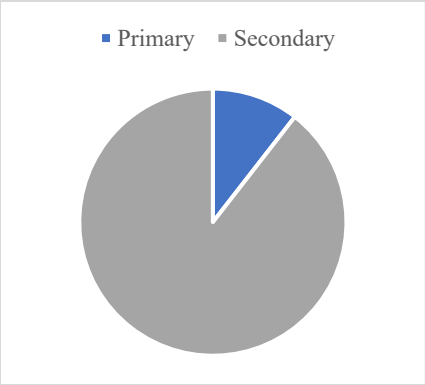


Figure 4. Age of students



Age of participants

Figure 4 shows the breakdown of target age groups for the included studies. Most of the research (89%) focused on students enrolled in secondary schools, with only two studies[11,19] reporting on language programs in primary education (ages 5-10).

4.1.3 Study design

Data type

Only four studies collected qualitative data: one study[11] reported only qualitative data and three studies[3,8,19] employed a mixed methods design. On the other hand, 18 of the 19 studies included quantitative measures, with 15 of these reporting exclusively quantitative findings. No study involved allocation to treatment conditions, as all were observational studies of intact classes in schools.

Study duration

There was an almost even split of cross-sectional and longitudinal studies, with nine of the former and ten of the latter. Figure 5 illustrates study duration of the selected works, including a further classification of longitudinal designs, which ranged from 1-3 years in length.

Control groups

Three studies[5,11,19] did not elicit findings from a non-CLIL comparison group. The remaining 16 studies all included at least one comparison group, and two of these studies[9,12] compared CLIL performance with two different non-CLIL groups selected to control for different factors. Overall, 14 studies chose age-matched comparison groups; in contrast, two studies[12,13] matched groups on total hours of L3 instruction, one study[9] held initial age of L3 exposure constant, and a final study[8] did not control for any of these factors. A further discussion of the confounders controlled for by each study follows in 4.3.

Figure 5. Study duration

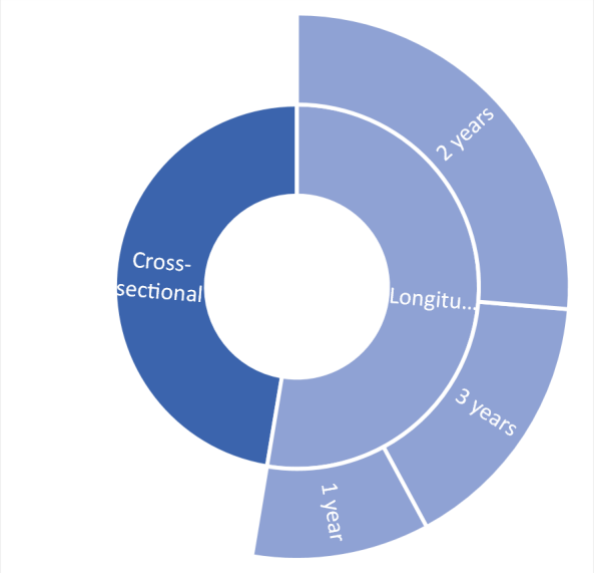
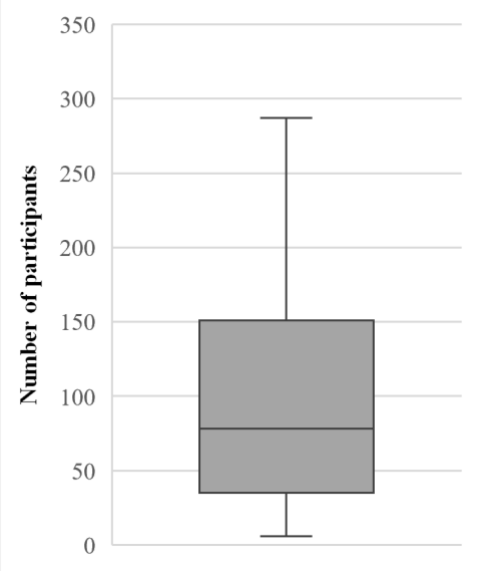


Figure 6. Sample size



Sample size

The sample size across all studies can be found in Figure 6. These values ranged from 6 to 287, with a median of 78 participants. Three studies included measures of only adult participants[11] or both and adult stakeholders and children[18,19], while the other 16 studies recruited only child participants (i.e., students under the age of 18). Furthermore, three of the four studies reporting qualitative findings[8,11,19] elicited data from fewer participants than the overall median of the group of included studies. The fourth study reporting qualitative data[3] did not state the number of participants sampled in elicitation of qualitative results. Though the three studies reporting sample size for qualitative data come from the bottom two quartiles of the whole group and lower the minimum from 19 to 6, the exclusion of these data only shifts the median from 78 to 82.5 participants and does not greatly affect the overall picture demonstrated in Figure 6.

4.1.4 Reported outcomes

Figure 7 offers a view of the types of outcomes—attitudes, content knowledge, L1 skills, and/or L3 skills—reported by each study. Six of the seven studies reporting results regarding attitudes did not report any other outcomes. In contrast, neither study reporting on content knowledge or L1 skills exclusively reported these outcomes; rather, both of these studies[10,17] also included measures of L3 skills. Therefore, while attitudes were frequently reported in isolation, content knowledge and L1 skills were always grouped with other measures, and L3 skills were flexibly reported on their own (10) or with at least one other outcome (3).

Figure 8 shows the total number of studies that report on each of four different outcomes (not mutually exclusive). Altogether, the 19 included studies offered 23 assessments of outcomes: seven on attitudes, one on content knowledge, two on L1 skills, and 13 on L3 skills. None of the included studies reported on L2 skills.

Figure 7. Outcome type by study

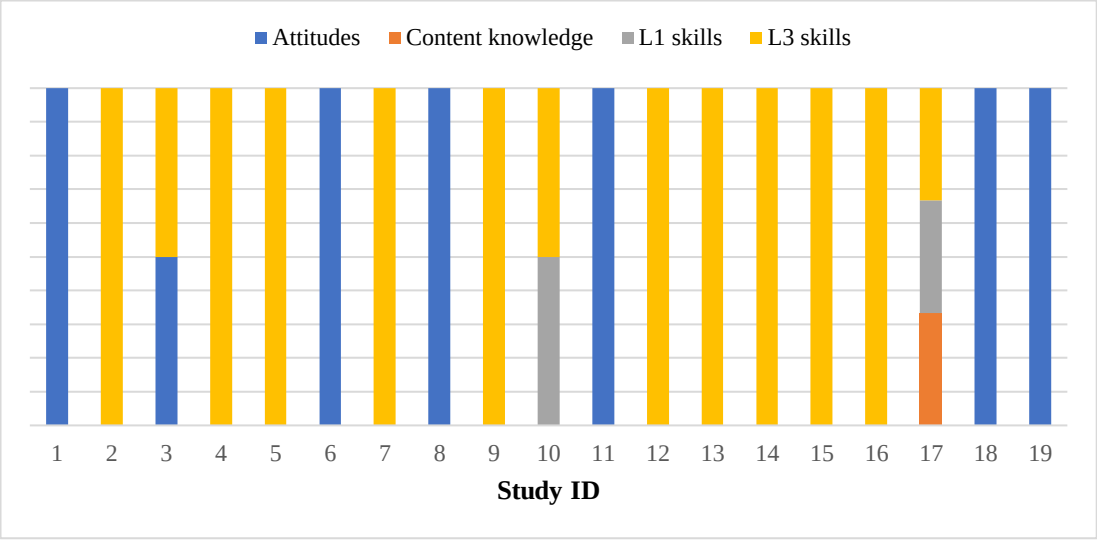


Figure 8. Outcome frequency

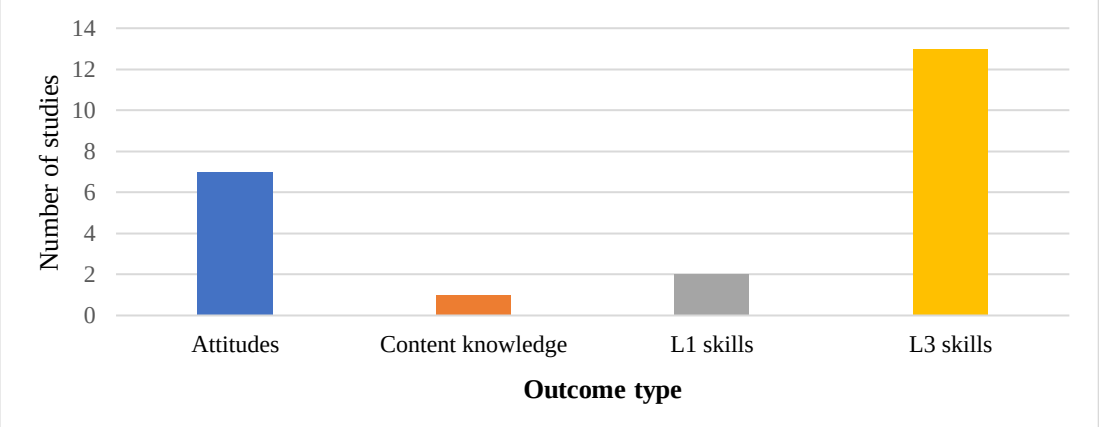
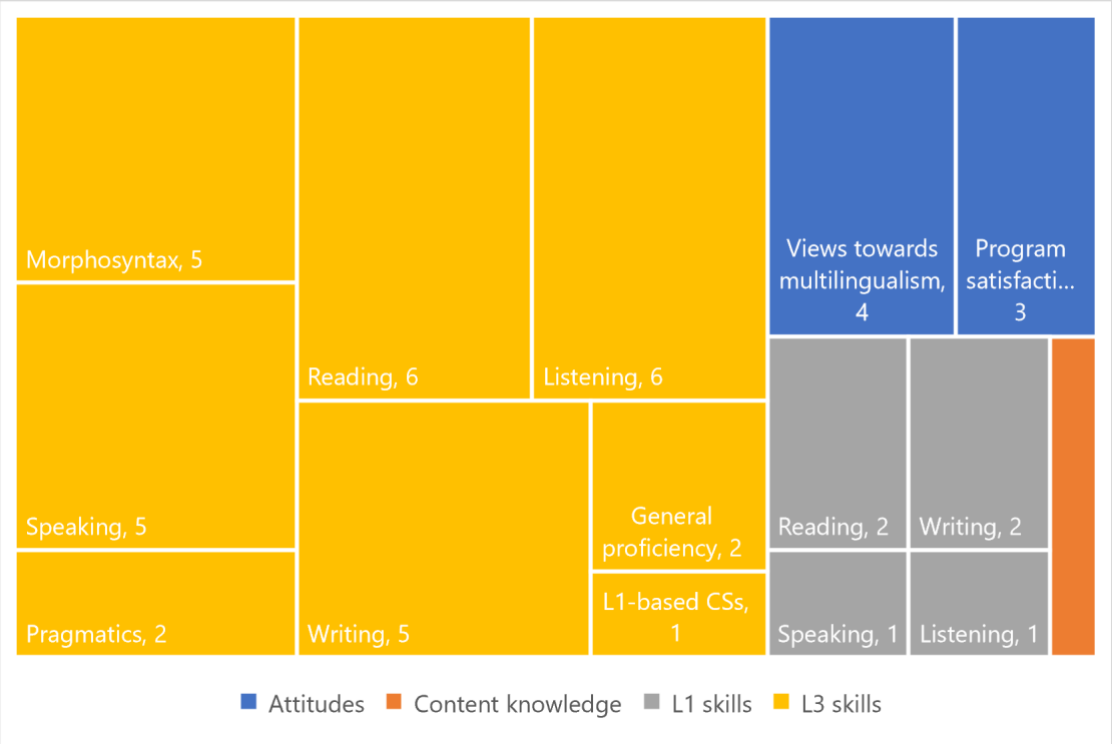


Figure 9 illustrates the variety of specific measures reported for each outcome. The 13 studies reporting outcomes in L3 skills investigated eight different sub-areas of L3 ability, with many assessing more than one of these areas. Overall, the most commonly reported measures of L3 skills were reading (6) and listening (6), followed closely by writing (5), speaking (5), and morphosyntax (5). The seven studies reporting attitudes each only investigated one measure of attitudes, either views towards multilingualism (4) or educational program satisfaction (3). With only two studies offering assessments of L1 skills, the studies investigated four different measures of L1 abilities. Finally, the one study reporting content knowledge outcomes included only one measure of subject-specific knowledge. As the proportional areas in Figure 9 show, measures of L3 skills were most common, representing more than two-thirds of the total reported outcome measures.

Figure 9. Specific measure and frequency of outcomes



4.2 Study findings

The above section revealed that all studies in this review were published since 2008, and that the majority of studies came from secondary schools in Spain. Most of the works reported quantitative outcomes, although one qualitative and three mixed methods studies were also retrieved.

Moving beyond the general characteristics of the included studies as a whole, this section now reviews the findings from individual studies. First, 4.2.1 offers brief narrative descriptions of the design, context and findings of each individual study. The descriptions elaborate upon the research questions, outcome measures, and direction of reported effect. Following the narrative reviews, 4.2.2 presents the effect sizes of studies. Appendix E summarizes the direction of effects by study as presented in this section and illustrated further in 4.4.

4.2.1 Narrative summaries of individual studies

1. **Amengual-Pizarro and Prieto-Arranz (2015)** conducted a longitudinal case study of 151 Catalan-Spanish bilingual students in five state-run secondary schools in the Balearic Islands. Roughly half of the students (85) were enrolled in CLIL sections with six hours of L3 English each week, while the non-CLIL control group students (66) received only three hours of EFL lessons per week. Administering a questionnaire twice over two years, the investigators explored whether learning context (CLIL vs. non-CLIL) would influence students' affective factors concerning views towards the content subjects taught through English and interest in language learning. The researchers reported that both CLIL and non-CLIL students presented positive attitudes towards language learning. Among the 55 questionnaire items, only two demonstrated significant between-group differences. The authors concluded that there were no significant differences in attitudes towards L3 learning between CLIL and non-CLIL groups.
2. **García Mayo and Villarreal Olaizola (2011)** investigated whether there were differences in accuracy of L3 oral English morphosyntax in Basque-Spanish bilingual secondary school students across two time points three years apart. While 50 students (27 CLIL/23 non-CLIL) participated initially, only 28 students (13 CLIL/15 non-CLIL) were tested three years later. Among the non-CLIL control group at the second point of testing, only four were from the original cohort of participants, indicating a high rate of attrition and mixing in of new participants. Through the analysis of an oral narration task at the two time points, the investigators calculated each group's proportion of omission of four English morphemes. Significant between-group differences were only found for one morpheme at one time point, thus the authors concluded that

there was not sufficient evidence that CLIL students performed differently than the comparison group.

3. **Grisaleña et al. (2009)** conducted a study with 229 Basque-Spanish bilingual secondary students in the BAC, evaluating whether exposure to CLIL pedagogy was more effective than a traditional foreign language teaching approach with respect to L3 English competence. Participants were tested twice over two years and spanned three age groups and two conditions (CLIL vs non-CLIL). Quantitative data regarding English skills in reading, listening, writing, and speaking were synthesized with qualitative findings from interviews with students, parents, and teachers to assess the effectiveness of the CLIL program. Overall, the report concluded that the CLIL program had a positive impact on all measured L3 skills and attitudes towards CLIL methodology.
4. **Gutiérrez Mangado and Martínez-Adrián (2018)** studied the effects of CLIL on the acquisition of L3 English nominal morphology and article use by 35 Basque-Spanish bilingual secondary students in the BAC. Both a CLIL and a non-CLIL group participated in an oral production task to gauge article omission and overuse errors. The results indicated that there was no significant difference in the omission of articles between groups, but that the CLIL group overused definite articles significantly less than their non-CLIL counterparts. The authors concluded that CLIL led to pragmatic benefits (syntax-semantics-discourse-interface), where participants formed more target-like utterances and exhibited less overuse, though there was no evidence that CLIL groups outperformed the non-CLIL cohorts in measures of morphosyntax such as omission of required articles in oral tasks.
5. **Herraiz Martinez and Sánchez Hernández (2019)** studied the frequency and type of pragmatic markers employed by CLIL students in oral production tasks in L1 Catalan and Spanish and L3 English. Participants were 19 Catalan-Spanish bilingual secondary school students from the Valencian Community, Spain. All students were members of the CLIL section of the school, learning through Catalan, Spanish, and English. The investigators reported that participants produced significantly more pragmatic markers in L3 English than in either L1 Catalan or Spanish, and concluded that students participating in

CLIL pedagogy were able to achieve a high level of L3 proficiency with reference to pragmatic skills in oral productive environments.

6. **Lasagabaster (2009)** studied whether CLIL fostered positive attitudes towards multilingualism in a sample of 277 Basque-Spanish bilingual secondary students in the BAC, Spain. A total of 168 students received CLIL pedagogy, while the 109 students in the non-CLIL control only received L3 EFL classes. Using a survey and the Varimax rotated factor matrix, the author identified five factors that together explained 56.090% of the total variance between scores. The CLIL intervention group presented significantly more positive views towards bilingualism in each of the factors when considered as a whole, and 16 out of 25 of the individual survey questions. While both sets of students exhibited generally positive views towards multilingualism, students who had received CLIL instruction were significantly more positive than their non-CLIL counterparts.
7. **Lázaro Ibarrola (2011)** investigated the differences in the development of L3 English morphosyntax between CLIL and non-CLIL secondary students in the BAC, Spain. Participants were 26 Basque-Spanish bilinguals who completed two oral production tasks, once after one year of CLIL instruction and again two years later. Measuring rate of accuracy of lexical verb inflection, use of pronouns, and frequency of subordinate clauses, the author reported that CLIL students exhibited significant improvement over time in all areas and outperformed the non-CLIL group at both time points. The author contended that the results demonstrated the superiority of CLIL pedagogy in developing grammar skills in oral tasks.
8. **López-Deflory and Juan-Garau (2017)** conducted a mixed methods study with 73 Catalan-Spanish bilingual secondary students in the Balearic Islands to assess whether CLIL pedagogy influenced multilingual language attitudes. All participants, 43 CLIL and 30 non-CLIL students, took a Likert-scale survey indicating their attitudes towards multilingualism. The CLIL students presented significantly more positive attitudes in half of the pre-determined categories and no differences in the other categories. In a qualitative analysis of open-ended responses, non-CLIL students reported more instrumental, career-related motivators than their CLIL counterparts. In contrast, CLIL students highlighted more cultural value of languages, both L1 Catalan and L3

English. The authors concluded that the quantitative and qualitative findings converged, indicating that exposure to CLIL pedagogy fostered more positive views towards multilingualism.

9. **Martínez Adrián and Gutiérrez Mangado (2015)** studied whether disparities existed in general L3 competence and morphosyntax between CLIL and non-CLIL groups when both age and total instructional hours were controlled. A total of 44 Basque-Spanish bilingual secondary students were assessed: 13 from a CLIL cohort, 19 from an age-matched non-CLIL control that had received equal L3 instructional hours but had an earlier age of first exposure, and 12 from a non-CLIL control that were two years older but had received a comparable number of L3 instructional hours and had the same age of first exposure. The CLIL group showed significantly higher scores than the control group (matched in age and hours of exposure) on English proficiency tests and comparable results to the older group. However, regarding the omission of specific L3 inflectional morphemes, the CLIL participants did not outperform the control groups, and performed worse than the older non-CLIL group. The authors concluded that CLIL pedagogy offers benefits to general L3 linguistic competence but not for morphosyntax when age is held constant.
10. **Merino and Lasagabaster (2018)** conducted a longitudinal study of 285 Basque-Spanish bilingual secondary school students over two years, with an experimental group that had just begun CLIL education and a non-CLIL control group. They found that the CLIL students outperformed the non-CLIL control group across all L3 English competencies tested (speaking, writing, reading, and listening) and in tests of Basque and Spanish reading and writing. However, they also highlighted that the linguistic development between groups over the two-year study did not exhibit statistically significant differences. Given that students self-selected into the CLIL group and entered the study with a higher level of English to begin with, the authors concluded that there was not sufficient evidence of the superiority of a CLIL model over traditional language education.
11. **Ní Dhiorbháin (2020)** studied the implementation of trilingual CLIL across six primary schools in Ireland over nine months and investigated the potential benefits and drawbacks. All participating schools were Irish-medium primary schools that offered most subjects in Irish, with limited hours of L1 English to

support literacy. The additional languages of German or French were introduced in physical education, art, or music classes as part of a pilot scheme. At the conclusion of the year, the investigator conducted a group interview with five participating teachers and one principal to gather their opinions. The educators held very positive views towards the CLIL program and saw marked progress in their students' L3 acquisition. Overall, teachers did not view CLIL pedagogy as an impediment to content learning, though teachers' perceptions of their own shortcomings in the L3 could be a possible barrier to program effectiveness. The study concluded that multilingual CLIL has the potential to be highly effective in Irish-medium primary schools and should be developed further.

12. **Ollo Jiménez and Martínez-Adrián (2019)** investigated the L3 English proficiency and use of L1-based communication strategies (CSs) in 78 secondary students in Navarre, Spain. Participants were from four intact classrooms: CLIL 1 (*n*= 23), a class of 12/13-year-old learners; non-CLIL 1 (*n*= 14), a class of students the same age as CLIL 1 but with fewer hours of English exposure; CLIL 2 (*n*= 22), a class of students three years older than the younger groups; and non-CLIL 2 (*n*= 19), a class of students the same age as CLIL 2 and with similar hours of English exposure as CLIL 1. Regarding English proficiency, results showed that both CLIL groups had significantly higher scores than their age-matched peers in non-CLIL sections and that the younger CLIL group did not exhibit significant differences when compared to the older non-CLIL class. Concerning L1-based CSs such as borrowing, significant differences were only found with the older CLIL group, which used significantly fewer of these strategies than either the non-CLIL 2 or CLIL 1 students. The authors concluded that CLIL students had higher L3 proficiency and relied less on the L1 when speaking than their non-CLIL counterparts.
13. **Pérez-Vidal and Roquet (2015)** conducted a longitudinal study with 100 Catalan-Spanish bilingual secondary students in Spain, assessing relative gains in L3 English skills over one academic year. Half of the participants were in a CLIL section studying science in English, and the other half were from classes that were a year older and did not have an option to enroll in CLIL, and thus only had traditional EFL classes. In this way, the investigators sought to control for the total hours of exposure to English and potential effects of higher

motivation in CLIL sections, while still taking into account age differences by measuring relative gains over the year. The authors concluded that CLIL students exhibited significantly greater relative improvement in English reading, grammar skills and writing accuracy, though no differences were observed in listening.

14. **Prieto-Arranz et al. (2015)** investigated the changes in L3 English reading and listening scores in 87 Catalan-Spanish bilingual secondary school students in Spain. Of the participants, 50 were in a CLIL section studying social science or natural science in English and 37 received only EFL classes. Measuring scores at four points across three years, the investigators reported growth across both skills for both sets of students, but more so for the CLIL group in reading tests. A two-way ANOVA revealed significant main effects of group and time, and an interaction between the two. The authors concluded that CLIL pedagogy had a positive effect on L3 reading development and no impact on listening skills.
15. **Ruiz de Zarobe (2008)** conducted a three-year longitudinal study with Basque-Spanish bilinguals in Spain investigating L3 English speaking skills. The initial cohort consisted of 89 secondary school students from three schools split across a non-CLIL section studying EFL, a CLIL section with one additional subject in English, and a CLIL section with two additional subjects in English. Speaking skills were measured at three time points, although at final testing only 21 of the original 89 students were tested. The authors reported that across all three test times the CLIL sections performed significantly better than the non-CLIL comparison group in every competency tested (pronunciation, vocabulary, grammar, fluency, and content).
16. **San Isidro (2010)** conducted a cross-sectional analysis of L3 English skills of 287 Galician-Spanish bilingual secondary school students from ten schools in Spain. Half of the students had participated in a CLIL program for two years while the other half had only received traditional EFL classes during this time. The investigator found significant between-group differences for all the L3 skills tested: overall proficiency, reading/writing, listening, and speaking. Moreover, San Isidro (2010) found no significant differences among participants based on sex, however he did find that on average students from

urban areas significantly outperformed their peers from rural zones in all areas except reading/writing.

17. **San Isidro and Lasagabaster (2019)** conducted a longitudinal study with 44 secondary students in Spain to track differences across two academic years between CLIL and non-CLIL students in L1 Galician and Spanish, L3 English, and CLIL subject (social science) content knowledge. At the beginning of the study, students were admitted to the CLIL section on a first-come, first-served basis, and there were no statistically significant differences between the two groups in any of the tested competencies. Two years later, both groups had made significant progress in L3 English, though the CLIL group had done so to an even greater extent. Regarding the L1s of Galician and Spanish, the CLIL section exhibited significant improvements over the two-year period, while the non-CLIL group made limited gains in Galician and presented no difference in Spanish. Finally, the CLIL cohort did not demonstrate significant changes in social science knowledge, while the non-CLIL students experienced a small yet significant drop in scores between the first year and the second. The authors concluded that the students receiving CLIL methodology made greater improvements across all three languages than their non-CLIL peers and presented no detrimental effects regarding their learning of content knowledge.
18. **San Isidro and Lasagabaster (2020)** investigated the views towards multilingualism of 44 secondary students and their parents across two academic years. The participants were the same cohort of Spanish students from Galicia studied in San Isidro and Lasagabaster (2019), though this second study also surveyed one parent of each student. Notably, the CLIL parents were more than four times as likely to have attended university than the non-CLIL parents ($p< 0.05$). While all students and parents began with relatively positive attitudes towards learning other languages and exhibited significant trends of growing more positive over time, these changes were significantly greater for both CLIL students and parents when compared to their non-CLIL counterparts. The investigators concluded that participation in a multilingual CLIL model had positive benefits on attitudes towards multilingualism in both participating students and their parents.
19. **Wang and Kirkpatrick (2013)** conducted a mixed methods study investigating stakeholder satisfaction and opinions regarding trilingual CLIL

pedagogy as took place in a primary school in Hong Kong offering Cantonese, English, and Mandarin as MOIs. The investigators conducted interviews with school staff (n = 13) and parents (n= 10) to assess their views of successes and challenges of the trilingual curriculum. Generally, teachers valued the school's trilingual curriculum and considered it to be highly successful, although several teachers highlighted the challenge of teaching such young students through English and frequently switched into Cantonese to explain concepts while teaching. Parents expressed pride that their students were learning in three languages and fully supported this educational model. In addition to the interviews, 121 students responded to a survey with Likert-scale questions regarding their perceptions of the school. Students expressed positive views towards having three MOIs in the school, though reported more confidence in the development of their Cantonese skills than their English skills. Overall, the authors concluded that the quantitative and qualitative findings indicated that key stakeholders harbored positive views towards the use of three languages in the school.

4.2.2 Effect sizes

Of the 18 studies reporting quantitative data, only Merino and Lasagabaster (2018) reported effect sizes for their findings. An additional 14 studies included enough data (or provided it upon request) to calculate an effect size. Table 5 presents the effect sizes for any statistically significant differences reported. The size of effect (small, medium, high) was determined based on the standards advised by Cohen (1988). Three studies[3,13,17] neither reported effect size nor sufficient information to calculate one and did not reply to an email requesting further data.

As described in 4.2.1, four studies[1,2,9,10] did not find sufficient evidence of differences between CLIL and non-CLIL groups. Two of these studies[1,2] found a small or medium effect size but only for a limited number of the many tasks conducted. One study[9] found large, but contradictory, effect sizes. A final study[10] also found large effect sizes, though these disparities did not hold when considering the longitudinal changes in performance. The remaining studies found an overall positive impact of CLIL and reported a range of effect sizes. The majority of these effect sizes were large, especially for L3 skills and attitudes. The only effect size for L1 skills was small, and there were no reported effect sizes for content knowledge.

Table 5. Effect sizes by study

Study	Outcome	Outcome measure and effect size[1]	Size of effect
1. Amengual-Pizarro & Prieto-Arranz (2015)*	Attitudes	Intrinsic motivation towards L3: 0.38	small
2. García Mayo & Villarreal Olaizola (2011)*	L3 skills	Omission of the copula: 0.78	medium
4. Gutiérrez Mangado & Martínez-Adrián (2018)	L3 skills	General proficiency: 1.26 Correct article usage: 0.84	large
5. Herraiz Martinez & Sánchez Hernández (2019)	L3 skills	*Pragmatic marker use:* English-Catalan: 1.03 English-Spanish: 1.76	large
6. Lasagabaster (2009)	Attitudes	Views towards multilingualism: 0.32-0.60	small-medium
7. Lázaro Ibarrola (2011)	L3 skills	Various measures of correct English pronoun use and inflection: 0.88-2.92	large
8. López-Deflory & Juan-Garau (2017)	Attitudes	International attitudes: 2.97 L3 views: 1.41 L1 views: 0.87	large
9. Martínez Adrián & Gutiérrez Mangado (2015)*	L3 skills	General proficiency (age-matched control): 2.50 Omission of inflection (non-age-matched control): -1.19	large
10. Merino & Lasagabaster (2018)*	L3 skills L1 skills	English competences: r= 0.63-0.68 Basque reading: r= 0.31 Basque writing: r= 0.22 Spanish writing: r= 0.26	large small
12. Ollo Jiménez & Martínez-Adrián (2019)	L3 skills	*General proficiency:* Younger, CLIL-NCLIL: 1.93 Older, CLIL-NCLIL: 3.39 Older-younger, CLIL: 3.24 Older-younger, NCLIL: 1.96 *L1-based CSs:* Older CLIL-NCLIL: 0.37 Older-younger CLIL: 0.56	large small medium
14. Prieto-Arranz et al. (2015)	L3 skills	General reading: η^2= 0.01 Specific reading: η^2= 0.15 News listening: η^2= 0.03	small large small

[1] Effect sizes refer to Cohen's *d*, unless indicated otherwise

15. Ruiz de Zarobe (2008)	L3 skills	Various writing tasks: 0.61-0.98	medium-large
16. San Isidro (2010)	L3 skills	Reading/writing: 1.19 Listening: 0.95 Speaking: 1.20	large
18. San Isidro & Lasagabaster (2020)	Attitudes	*Student views towards CLIL:* T1: 0.90 T2: 1.58 T3: 1.70 *Parent views towards CLIL:* T2: 1.09 T3: 1.16	large
19. Wang & Kirkpatrick (2013)	Attitudes	Confidence (Cantonese-English): 0.89	large

*reported overall no significant effects of CLIL

4.3 Risk of bias

A summary of the individual risk of bias (RoB) ratings for each study can be found in Table 6, with data points of Yes (green), Can't tell (yellow), or No (red) for each relevant question relating to low, moderate, or high RoB. Regarding overall weight of evidence, three studies[8,11,17] received a 'strong' rating, seven studies[1,6,9,10,12,14,18] received a 'moderate' rating, and the remaining nine studies[2,3,4,5,7,13,15,16,19] received a 'weak' rating. The following subsections review the overall component ratings for the quantitative, qualitative, and mixed methods designs. Finally, this section concludes with an assessment of the cumulative confidence across studies in 4.3.4.

4.3.1 Studies reporting quantitative data

Of the 19 included studies, 18 reported quantitative findings and could be assessed using the quantitative section of the MMAT. In total, 'Yes' represented 50% or more of total responses in three of the five categories, while 'Can't tell' was most common for selection bias, and 'No' was most common for confounder bias.

Table 6. Risk of bias of individual studies

Study ID	Quantitative					Qualitative					Mixed methods					
	Selection bias	Intervention and outcome measures	Complete outcome data	Confounders	Intervention administration	Rationale for approach	Data collection methods	Findings derived from data	Interpretation supported by data	Coherence among steps	Rationale for approach	Different components integrated	Outputs of each well-interpreted	Divergences addressed	Quality of each component	Global strength of evidence rating
1																moderate
2																weak
3																weak
4																weak
5																weak
6																moderate
7																weak
8																strong
9																moderate
10																moderate
11																strong
12																moderate
13																weak
14																moderate
15																weak
16																weak
17																strong
18																moderate
19																weak
	8	9	13	6	11	4	2	2	2	1	1	1	1	1	1	3	Strong
	9	5	2	5	3	0	2	2	1	1	0	0	0	1	0	7	Moderate
	1	4	3	7	4	0	0	0	1	2	2	2	2	1	2	9	Weak
Total	18					4					3					19	Total

Selection bias

With regard to selection bias, only one study[15] exhibited a high risk of selecting a non-representative sample, namely due to the significant attrition and lack of reporting of reasons for dropouts. Just under half the studies (8) presented low RoB, while the remaining nine studies had small sample sizes or reported too few details to confidently confirm a low RoB.

Intervention and outcome measures

The category concerning intervention and outcome measures was a relatively strong category, with nine studies clearly outlining procedures that included validated, reliable instruments. Four studies[5,13,15,19] offered little to no information regarding the procedure and were deemed to have a high RoB, while a further five[2,7,12,14,16] with moderate RoB lacked details regarding validation of outcome measures.

Reporting of data

Concerning the reporting of data, 13 studies provided complete data with negligible attrition and clear reporting of reasons where dropouts occurred. Petticrew and Roberts (2006) recommend rates no higher than 20% attrition, which was used as a benchmark for this review. Three studies[2,13,15] either did not report complete data or had rates of attrition over 20%. The remaining two studies[1,10] had low rates of attrition but did not report the reasons for participants leaving the study.

Potential confounders

The section with the greatest risk of bias was that of controlling of confounders, where only six studies adequately controlled for risk of bias from confounding factors. Appendix F provides a list of seven potential confounders considered in this review and indicates whether a study employed a longitudinal design. Six studies[1,9,12,13,14,17] had low RoB in this area, employing a longitudinal (or pseudo-longitudinal) design and controlling for at least half of the considered confounders. Despite meeting these latter criteria, one study[18] was rated as having moderate RoB due to statistically significant differences in family SES between the intervention and control group students. Since parent opinions were a core measure of this study, differences in parent income and educational background were considered to add a significant risk of bias. Four more moderate-risk studies[4,8,10,15] either controlled for more than half of confounders with a cross-sectional design or only controlled for two or three of the confounders with a longitudinal design. Finally, seven studies[2,3,5,6,7,16,19] controlled for only one confounder

with a longitudinal design or two or three confounders with a cross-sectional design and were rated as having high RoB.

Administration of intervention

Eleven studies were assessed as having low RoB in their administration of intervention. While all these studies were observational and did not administer the intervention themselves, the studies were appraised on the likelihood of equal delivery of pedagogy or switching between groups. Three studies[1,15,16] had moderate RoB, due to participants being split across schools, where investigators did not make reference to alignment of teaching practices. Lastly, four studies[2,4,9,14] were rated as having high RoB, where the focus of pedagogy between classes/schools was explicitly different.

4.3.2 Studies reporting qualitative data

A total of four studies[3,8,11,19] reported qualitative data and were assessed in the relevant section of the MMAT. All four studies presented adequate rationale for pursuing a qualitative design.

Regarding data collection methods and findings derived from the data, two studies[8,11] expressed clear processes of data collection and analysis and logical methods for interpreting the data. The other two studies[3,19] offered insufficient explanation of the procedure and were assessed as having moderate RoB in these two categories.

Two studies[11,19] effectively reported evidence to justify their conclusions. One study[8] cited selective findings but did not convincingly evidence their conclusions and so had a moderate RoB. The final study[3] made no reference to specific data in the presentation of its conclusions and was assessed as having a high RoB.

With respect to coherence among steps, only one study[8] presented a low RoB. One study[11], despite having strong scores in the other sections, occasionally quoted participants that had not been introduced in the methodology, introducing some ambiguity. This study was rated with moderate RoB. The final two studies[3,19] reported very little of the qualitative methods and had high RoB due to lack of coherence.

4.3.3 Studies with a mixed methods design

Of the three studies with a mixed methods design, two[3,19] were given a global rating of 'weak,' due in general to the lack of integration between quantitative and qualitative findings. For both of these studies, very little description was offered regarding the qualitative part of the study.

The third study[8] was given a global rating of 'strong' for its cogent integration of quantitative and qualitative findings. There was clear justification of the use of a mixed methods design, and moreover, thoughtful synthesis of the two areas of findings.

4.3.4 Cumulative confidence across studies

Of the 19 studies included in this review, nine had a 'weak' global strength of evidence rating. Among the other half, only three were 'strong' and seven were 'moderate.' Despite roughly half of the included studies being evaluated as having high risk of bias, the individual components of the MMAT exhibited few systematic patterns of weakness across studies. Of all the aspects assessed, controlling of confounders in quantitative designs was the greatest source of bias across studies. However, for instance, it is impossible to have a control group matched for hours of in-school instruction in a longitudinal design because hours of instruction is the core difference between CLIL and non-CLIL groups, and so a certain degree of bias from confounders is inevitable. As a result, this group of studies overall has a moderate strength of evidence. The following section synthesizes the findings of this chapter and provides an initial answer to the research questions of this review.

4.4 Review of overall results

This final section reviews the findings of this chapter with the aim of providing a preliminary answer to this review's research questions. Chapter 5 picks up on these questions and explores in more depth the potential causes and implications of these results.

RQ 1: What is the extent and nature of empirical research investigating educational outcomes in multilingual CLIL school settings?

A total of 19 studies have been published in the past 27 years, although the first relevant study was not published until the year 2008. Overall, a total of 17 studies investigated secondary schools in Spain, while the other two were conducted in primary schools in Ireland and Hong Kong. Thus, the research output examining secondary schools in Spain is strong, however little to no work in this area has been conducted in primary schools or other geographic contexts.

RQ 2: What is the effectiveness of CLIL in multilingual schools?

Figure 10 plots the reported effect of CLIL and weight of evidence rating of every study. Each bubble in the graph represents one study. The color of the data point indicates the

reported outcomes of the study, while the shape relates to study design (circle-quantitative; triangle-qualitative; diamond-mixed methods), and the size of each shape is proportional to the effect size (a black outline denotes a study without calculable effect sizes).

Most studies (15/19) found positive effects of multilingual CLIL programs. Indeed, no study reported negative outcomes for CLIL students when juxtaposed with peers in comparable non-CLIL sections. Notably, all three studies with a strong rating found unequivocal benefits of CLIL, albeit with a smaller average effect size than the weak studies. Moreover, there was no systematic pattern among the studies reporting no net effects of CLIL, as these studies reported a mix of L3 outcomes (3), L1 outcomes (1), and attitudes (1).

Figure 10. Reported overall effect of CLIL by study

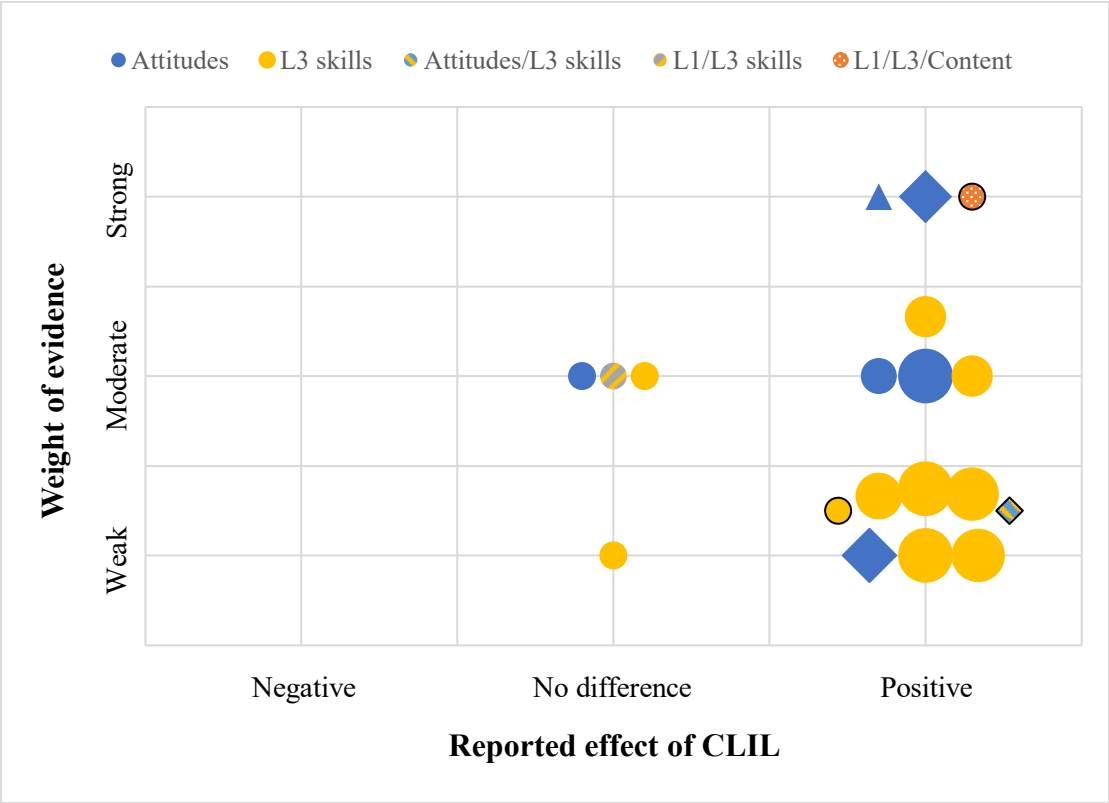
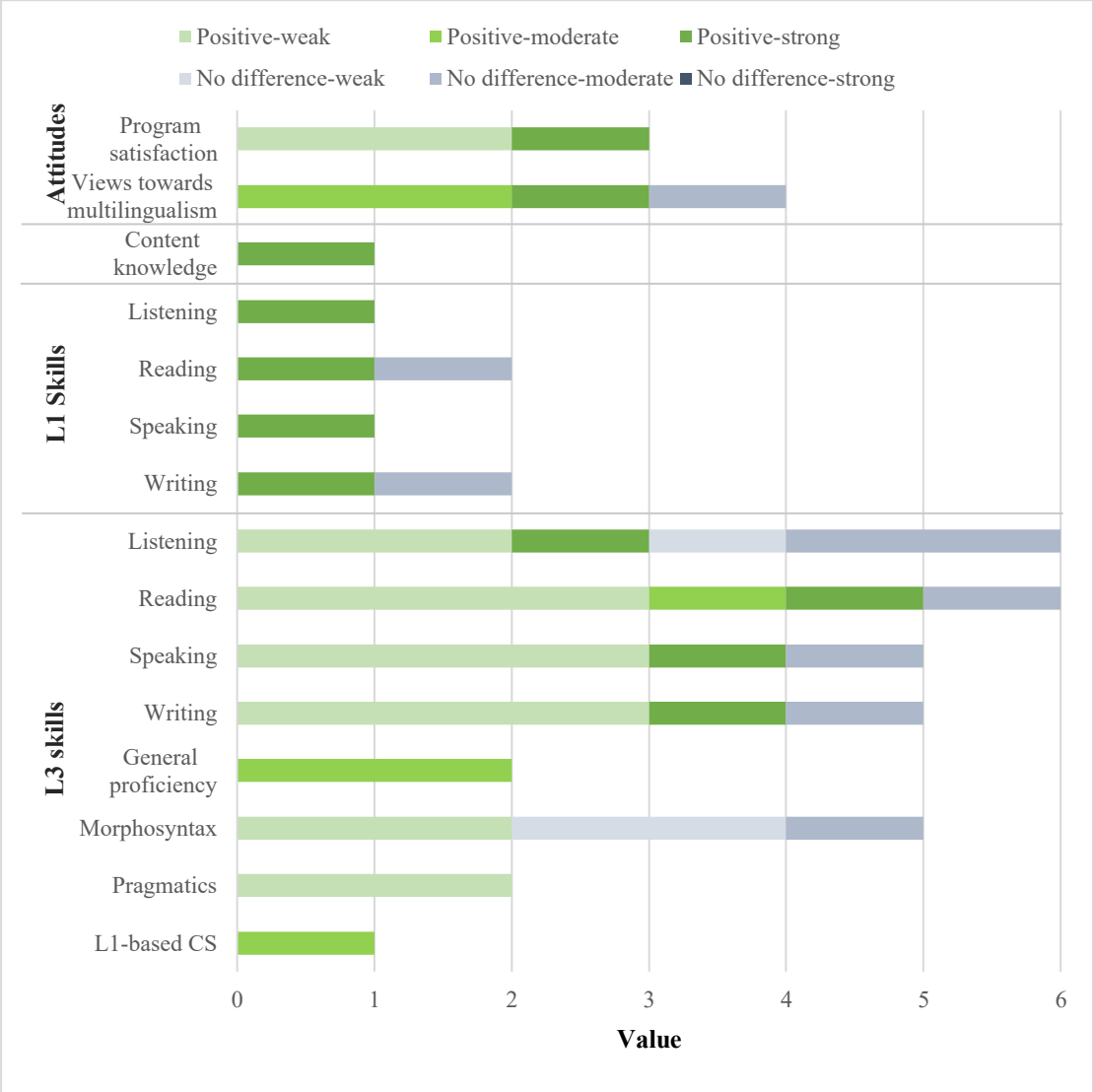


Figure 11 illustrates effects of CLIL for each of the specific measures reported by the studies in this review. Green and gray denote positive effects and no difference of CLIL pedagogy, respectively. The deeper the color, the stronger the weight of evidence of the study reporting those findings. In general, the findings suggest that multilingual CLIL has a very positive impact on stakeholder attitudes and some areas of L3 development such as reading, writing, and speaking. In contrast, this form of educational program

seems to have less of an influence on the L3 skills of listening and morphosyntax. The following chapter will analyze these results in more depth and discuss potential causes and implications of these findings.

Figure 11. Reported effect of CLIL by outcome measure and strength of evidence



CHAPTER 5. Discussion

5.1 RQ 1: Extent of the literature on multilingual CLIL

This review found a total of 19 studies from 1994 to 2021 that reported on educational outcomes in multilingual CLIL programs, although the first study in this range was published in 2008. This indicates that, despite the existence of trilingual programs for several decades (Cenoz, 2009; Eurydice, 2006), research has not flourished in this field. The inclusion criteria of this review were intentionally broad to capture published work or gray literature with any study design reporting on a variety of educational outcomes. Even so, the search only resulted in an average of a little over one study per year in the last 13 years. Other systematic reviews investigating bilingual immersion have lamented the scarcity of empirical studies outside of North America (Banegas et al., 2020) or criticized the validity of studies, calling for more rigorous methodologies and longitudinal designs (Bruton, 2011; Goris et al., 2019). However, even when considering a 21-year span and including all studies regardless of publication status or methodological rigor, this review still found relatively few studies investigating multilingual CLIL programs.

5.1.1 Literature by age group

Only two studies in this review were conducted in primary schools. Even in Spain, where this review found 17 empirical studies, no work investigated outcomes of multilingual education in primary schools. This is not to say that multilingual CLIL programs do not exist in primary schools; on the contrary, instantiations of such programs targeted at young learners can be found all over the globe (Cenoz, 2009; Murphy, 2014). What is more, the two studies in this review that did focus on a primary age group employed qualitative and mixed methods designs aimed at assessing stakeholder satisfaction with the educational program. Both studies, in very different geographic contexts, found positive impressions by the students, teachers, and parents involved in the programs. Overall, more research should consider the educational outcomes for students in primary multilingual CLIL. While research into attitudes is useful, there is also a need for quantitative empirical studies that evaluate outcomes other than attitudes.

5.1.2 Literature by geographic region

With 17 of the 19 studies in this review coming from Spain, it is clear that the literature concerning multilingual CLIL is lacking for other geographic regions. Even though this

review aimed to capture a broad scope of literature by using a variety of search terms to include diverse global contexts (see 3.4), this dissertation's findings align with other systematic reviews investigating bilingual CLIL (e.g. Goris et al., 2019; Graham et al., 2018) that also reported a disproportionate research output from Spain when compared to other countries.

Despite the wide variety of multilingual educational settings described in 2.4, many of which include three MOIs, the research found in this review does not reflect this diversity. These findings of a wide body of research from Spain may reveal a true imbalance in the global research output. However, it should be noted that the search conducted for this review was piloted with a pre-determined exemplar study (Merino & Lasagabaster, 2018) that was conducted in Spain. Moreover, six studies in this review were found via forward and backward citation searching, and all six were likewise investigating schools in Spain. Therefore, the search may have been biased to select for studies conducted in this context.

Nonetheless, the search did identify several works describing educational outcomes in multilingual schools outside of Spain that met this review's criteria. However, none of these works were primary empirical studies that reported research questions and methods, and so they were excluded. For instance, Björklund (2005) describes the situation of trilingual CLIL in Swedish, Finnish and English in Vaasa, Finland. Furthermore, Gorter (2005) similarly describes the context of Friesland, Netherlands where Frisian (a regional language), Dutch, and English are used as LOIs. While both works report small pieces of relevant data, neither is a primary empirical study. Each of these works also cites a book chapter that further describes the educational context and selectively references data, though again without reporting on primary research (Björklund & Suni, 2000; Ytsma, 2000). As indicated in 3.6.2, one study concerning trilingual education in Friesland (Ytsma, 1997) could not be retrieved. Thus, no primary studies in either the Netherlands or Finland were found. Despite research apparently being conducted in these regions (Cenoz, 2009), this review turned up very few relevant results, nor was the reviewer able to find any through independent searching.

One study (Schietroma, 2019) reported relevant outcomes from research in a multilingual school in Italy, however complete results and information concerning the methodology were not available. The author did not reply to an email request for more

information from the original dissertation conducting this research, nor could this unpublished dissertation (Schietroma, 2008) be retrieved online.

A further study (Terlević Johansson, 2013) did meet the requirements of reporting on primary research, however the school in the study only offered L3 Swedish and L1 German as MOIs, while L2 English was offered as a subject. This review adopted stringent criteria for a multilingual program and so the study was excluded. However, given the generally high proficiency of German learners of English (European Commission, 2012), this research may well have represented a successful additive trilingual program.

Finally, a study in the Philippines (Walter & Dekker, 2011) offered results from a primary school that used Lilubuagen (a local language) and Filipino as MOIs, with English taught as a subject. In a context like Spain, it is easy to demarcate a line between majority and non-majority languages, given the relative linguistic homogeneity of the country (Cenoz, 2009). While Spain does have three regional minority languages, the situation is less complex than the linguistic context of the Philippines or other post-colonial regions with a high degree of multilingualism (Benson, 2021). With English as an official language of the Philippines, the school in Walter and Dekker's (2011) study presents another potential situation where additive trilingualism may be promoted.

Thus, this review highlights the large research output from Spain and underscores the paucity of work conducted in other regional contexts. While this may stem from a bias in the design of the search itself, similar findings from other systematic reviews conducted in the past five years suggest that this disparity may reflect a genuine imbalance in the literature. In contrast, a closer analysis of some of the excluded works from this review reveals that research in other geographic contexts such as the Netherlands and Finland seems to be escaping searches, while other schools that may be successfully developing multilingualism are likewise omitted due to the particular inclusion criteria of this study. In reference to this latter point, schools like those found in Germany and the Philippines would be captured in a review that considers programs that offer two MOIs and a third, non-majority language as a subject. However, this adjustment of inclusion criteria would simultaneously open the search to many schools that do not promote additive trilingualism. Thus, while it may be necessary to rethink the definition of a multilingual school to better fit a variety of contexts that differ from Spain, the development of objectively measurable criteria still proves difficult.

5.2 RQ 2: Effectiveness of multilingual CLIL

Regarding the effects of multilingual CLIL on educational outcomes, this review found generally positive outcomes for attitudes and L3 skills and no reported findings of any detrimental effects for L1 or academic content knowledge.

The literature in this field is relatively split between studies investigating attitudes on one hand and those reporting on language and content knowledge outcomes on the other. While Grisaleña et al. (2009) reported findings on attitudes alongside an assessment of L3 skills, this mixed methods study exhibited a high degree of risk of methodological bias and included few details regarding the qualitative elicitation of student opinions. Coyle et al. (2010) highlight that performance evidence (language and academic content knowledge outcomes) and affective evidence (attitudes) are the two most frequently considered facets when evaluating the impact of CLIL programs. However, the authors comment that there is a disjunction between these two areas and that these results are seldom reported together. More assessments of attitudes and language skills together would be a fruitful addition to this field, as the synthesis of these findings could further reveal strengths or weaknesses of CLIL pedagogy.

5.2.1 Language outcomes

L1 outcomes

Two studies reported findings on L1 outcomes. One of the studies with strong weight of evidence, San Isidro and Lasagabaster (2019), indicated that students in the CLIL section had higher scores in both L1 Spanish and Galician across all skills. In contrast, the moderate-strength study, Merino and Lasagabaster (2018), found significant baseline differences between CLIL and non-CLIL students and did not find a significant interaction of time and intervention. Given these contradictory findings from only two studies, it is difficult to offer a conclusive answer regarding L1 development in multilingual CLIL programs. One point of difference between these two studies is that participants were Galician-Spanish bilinguals in the former and Basque-Spanish bilinguals in the latter. Some research has shown that more closely related languages are typically easier to learn than completely unrelated languages (Bild & Swain, 1989). Therefore, it is possible that the null effect reported by Merino and Lasagabaster (2018) stems from the greater typological distance of Basque and Spanish when compared to Galician and Spanish. The factor of linguistic distance and its relationship with CLIL in a learner's developing multilingual language system should be further investigated.

Nonetheless, in alignment with research on bilingual programs, no research from this review on multilingual CLIL found any negative consequences of reduced exposure to the L1 in school on native language outcomes.

L2 outcomes

No study investigating language outcomes reported on L2 development, as all of these studies were conducted in Spain and none targeted students with three sequentially acquired languages. The participants in each study were all bilingual from birth in Spanish and Basque, Catalan, or Galician. Nonetheless, only seven of the studies addressed home language practices (see Appendix F), and the remaining studies grouped students together regardless of what languages they used outside of school. A closer analysis of the micro sociolinguistic context would be useful in future research to investigate whether discrepancies in this area lead to overall language proficiency differences in a multilingual CLIL environment.

L3 outcomes

As noted in 4.4, the results for L3 outcomes were strongly positive for the skills of reading, speaking, and writing. In each of these areas, only Merino and Lasagabaster (2018) reported no impact of multilingual CLIL on L3 performance. However, this latter study did in fact find differences between the CLIL and non-CLIL sections, with CLIL students significantly outperforming their non-CLIL peers in all areas. Merino and Lasagabaster (2018) argued that the baseline differences between groups precluded the determination of a causal relationship in favor of CLIL. Nonetheless, the positive findings from other studies that also use a longitudinal design (e.g., Pérez-Vidal & Roquet, 2015; Prieto-Arranz et al., 2015; San Isidro & Lasagabaster, 2019) provide compelling evidence that multilingual CLIL programs do benefit the skills of L3 reading, writing, and speaking when compared to traditional foreign language teaching.

The findings concerning L3 listening are more mixed, with two weak and one strong study reporting positive effects of CLIL and one weak and two moderate studies reporting no differences between CLIL and non-CLIL sections. Morphosyntax in oral production tasks represents another area where the benefits of CLIL are less clear, with more studies reporting no significant differences between groups than studies that reported a benefit for CLIL sections. Martínez Adrián and Gutiérrez Mangado (2015) found that the CLIL section in their study performed significantly worse than a non-CLIL control group, however these latter students were two years older and so not a

comparable group for analysis. The findings for listening and morphosyntax align with previous research on bilingual CLIL programs, that has also presented a combination of findings ranging from no significant differences between groups to benefits for CLIL students (see Merino & Lasagabaster, 2018).

Finally, a couple of studies reported significantly stronger results for CLIL students in pragmatic oral skills (Gutiérrez Mangado & Martínez-Adrián, 2018; Herraiz Martinez & Sánchez Hernández, 2019) and the use of L1-based CSs (Ollo Jiménez & Martínez-Adrián, 2019). These initial findings indicate potential benefits of multilingual CLIL for these target language skills, however more research must be conducted to confirm the results of these studies.

Overall, the picture of L3 outcomes is not significantly different from that of L2 outcomes in bilingual CLIL programs. The findings retrieved by this review reveal little to no evidence indicating that participation in CLIL has any negative impacts on L3 development, and indeed several studies indicate that CLIL has very positive effects for the skills of reading, speaking, and writing.

5.2.2 Academic content knowledge outcomes

With only one study reporting measures of academic content knowledge, little can be said as to the impact of multilingual CLIL on this educational outcome. This represents an important area of future research. Moreover, only half of the studies in this review controlled the academic subject(s) taught in a target language. In contrast, the other half of studies did not, and students in the CLIL intervention group did not all study the same subject in the target language (see Appendix F). The investigation of whether subject allocation influences CLIL outcomes was beyond the scope of this study, though this is another area that should be further investigated.

5.2.3 Attitudes

The affective outcomes of attitudes were some of the strongest positive findings of this review. With only Amengual-Pizarro and Prieto-Arranz (2015) finding few significant differences between CLIL and non-CLIL students, six other studies found clear positive benefits of multilingual CLIL models, both in terms of attitudes towards multilingualism and program satisfaction. Moreover, these findings came from three countries (two continents) in both primary and secondary schools, which provides a preliminary indication that these positive findings may be applicable in a wider variety

of contexts. The studies reported herein corroborate the generally positive findings from bilingual CLIL programs and suggest that multilingual CLIL models are also conducive to the development of positive attitudes.

5.3 Limitations

This review made every effort to limit the impact of potential bias, however there are inevitably some limitations to this work. This section reviews some of the most significant potential sources of bias, before turning to overall conclusions in 5.4.

In designing this review, search terms were piloted against an exemplar article taken from the Spanish context. As discussed in 5.1.2, this may have biased the search to disproportionately select for other studies conducted in this same region. In addition, two studies (Schietroma, 2008; Ytsma, 1997) could not be retrieved, despite indications that they may have met the inclusion criteria of this review. Therefore, it may be that some research from other relevant multilingual schools was not captured through this review's search.

One study (Ní Dhiorbháin, 2020) was published in Irish, a language not spoken by the main reviewer. While a proficient speaker of Irish was consulted to corroborate that the Google translation offered a high standard of accuracy, this speaker only assessed the translation of the study's abstract. Google Translate has been shown to systematically produce grammatical errors when used for academic texts, though less research has investigated the faithful transmission of intelligible messages (Groves & Mundt, 2015). Even though a partial translation of Ní Dhiorbháin's (2020) study was evaluated for accuracy, the use of an online translation tool introduces a risk of bias given the unconfirmed reliability of the full-text translation.

Regarding the assessment of bias of each study, this review used the MMAT so as to make direct comparisons between the weight of evidence of studies with quantitative, qualitative, and mixed methods designs. Despite the advantage of being able to make direct comparisons, there are still inherent differences to each research design that are obscured by their direct juxtaposition. Moreover, while the two reviewers reached a sufficient level of agreement for two of the three studies that were double screened, the interrater reliability was below 0.7 for the mixed methods study assessed. Therefore, the subjectivity of the risk of bias assessments of individual studies is a source of potential weakness in this review.

5.4 Conclusion

This systematic review investigated the extent of research into multilingual CLIL environments across the world. The resulting studies represent a small but steady output of research beginning in 2008. The research comes almost exclusively from secondary schools in Spain, and work with other age groups and in diverse geographic contexts is sorely needed.

The 19 studies found by this search presented different levels of risk of bias, though there were few systematic patterns of weakness across the group. As such, the overall weight of evidence is moderate. However, the generalizability of the findings is extremely low, as research on multilingual schools outside of Spain was largely not found by this review.

The research on multilingual CLIL programs closely echoes the findings from research in comparable bilingual programs. Despite claims of CLIL only producing positive outcomes due to being an elitist or self-selecting form of education, even studies that controlled for baseline differences and matched participants between sections found promising results. In particular, CLIL appears to foster the development of positive attitudes towards other languages and cultures and promote target language development to a greater extent than traditional foreign language education. Crucially, there is still no evidence to indicate that CLIL programs have any negative repercussions on L1 abilities or academic content knowledge, although more work in this area is required.

This review has revealed a small, though moderately strong body of research on multilingual CLIL programs as they occur in Spain. Given the narrow geographic pool of this research, it would be unwise to generalize these findings to other contexts. However, the fact that outcomes from bilingual and multilingual schools are very similar bodes well for other regions where bilingual education exists. Moreover, the qualitative reports from Ireland and Hong Kong indicate high satisfaction with budding multilingual educational programs and a desire to develop these further. Given the growth of multilingualism worldwide, there is a critical need for research that investigates a broader range of school contexts to add to the burgeoning literature on educational programs that foster multilingualism and multiliteracy.

References[2]

Admiraal, W., Westhoff, G., & de Bot, K. (2006). Evaluation of bilingual secondary education in the Netherlands. *Educational Research and Evaluation*, *12*(1), 75–93.

Amengual-Pizarro, M., & Prieto-Arranz, J. I. (2015). Exploring affective factors in L3 learning: CLIL vs. non-CLIL. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 197–220). Springer.

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). DOI: 10.1037/0000165-000

Armijo-Olivo, S., Stiles, C. R., Hagen, N. A., Biondo, P. D., & Cummings, G. G. (2012). Assessment of study quality for systematic reviews: A comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: Methodological research. *Journal of Evaluation in Clinical Practice*, *18*(1), 12–18.

Baetens Beardsmore, H. & Lebrun, N. (1991). Trilingual education in the Grand Duchy of Luxembourg. In O. García (Ed.), *Bilingual Education* (pp. 107–120). John Benjamins.

Bale, J. (2010). International comparative perspectives on heritage language education policy research. *Annual Review of Applied Linguistics*, *30*, 42–65.

Baker, C., & Wright, W. E. (2017). *Foundations of bilingual education and bilingualism* (6th ed.). Multilingual Matters.

Banegas, D., Poole, P., & Corrales, K. (2020). Content and language integrated learning in Latin America 2008-2018: Ten years of research and practice. *Studies in Second Language Learning and Teaching*, *10*(2), 283–305.

Benson, C. (2021). L1-based multilingual education: What is working and what is slowing us down. In P. Harding-Esch & H. Coleman (Eds.), *Language and the Sustainable Development Goals* (pp. 17–29). British Council.

Bild, E.R. & Swain, M. (1989). Minority language students in a French Immersion programme: Their French proficiency. *Journal of Multilingual and Multicultural Development*, *10*(3), 255–274.

Björklund, S. (2005). Towards trilingual education in Vaasa/Vasa, Finland. *International Journal of the Sociology of Language, 2005*(171), 23–40.

Björklund, S. & Suni, I. (2000). The role of English as L3 in a Swedish immersion program in Finland: Impacts on language teaching and language relations. In U. Jessner and J. Cenoz (Eds.), *English in Europe: The acquisition of a third language* (pp. 198–221). Multilingual Matters.

Boland, A., Cherry, M., & Dickson, R. (2014). *Doing a systematic review: A student's guide*. SAGE Publications.

[2]This dissertation adheres to the citation conventions of the APA style guide, 7th edition (American Psychological Association, 2020).

Brunton, G., Stansfield, C., Caird, J., & Thomas, J. (2017). Finding relevant studies. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd edition, pp. 93–122). SAGE Publications.

Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, *39*(4), 523–532.

The Campbell Collaboration. (n.d.). Evidence synthesis tools for Campbell authors. Retrieved March 23, 2021, from https://www.campbellcollaboration.org/research-resources/resources.html

Cenoz, J. (1998). Multilingual education in the Basque Country. In J. Cenoz and F. Genesee (Eds.), *Beyond bilingualism: Multilingualism and multilingual education* (pp. 175–191). Multilingual Matters.

Cenoz, J. (2009). *Towards multilingual education: Basque educational research from an international perspective*. Multilingual Matters.

Cenoz, J. (2015). Content-based instruction and content and language integrated learning: The same or different? *Language, Culture, and Curriculum*, *28*(1), 8–24.

Cenoz, J. (2017). Translanguaging in school contexts: International perspectives. *Journal of Language, Identity & Education*, *16*(4), 193–198.

Cenoz, J., Genesee, F., & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. *Applied Linguistics*, *35*(3), 243–262.

Cenoz, J. & Valencia, J. F. (1994). Additive trilingualism: Evidence from the Basque Country. *Applied Psycholinguistics*, *15*(2), 195–207.

Clyne, M., Hunt, C.R. & Isaakidis, T. (2004). Learning a community language as a third language. *International Journal of Multilingualism*, *1*(1), 33–52.

Cochrane Effective Practice and Organisation of Care (EPOC). (2017). Data collection form. Retrieved 11 August, 2021, from https://epoc.cochrane.org/resources/epoc-resources-review-authors#conducting

Cohen, J (1988). *Statistical Power Analysis for the Social Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge University Press.

Crawford, J. (2008). *Advocating for English language learners: Selected essays.* Multilingual Matters.

Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. *Journal of Clinical Epidemiology*, *64*(1), 79–89.

Dai, Q. & Cheng, Y. (2007). Typology of bilingualism and bilingual education in Chinese minority nationality regions. In A. Feng (Ed.), *Bilingual education in China* (pp. 75–93). Multilingual Matters.

Dalton-Puffer, C., Nikula, T., & Smit, U. (Eds.). (2010). *Language use and language learning in CLIL classrooms*. John Benjamins.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research & Policy*, *10*(1), 45–53.

Duarte, J. (2020). Translanguaging in the context of mainstream multilingual education. *International Journal of Multilingualism, 17*(2), 232–247.

EPPI-Centre. (n.d.). EPPI-Centre database of education research. Retrieved March 23, 2021, from http://eppi.ioe.ac.uk/cms

Etxeberria, F. (1999). *Bilingüismo y educación en el País del Euskara.* Erein.

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

Eurydice. (2006). *Content and Language Integrated Learning (CLIL) at school in Europe.* European Commission.

Fernández-Sanjurjo, J., Fernández-Costales, A., & Arias Blanco, J. M. (2019). Analysing students' content-learning in science in CLIL vs. non-CLIL programmes: Empirical evidence from Spain. *International Journal of Bilingual Education and Bilingualism*, *22*(6), 661–674.

Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. SAGE Publications.

Fung, D., & Yip, V. (2014). The effects of the medium of instruction in certificate-level physics on achievement and motivation to learn. *Journal of Research in Science Teaching*, *51*(10), 1219–1245.

García, O. (2009). *Bilingual education in the 21st Century: A global perspective.* Wiley-Blackwell.

García, O. (2014). Multilingualism and language education. In C. Leung & B. V. Street (Eds.), *The Routledge companion to English studies* (1st ed., pp. 84–99). Routledge.

García, O. & Wei, L. (2014). *Translanguaging: Language, bilingualism, and education.* Palgrave Macmillon.

García Mayo, M. del P., & Villarreal Olaizola, I. (2011). The development of suppletive and affixal tense and agreement morphemes in the L3 English of Basque-Spanish bilinguals. *Second Language Research*, *27*(1), 129–149.

Genesee, F. (1987). *Learning through two languages: Studies of immersion and bilingual education.* Newbury House.

Genesee, F. (2006). Bilingual first language acquisition in perspective. In P. McCardle & E. Hoff (Eds.), *Childhood bilingualism: Research on infancy through school age* (pp. 45–67). Multilingual Matters.

Genesee, F. (2008). Dual language in the global village. In T. W. Fortune & D. J. Tedick (Eds.), *Pathways to multilingualism: Evolving perspectives on immersion education* (pp. 22–45). Multilingual Matters.

González Gándara, D. (2015). CLIL in Galicia: Repercussions on academic performance. *Latin American Journal of Content & Language Integrated Learning*, *8*(1), 13–24.

Goris, J., Denessen, E., & Verhoeven, L. (2019). Effects of content and language integrated learning in Europe: A systematic review of longitudinal experimental studies. *Behavioural Science Institute*, *18*(6), 675–698.

Gorter, D. (2005). Three languages of instruction in Fryslân. *International Journal of the Sociology of Language, 2005*(171), 57–73.

Gough, D., Oliver, S., & Thomas, J. (2017). Introducing systematic reviews. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An Introduction to Systematic Reviews* (2nd ed., pp. 1–17). SAGE Publications.

Graham, K. M., Choi, Y., Davoodi, A., Razmeh, S., & Dixon, L. Q. (2018). Language and content outcomes of CLIL and EMI: A systematic review. *Latin American Journal of Content & Language Integrated Learning*, *11*(1), 19–37.

Grisaleña, J., Alonso, E., & Campo, A. (2009). Enseñanza plurilingüe en centros de educación secundaria: Análisis de resultados [Plurlingual education in secondary schools: Analysis of results]. *Revista Iberoamericana de Educación, 49*(1), 1–12.

Groves, M. & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes, 37,* 112–121.

Gutiérrez Mangado, M., & Martínez-Adrián, M. (2018). CLIL at the linguistic interfaces. *Journal of Immersion and Content-based Language Education*, *6*(1), 85–112.

Hélot, C. & de Mejía, A. (2008). *Forging multilingual spaces: Integrated perspectives on majority and minority bilingual education*. Multilingual Matters.

Herraiz Martinez, A. & Sánchez Hernández, A. (2019). Pragmatic markers produced by multilingual speakers: Evidence from a CLIL context. *English Language Teaching*, *12(*2), 68–76.

Heugh, K., Benson, C., Bogale, B., & Gebre Yohannes, M. (2012). Implications for multilingual education: Student achievement in different models of education in Ethiopia. In T. Skutnabb-Kangas & K. Heugh (Eds.), *Multilingual education and sustainable diversity work: From periphery to centre* (pp. 239–262). Routledge.

Hong, Q.N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I. (2019). Improving the content validity of the Mixed Methods Appraisal Tool (MMAT): A modified e-Delphi study. *Journal of Clinical Epidemiology*, *111*, 49–59.

Hughes, S. P., & Madrid, D. (2020). The effects of CLIL on content knowledge in monolingual contexts. *Language Learning Journal*, *48*(1), 48–59.

International Database of Education Systematic Reviews. (n.d.). About IDESR. Retrieved March 23, 2021, from https://idesr.org/#aboutuspage

Juan-Garau, M. & Salazar-Noguera, J. (2015). *Content-based language learning in multilingual educational environments*. Springer.

Kalizhanova, A., Maryshkina, T., Ishmuratova, M., Ibrayeva, B., & Sembiyev, K. (2021). A trilingual e-dictionary of biological terms for paleobiological and historical research in Kazakhstan. *Historical Biology*, *33*(5), 639–642.

Lambert, W. E. (1981). Bilingualism and language acquisition. In J. Winitz (Ed.), *Native language and foreign language acquisition* (pp. 9–22). The New York Academy of Sciences.

Lambert, W. E. & Tucker, G. R. (1972). *The bilingual education of children: The St. Lambert experiment*. Newbury House.

Lasagabaster, D. (2009). The implementation of CLIL and attitudes towards trilingualism. *International Journal of Applied Linguistics*, 157(1), 23–43.

Lasagabaster, D., & Sierra, J. M. (2010). Immersion and CLIL in English: More differences than similarities. *ELT Journal*, *64*(4), 367–375.

Lázaro Ibarrola, A. (2011). Faster and further morphosyntactic development of CLIL vs. EFL Basque-Spanish bilinguals learning English in high-school. *International Journal of English Studies*, *12*(1), 79–96.

Li, D., & Tong, C. L. (2020). A tale of two Special Administrative Regions: The state of multilingualism in Hong Kong and Macao. In H. Klöter & M. S. Saarela (Eds.), *Language diversity in the Sinophone world* (1st ed., pp. 142–163). Routledge.

Li, Y., Chen, W., Liu, G., Wei, L., Liu, C., Si, L., Zhang, J., & Yang, K. (2017). Analysis of quality assessment tools in Campbell systematic reviews. *Abstracts of the Global Evidence Summit*, *9 Supp 1*.

Liabo, K., Gough, D., & Harden, A. (2017). Developing justifiable evidence claims. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 251–277). SAGE Publications.

Lightfoot, A., Balasubramanian, A., Tsimpli, I., Mukhopadhyay, L, & Treffers-Daller, J. (2021). Measuring the multilingual reality: Lessons from classroom in Delhi and Hyderabad. *International Journal of Bilingual Education and Bilingualism*, 1–21. DOI: 10.1080/13670050.2021.1899123

Lindholm-Leary, K. (2016). Students' perceptions of bilingualism in Spanish and Mandarin dual language programs. *International Multilingual Research Journal*, *10*(1), 59–70.

L'nyavskiy-Ekelund, S. & Siiner, M. (2017). Fostering social inclusion through multilingual habitus in Estonia: A case study of the Open School of Kalamaja and the Sakala Private School. *Social Inclusion*, *5*(4), 98–107.

Lo, Y. Y., & Fung, D. (2020). Assessments in CLIL: The interplay between cognitive and linguistic demands and their progression in secondary education. *International Journal of Bilingual Education and Bilingualism*, *23*(10), 1192–1210.

Maljers, A., Marsh, D., & Wolff, D. (2007). *Windows on CLIL: Content and language integrated learning in the European spotlight*. European Platform for Dutch Education.

Marsh, D. (2000). *Using Languages to Learn and Learning to Use Languages*. University of Jyväskylä.

Martínez Adrián, M. & Gutiérrez Mangado, J. (2015). Is CLIL instruction beneficial in terms of general proficiency and specific areas of grammar? *Journal of Immersion and Content-Based Language Education*, *3*(1), 51–76.

Meganathan, R. (2011). Language policy in education and the role of English in India: From library language to language of empowerment. In H. Coleman (Ed.), *Dreams and realities: Developing countries and the English language* (pp. 57–86). British Council.

Menken, K. & Kleyn, T. (2010). The long-term impact of subtractive schooling in the educational experiences of secondary English language learners. *International Journal of Bilingual Education and Bilingualism*, *13*(4), 399–417.

Merino, J. A., & Lasagabaster, D. (2018). CLIL as a way to multilingualism. *International Journal of Bilingual Education and Bilingualism*, *21*(1), 79–92.

Merisuo-Storm, T. (2007). Pupils' attitudes towards foreign-language learning and the development of literacy skills in bilingual education. *Teaching and Teacher Education*, *23*(2), 226–235.

Meyerhöffer, N., & Dreesmann, D. C. (2019). The exclusive language of science? Comparing knowledge gains and motivation in English-bilingual biology lessons between non-selected and preselected classes. *International Journal of Science Education*, *41*(1), 1–20.

Ministerio de Educación y Ciencias. (2018). Resolución 31,531. Retrieved 21 July, 2021, from https://www.schooloftomorrowparaguay.com/mec

Mohanty, A. (2006). Multilingualism of the unequals and predicaments of education in India: Mother tongue or other tongue? In O. Garcia, T. Skutnabb-Kangas, & M. Torres Guzman (Eds.), *Imagining Multilingual Schools: Language in Education* (pp. 262–283). Multilingual Matters.

Moher, D., Liberati A., Tetzlaff J., Altman D. G., The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. PLoS Med *6*(7): e1000097. DOI:10.1371/journal.pmed1000097

Moher, D., Pham, B., Lawson, M. L., & Klassen, T. P. (2003). The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technology Assessment*, *7*(41), 1–90.

Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., Mierzwinski-Urban, M., Clifford, T., Hutton, B., & Rabb, D. (2012). The effect of English-language restriction on systematic review-based meta-analyses: A systematic review of empirical studies. *International Journal of Technology Assessment in Health Care*, *28*(2), 138–144.

Murphy, V. A. (2014). *Second language learning in the early school years: Trends and contexts*. Oxford University Press.

Ní Dhiorbháin, A. (2020). Múineadh ábhar tríd an tríú teanga i mbunscoileanna Lán-Ghaeilge [Teaching a subject through a third language in Irish-medium primary schools]. *TEANGA, The Journal of the Irish Association for Applied Linguistics*, *27*, 1–21. DOI: 10.35903/teanga.v27i.484

Nikula, T. (2017). CLIL: A European approach to bilingual education. In N. Van Deusen-Scholl & S. May (Eds.), *Second and foreign language education* (pp. 111–124). Springer International Publishing.

Ollo Jiménez, P. & Martínez-Adrián, M. (2019). A study of self-reported opinions of L1-based communications strategies in CLIL and non-CLIL secondary-school learners of L3 English. *RAEL: Revista Electrónica de Lingüistica Aplicada*, *18*(1), 72–90.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 1–10.

Padilla, A., Fan, L., Xu, X., & Silva, D. (2013). A Mandarin/English two-way immersion program: Language proficiency and academic achievement. *Foreign Language Annals*, *46*(4), 661–679.

Pérez Cañado, M. L. (2020). What's hot and what's not on the current CLIL research agenda: Weeding out the non-issues from the real issues. A response to Bruton (2019). *Applied Linguistics Review*. 1–21. DOI: 10.1515/applirev-2020-0033.

Pérez-Vidal, C. & Roquet, H. (2015). The linguistic impact of a CLIL *Science* programme: An analysis measuring relative gains. *System*, *54*, 80–90.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing.

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

Pluye, P. & Hong, Q. N. (2014). Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. *Annual Review of Public Health, 35*, 29–45.

Prieto-Arranz, J. I., Rallo Fabra, L., Calafat-Ripoll, C., & Catrain-González, M. (2015). Testing progress on receptive skills in CLIL and non-CLIL contexts. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 123–137). Springer.

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

Ruiz de Zarobe, Y., & Jiménez Catalán, R. M. (2009). *Content and language integrated learning: Evidence from research in Europe*. Multilingual Matters.

San Isidro, X. (2010). An insight into Galician CLIL: Provision and results. In D. Lasagabaster & Y. Ruiz de Zarobe (Eds.), *CLIL in Spain: Implementation, results and teacher training* (pp. 55–78). Cambridge Scholars.

San Isidro, X. (2017). *CLIL in a multilingual setting: A longitudinal study on students, families and teachers* [Doctoral thesis, University of the Basque Country: Vitoria-Gasteiz].

San Isidro, X. & Lasagabaster, D. (2019). The impact of CLIL on pluriliteracy development and content learning in a rural multilingual setting: A longitudinal study. *Language Teaching Research*, *23*(5), 584–602.

San Isidro, X. & Lasagabaster, D. (2020). Students' and families' attitudes and motivations to language learning and CLIL: A longitudinal study. *The Language Learning Journal*. DOI: 10.1080/09571736.2020.1724185

Schietroma, E. (2008). *Strategie didattiche per l'insegnamento delle scienze della terra* [Unpublished dissertation, Sapienza Università di Roma: Roma].

Schietroma, E. (2019). Innovative STEM lessons, CLIL and ICT in multicultural classes. *Journal of e-Learning and Knowledge Society*, *15*(1), 183–193.

Shameem, N. (2007). Language education needs for multilingualism in Fiji primary schools. *International Journal of Educational Development*, *27*(1), 39–60.

Sierra, J. & Olaziregi, I. (1989). *EIFE 2: Influence of factors on the learning of Basque*. Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia.

Skutnabb-Kangas, T. & Heugh, K. (2013). *Multilingual education and sustainable diversity work: From periphery to center.* Routledge.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5–11.

Terlević Johansson, K. (2013). Successful learning in L3 German through CLIL? Findings from a study on the oral production of Swedish pupils in lower secondary school. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, *18(*2), 15–26.

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., . . . Kavanagh, J. (2004). Integrating qualitative research with trials in systematic reviews. *British Medical Journal, 328*(7446), 1010–1012.

Thomas, W. P. & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Center for Research on Education, Diversity and Excellence.

Trujillo, O. (1997). A tribal approach to language and literacy in a trilingual setting. In J. Reyhner (Ed.), *Teaching Indigenous Languages. Selected Papers from the Annual Symposium on Stabilizing Indigenous Languages* (pp. 10–21). Northern Arizona University: Center for Excellence in Education.

UK Department for Education. (2021). Schools, pupils and their characteristics. Retrieved August 1, 2021, from https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics.

University of York. (n.d.). About PROSPERO. Retrieved March 23, 2021, from https://www.crd.york.ac.uk/prospero/#aboutpage

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

Walter, S. & Dekker, D. (2011). Mother tongue instruction in Lubuagan: A case study from the Philippines. *International Review of Education*, *57*(5), 667–683.

Wang, L. & Kirkpatrick, A. (2013). Trilingual education in Hong Kong primary schools: A case study. *International Journal of Bilingual Education and Bilingualism*, *16(*1), 100–116.

Wedikkarage, L. (2018). The effectiveness of teaching science subjects through English medium: A narrative analysis of teacher experiences and perceptions in Sri Lankan secondary schools. *International Review of Education*, *64*, 779–801.

Williams, C. (2001) Welsh in Great Britain. In G. Extra & D. Gorter (Eds.), *The other languages of Europe* (pp. 59–81). Multilingual Matters.

Ytsma, J. (1997). The Trilingual School. *De Pompebleden*, *68*(5), 70–71.

Ytsma, J. (2000). Trilingual primary education in Friesland. In U. Jessner and J. Cenoz (Eds.), *English in Europe: The acquisition of a third language* (pp. 222–235). Multilingual Matters.

Ytsma, J. (2001). Towards a typology of trilingual primary education. *Journal of Bilingual Education and Bilingualism*, *4*(1), 11–22.

Appendix A. IDESR protocol document

Title

Educational outcomes in multilingual CLIL school settings. Protocol for a systematic review.

Review Question

1. How many original research papers have been published on educational outcomes in multilingual school settings?
    a. How many papers have been published in public/private educational contexts?
    b. How many papers have been published in each region of the world?
2. What is the effectiveness of content-based instruction (CBI) in multilingual schools?
    a. What impact does CBI have on language outcomes (L1, L2, or L3)?
    b. What impact does CBI have on content knowledge outcomes?
    c. What impact does CBI have on other educational outcomes such as motivation or views towards multiculturalism?

Rationale

There exists a breadth of research on content-based instruction in bilingual environments, though less work has been done in multilingual settings, or schools that actively encourage the development of trilingualism and triliteracy, usually through the inclusion of three languages as media of instruction. This review seeks to assess the extent of the literature in this area and synthesize any results. This research is relevant given the prevalence of multilingual communities worldwide and given the increasing global pressure to master several languages. Content-based instruction is a pedagogical method that could be incredibly promising for multilingual communities, as an effective way of fostering multilingualism and multiliteracy. However, without research, the outcomes of such programs are still unclear.

Inclusion Criteria

Bibliographic information

Include 1: Studies with a full reference or sufficient information.

Exclude 1: Studies with insufficient bibliographic information.

Rationale: Without sufficient bibliographic information, retrieval of works is infeasible.

Date of publication

Include 2: Studies published on or after January 1, 1994.

Exclude 2: Studies published before January 1, 1994.

Rationale: The term CLIL was first used in the European context in 1994 (Nikula, 2017). While other forms of content and language integrated learning existed before this time, the European Union's renewed focus on language education increased research efforts in the field, particularly in multilingual contexts. This study aims to assess modern research in this area.

Participants

Include 3: Studies on typically developing foreign language learners. Include studies even if no explicit reference is made to learning ability if reasonable assumption can be made that participants are comprised mainly of typically developing individuals.

Exclude 3: Studies that exclusively target non-typically developing learners or learners with Developmental Language Disorder.

Rationale: This review seeks to assess effectiveness of CLIL methodology as applies to typically developing school populations. The findings for non-typically developing individuals may not hold for a larger population, and so these latter results should not be extrapolated, nor will they be included in this review.

Include 4: Studies conducted in primary or secondary schools (students aged 5-18).

Exclude 4: Studies in early years, university, or adult educational contexts.

Rationale: The language learning capacity and the educational goals for very young (under 5) or adult (over 18) learners are quite different from those learners in primary or secondary education. This study focuses on the outcomes of CLIL programs in these middle year programs, where learners are well suited to learn another language.

Intervention

Include 5: Studies involving schools where three languages are used as language of instruction for 2+ hours a week beyond traditional language arts classes. Alternatively, two languages may be used where they are L2 and L3.

Exclude 5: Studies on schools that support language proficiency and literacy in fewer than three languages for all students.

Rationale: The focus of this study is on multilingual CLIL programs, where schools actively promote language proficiency and literacy in more than two languages. Schools offering only two languages of instruction (L2 and L3) will be included with the understanding that, as the majority community language, L1 proficiency and literacy are sufficiently supported in language arts classes.

Include 6: Studies where learners received regular educational intervention (CLIL program) for at least one year.

Exclude 6: Studies where participants only engage in short-term CLIL projects or units.

Rationale: This study aims to highlight the long-term impact on educational outcomes for students who partake in CLIL educational models. Requiring at least a year of exposure allows for the effects of the pedagogical form to take shape.

Outcome

Include 7: Primary research studies reporting any measure of CLIL program effectiveness, including but not limited to language outcomes, content knowledge outcomes, or stakeholder attitudes. Include studies reporting either quantitative or qualitative outcomes.

Exclude 7: Systematic reviews and studies that provide narrative evaluation of an educational program but provide no explicit measures of program success.

Rationale: A synthesis of empirical findings in this field of literature is impossible without the reporting and evaluation of concrete data.

Publication status

Do not exclude studies based on publication status. Include gray literature. This paper seeks to offset potential publication bias by including a wider range of research, including gray literature.

Study design

Do not exclude studies based on the research design. A broad range of study designs are employed in assessing CLIL methodology, and the exclusion of any one may provide a limited view of the research in this area.

Language of publication

Do not exclude studies based on the language of publication. Limiting to studies written in English may result in a systematic neglect of a certain body of research.

Information Sources

The list of databases consulted for this systematic review can be found below. This list includes prominent databases in the fields of education, linguistics, psychology, as well as multidisciplinary sources to cover wider fields of the social sciences. Furthermore, searches were conducted on platforms such as OpenGrey and ProQuest Dissertations & Theses Global to encompass any gray literature not present in the other databases. All databases were accessed electronically via subscription from the University of Oxford's Bodleian Library.

Education: ProQuest Education Collection (including ERIC)

Linguistics: ProQuest Linguistics Collection (including LLBA)

Psychology: PsychInfo

Multidisciplinary: Web of Science, Scopus

Gray literature: OpenGrey, ProQuest Dissertations & Theses Global

The initial search was followed by both forward and backward citation searches, whereby any study that met all the inclusion criteria was reviewed further to find other relevant matches. In the backward citation search, an eligible study's references were reviewed and compared against the selection criteria. For the forward citation search, all studies found on Web of Science to have cited an eligible paper were similarly reviewed.

Search Strategy

For this study, an experienced librarian at the University of Oxford's Department of Education was consulted to formulate the initial search. The two elements of mode of instruction (CLIL) and multilingualism were identified as the crux that should underpin the search. Following pilot scoping searches in the Proquest Education Collection, it was found that many results targeted university or adult participants. To account for this, a field specifying target participants was added, narrowing down the focus of results. A number of similar labels were included within each of these three categories to encompass the variability of terminology. Different terms within each category were connected with the operator 'OR' and each field was joined with 'AND.'

The first two search fields were applied to abstracts only, after piloting revealed a large number of articles being flagged exclusively due to journal title, such as Multilingual Matters. The field of participant was searched in all fields except full text, where possible. The resulting complete Boolean search string follows: ab(immersion OR CLIL OR "content and language integrated learning" OR CBI OR "content based instruction" OR CBLT OR "content based language teaching" OR EMI OR "English Medium Instruction" OR "language of instruction" OR "medium of instruction") AND ab(plurilingual* OR multilingual* OR trilingual* OR L3 OR "third language") AND noft(primary OR secondary OR "high school" OR elementary OR adolescent* OR child*)

Data Management

Upon completing the final search, all relevant information will be uploaded to Rayyan, a software program for systematic reviews where multiple collaborators can compare abstracts and other information against eligibility criteria (Ouzzani et al., 2016). There, duplicates will be eliminated and initial screening was conducted. Bibliographic information will be uploaded and organized on Mendeley reference manager. Detailed notes documenting the search process and the origin of each study meeting the eligibility criteria will be kept in a physical research journal and a Microsoft Excel document, respectively.

All research will be conducted on a MacBook Air running Mac OS Big Sur Version 11.2.1.

Selection Process

After the initial search, duplicates will be eliminated in Rayyan. The remaining titles and abstracts will screened against the eligibility criteria and marked for full-text screening. To be considered for full-text screening, a given abstract must not explicitly violate one of the exclusion criteria, meriting a deeper investigation. In the instance that a given work does not meet a criterion, the screening will stop and the criterion in question noted.

Following the initial screening, full reports will be obtained from the relevant databases. Where complete texts are not available or where the full report fails to unambiguously satisfy all the inclusion criteria, the author will be contacted for further information. Only studies that clearly meet all the inclusion criteria will be selected for the review. The resulting list will then used for backward and forward citation, and so on, until no further studies are identified.

In order to ensure reliability of the main author's screening, a second reader who is well-versed in applied linguistics has been recruited and informed of the aims and criteria of this study. Due to time and resource constraints, the second screener will review 10% of abstracts and full texts. Both reviewers will be blind to the other's decisions at time of screening. Where disagreements occur, the two screeners will confer to reach a conclusion. A Kappa value of 0.7 is generally considered acceptable in the field of education (Frey, 2018). If this threshold value is not met, the two screeners will confer to understand and clear up the causes of disagreements, and the second screener will subsequently review an additional 10% of abstracts/full articles.

Data Collection Process

Before executing the complete search, a data extraction form was specifically designed for this study. This document was modeled after the Cochrane good practice data extraction form, chosen in particular for its flexibility and suitability for a wide range of study designs (Cochrane Effective Practice and Organisation of Care, 2017). The data collected encompasses the essential categories of participants, intervention, comparison, and outcomes (PICO; Petticrew & Roberts, 2006), and has been reorganized for the purposes of this study. Fields were added to reflect the dynamic language background of the participants' community and to indicate the extent and duration of exposure to CLIL pedagogy.

Upon confirmation that a given study meets all eligibility criteria, data will be extracted and recorded in an Excel document by the primary investigator. In the case of incomplete data in a report, the main author of a study will be contacted via email.

Data Items

The following list was formed for data extraction and piloted against the prospective study Merino and Lasagabaster (2018). Where there are multiple outcomes to be reported, the relevant section of the data extraction form will be copied and pasted to include the additional results.

> General: Date form completed, ID of person extracting data, Reference citation, Study author contact details, Publication type, Document source
>
> Study overview: Research questions, Study design, Data type, Study duration, Location and language of publication
>
> Participants: School setting (social and educational context), Population description, Method of recruitment into CLIL, Languages spoken, Age, Gender, Other relevant sociodemographics
>
> Intervention: Languages of instruction (hours/week), Length of exposure to CLIL pedagogy, Non-CLIL control, Number of participants, Class grouping, Baseline imbalances, Attrition
>
> Outcomes: Outcome type, Outcome name, Unit(s) of measure, Time points measured, Descriptive outcomes, Effect sizes

Risk of bias/trustworthiness of individual studies

This review follows Slavin's (1986) best-evidence synthesis method, which does not discriminate against any research design in study selection, but rather assesses methodological rigor and potential bias in its evaluation of the overall work. In other words, studies are not weighed equally in a vote counting method, and instead each study is critically appraised and assigned a value of methodological quality to reflect its internal and external validity (Boland et al., 2014). Given the comprehensible paucity of RCTs in educational research, the more inclusive best-evidence synthesis method avoids the risk of prematurely concluding that no research exists in a field, without sacrificing review quality.

As this review actively seeks a variety of research designs, it employs an instrument specifically adapted to the needs of mixed methods research, the Mixed Methods

Appraisal Tool (MMAT; Pluye & Hong, 2014). Crowe and Sheppard (2011) found that the MMAT was the only critical appraisal tool at the time that was designed for a systematic mixed studies review. The MMAT, updated in 2018 with the feedback of over fifty experts to improve content validity, provides five different categories for assessing research methodology: qualitative, quantitative randomized controlled trials, quantitative non-randomized, quantitative descriptive, and mixed methods (Hong et al., 2019). The tool contains five criteria across each of the five categories, which are marked 'yes,' 'no,' or 'can't tell,' and an additional comment box to justify ratings. Therefore, a study receives a quality score of 0-5 in a given category and is scored on more than one category if it reports mixed methods outcomes.

As with initial screening, a second reader with conduct 10% of the risk of bias assessments.

Data Synthesis

Should there be enough data with comparable outcomes, meta analysis will be conducted using SPSS. If the resulting body of research is too small or diverse in outcome measures, a narrative synthesis of study quality and findings will be employed. Thomas et al. (2004) contend that the inclusion of qualitative information helps to triangulate the findings of quantitative data, and as such the resulting dialogue between these two outcomes in a narrative synthesis is highly informative.

In the latter case, the review would follow the three-step narrative synthesis structure outlined by Petticrew and Roberts (2006): (i) studies are grouped into logical categories; (ii) the findings and quality within each set is analyzed; and finally (iii) the findings among all groups are synthesized

Confidence in cumulative evidence

The strength of the cumulative evidence in this area of research will be assessed based on the results of the individual trustworthiness of each study as measured by the Mixed Methods Appraisal Tool.

Appendix B. Data extraction form

	Item	Data	Description
General	Date form completed		Use format dd/mm/yyyy
	Reference citation		List the full reference
	Study author contact details		
	Publication type		e.g. full report, abstract, letter
	Document source		Source database or website
Study Overview	Research questions		
	Study design		e.g. RCT, observational study
	Data type		Quantitative/qualitative
	Study duration		Include start date, end date, and duration, if possible
	Location and language of publication		
Participants	School setting (social and educational context)		e.g. primary, secondary, public, private
	Population description		Include any information regarding participants' learning disabilities, socioeconomic background, etc
	Method of recruitment into CLIL		How were students admitted into CLIL sections? (e.g. random allocation, interview, test scores)
	Languages spoken		Indicate L1/L2/L3, majority/minority/foreign, and proficiency level at beginning of the study in each language, as appropriate.
	Age		What is the age range and the number at each age?
	Gender		Include gender breakdowns where available.
	Other relevant sociodemographics		

Intervention	Languages of instruction (hours/week)		How many hours is each language used as a *language of instruction*? What subjects were taught in which languages?
	Length of exposure to CLIL pedagogy		Duration that intervention groups have been participating in CLIL lessons (indicate time before study and study intervention duration)
	Non-CLIL control		Was any control group included in the study? If so, what distinguished them from the intervention group?
	Number of participants		*n*= total number of participants
			n= CLIL intervention groups
			n= non-CLIL control groups
	Class grouping		Were participants grouped in classes? Describe differences.
	Baseline imbalances		Any significant differences at the beginning of the study?
	Attrition		Did any participants leave the study? How many, and for what reasons?
Outcomes	Outcome type		Language skills (L1/L2/L3), content knowledge, attitudes, or other
	Outcome name		e.g. vocabulary knowledge, speaking proficiency, understanding of biology terminology relating to the cell, attitudes towards learning through another language, etc.
	Unit(s) of measure		How is the outcome being operationalized?
	Time points measured		For longitudinal studies, how often were data taken?
	Descriptive outcomes		
	Effect sizes		

Appendix C. Mixed Methods Appraisal Tool (MMAT) - Version 2018

Category of study designs	Methodological quality criteria	Responses			
		Yes	No	Can't tell	Comments
1. Qualitative	1.1. Is the qualitative approach appropriate to answer the research question?				
	1.2. Are the qualitative data collection methods adequate to address the research question?				
	1.3. Are the findings adequately derived from the data?				
	1.4. Is the interpretation of results sufficiently substantiated by data?				
	1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation?				
2. Quantitative randomized controlled trials	2.1. Is randomization appropriately performed?				
	2.2. Are the groups comparable at baseline?				
	2.3. Are there complete outcome data?				
	2.4. Are outcome assessors blinded to the intervention provided?				
	2.5 Did the participants adhere to the assigned intervention?				
3. Quantitative non-randomized	3.1. Are the participants representative of the target population?				
	3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?				
	3.3. Are there complete outcome data?				
	3.4. Are the confounders accounted for in the design and analysis?				
	3.5. During the study period, is the intervention administered (or exposure occurred) as intended?				
4. Quantitative descriptive	4.1. Is the sampling strategy relevant to address the research question?				
	4.2. Is the sample representative of the target population?				
	4.3. Are the measurements appropriate?				
	4.4. Is the risk of nonresponse bias low?				
	4.5. Is the statistical analysis appropriate to answer the research question?				
5. Mixed methods	5.1. Is there an adequate rationale for using a mixed methods design to address the research question?				
	5.2. Are the different components of the study effectively integrated to answer the research question?				
	5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?				
	5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?				
	5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?				

Mixed Methods Appraisal Tool (MMAT) - Version 2018 User Guide

Qualitative studies

1.1. Is the qualitative approach appropriate to answer the research question?

Explanations
The qualitative approach used in a study (see non-exhaustive list on the left side of this table) should be appropriate for the research question and problem. For example, the use of a grounded theory approach should address the development of a theory and ethnography should study human cultures and societies.

This criterion was considered important to add in the MMAT since there is only one category of criteria for qualitative studies (compared to three for quantitative studies).

1.2. Are the qualitative data collection methods adequate to address the research question?

Explanations
This criterion is related to data collection method, including data sources (e.g., archives, documents), used to address the research question. To judge this criterion, consider whether the method of data collection (e.g., in depth interviews and/or group interviews, and/or observations) and the form of the data (e.g., tape recording, video material, diary, photo, and/or field notes) are adequate. Also, clear justifications are needed when data collection methods are modified during the study.

1.3. Are the findings adequately derived from the data?

Explanations
This criterion is related to the data analysis used. Several data analysis methods have been developed and their use depends on the research question and qualitative approach. For example, open, axial and selective coding is often associated with grounded theory, and within- and cross-case analysis is often seen in case study.

1.4. Is the interpretation of results sufficiently substantiated by data?

Explanations
The interpretation of results should be supported by the data collected. For example, the quotes provided to justify the themes should be adequate.

1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation?

Explanations
There should be clear links between data sources, collection, analysis and interpretation.

Quantitative non-randomized studies

3.1. Are the participants representative of the target population?

Explanations
Indicators of representativeness include: clear description of the target population and of the sample (inclusion and exclusion criteria), reasons why certain eligible individuals chose not to participate, and any attempts to achieve a sample of participants that represents the target population.

3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?

Explanations
Indicators of appropriate measurements include: the variables are clearly defined and accurately measured; the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure; validated and reliability tested measures of the intervention/exposure and outcome of interest are used, or variables are measured using 'gold standard'.

3.3. Are there complete outcome data?

Explanations
Almost all the participants contributed to almost all measures. There is no absolute and standard cut-off value for acceptable complete outcome data. Agree among your team what is considered complete outcome data in your field (and based on the targeted journal) and apply this uniformly across all the included studies. For example, in the literature, acceptable complete data value ranged from 80% (Thomas et al., 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder et al., 2003) and 30% for follow-up of more than one year (Viswanathan and Berkman, 2012).

3.4. Are the confounders accounted for in the design and analysis?

Explanations
Confounders are factors that predict both the outcome of interest and the intervention received/exposure at baseline. They can distort the interpretation of findings and need to be considered in the design and analysis of a non-randomized study. Confounding bias is low if there is no confounding expected, or appropriate methods to control for confounders are used (such as stratification, regression, matching, standardization, and inverse probability weighting).

3.5 During the study period, is the intervention administered (or exposure occurred) as intended?

Explanations
For intervention studies, consider whether the participants were treated in a way that is consistent with the planned intervention. Since the intervention is assigned by researchers, consider whether there was a presence of contamination (e.g., the control group may be indirectly exposed to the intervention) or whether unplanned co-interventions were present in one group (Sterne et al., 2016).

For observational studies, consider whether changes occurred in the exposure status among the participants. If yes, check if these changes are likely to influence the outcome of interest, were adjusted for, or whether unplanned co-exposures were present in one group (Morgan et al., 2017).

Mixed methods studies

5.1. Is there an adequate rationale for using a mixed methods design to address the research question?

Explanations
The reasons for conducting a mixed methods study should be clearly explained. Several reasons can be invoked such as to enhance or build upon qualitative findings with quantitative results

and vice versa; to provide a comprehensive and complete understanding of a phenomenon or to develop and test instruments (Bryman, 2006).

5.2. Are the different components of the study effectively integrated to answer the research question?

Explanations
Integration is a core component of mixed methods research and is defined as the "explicit interrelating of the quantitative and qualitative component in a mixed methods study" (Plano Clark and Ivankova, 2015, p. 40). Look for information on how qualitative and quantitative phases, results, and data were integrated (Pluye et al., 2018). For instance, how data gathered by both research methods was brought together to form a complete picture (e.g., joint displays) and when integration occurred (e.g., during the data collection-analysis or/and during the interpretation of qualitative and quantitative results).

5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?

Explanations
This criterion is related to meta-inference, which is defined as the overall interpretations derived from integrating qualitative and quantitative findings (Teddlie and Tashakkori, 2009). Meta-inference occurs during the interpretation of the findings from the integration of the qualitative and quantitative components, and shows the added value of conducting a mixed methods study rather than having two separate studies.

5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?

Explanations
When integrating the findings from the qualitative and quantitative components, divergences and inconsistencies (also called conflicts, contradictions, discordances, discrepancies, and dissonances) can be found. It is not sufficient to only report the divergences; they need to be explained. Different strategies to address the divergences have been suggested such as reconciliation, initiation, bracketing and exclusion (Pluye et al., 2009b). Rate this criterion 'Yes' if there is no divergence.

5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

Explanations
The quality of the qualitative and quantitative components should be individually appraised to ensure that no important threats to trustworthiness are present. To appraise 5.5, use criteria for the qualitative component (1.1 to 1.5), and the appropriate criteria for the quantitative component (2.1 to 2.5, or 3.1 to 3.5, or 4.1 to 4.5). The quality of both components should be high for the mixed methods study to be considered of good quality. The premise is that the overall quality of a mixed methods study cannot exceed the quality of its weakest component. For example, if the quantitative component is rated high quality and the qualitative component is rated low quality, the overall rating for this criterion will be of low quality

Appendix D. List of references of included studies

1. Amengual-Pizarro, M., & Prieto-Arranz, J. I. (2015). Exploring affective factors in L3 learning: CLIL vs. non-CLIL. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 197–220). Springer.
2. García Mayo, M. del P., & Villarreal Olaizola, I. (2011). The development of suppletive and affixal tense and agreement morphemes in the L3 English of Basque-Spanish bilinguals. *Second Language Research*, *27*(1), 129–149.
3. Grisaleña, J., Alonso, E., & Campo, A. (2009). Enseñanza plurilingüe en centros de educación secundaria: Análisis de resultados [Plurlingual education in secondary schools: Analysis of results]. *Revista Iberoamericana de Educación 49*(1), 1–12.
4. Gutiérrez Mangado, M., & Martínez-Adrián, M. (2018). CLIL at the linguistic interfaces. *Journal of Immersion and Content-based Language Education*, *6*(1), 85–112.
5. Herraiz Martinez, A. & Sánchez Hernández, A. (2019). Pragmatic markers produced by multilingual speakers: Evidence from a CLIL context. *English Language Teaching*, *12(*2), 68–76.
6. Lasagabaster, D. (2009). The implementation of CLIL and attitudes towards trilingualism. *International Journal of Applied Linguistics*, 157(1), 23–43.
7. Lázaro Ibarrola, A. (2011). Faster and further morphosyntactic development of CLIL vs. EFL Basque-Spanish bilinguals learning English in high-school. *International Journal of English Studies*, *12*(1), 79–96.
8. López-Deflory, E. & Juan-Garau, M. (2017). Going *glocal*: The impact of CLIL on English language learners' multilingual identities and attitudes in the Balearic Islands. *European Journal of Applied Linguistics*, *5*(1), 5–30.
9. Martínez Adrián, M. & Gutiérrez Mangado, J. (2015). Is CLIL instruction beneficial in terms of general proficiency and specific areas of grammar? *Journal of Immersion and Content-Based Language Education*, *3*(1), 51–76.
10. Merino, J. A. & Lasagabaster, D. (2018). CLIL as a way to multilingualism. *International Journal of Bilingual Education and Bilingualism*, *21*(1), 79–92.
11. Ní Dhiorbháin, A. (2020). Múineadh ábhar tríd an tríú teanga i mbunscoileanna Lán-Ghaeilge [Teaching a subject through a third language in Irish-medium primary schools]. *TEANGA, The Journal of the Irish Association for Applied Linguistics*, *27*, 1–21. DOI: 10.35903/teanga.v27i.484
12. Ollo Jiménez, P. & Martínez-Adrián, M. (2019). A study of self-reported opinions of L1-based communications strategies in CLIL and non-CLIL secondary-school learners of L3 English. *RAEL: Revista Electrónica de Lingüistica Aplicada*, *18*(1), 72–90.
13. Pérez-Vidal, C & Roquet, H. (2015). The linguistic impact of a CLIL *Science* programme: An analysis measuring relative gains. *System*, *54*, 80–90.
14. Prieto-Arranz, J. I., Rallo Fabra, L., Calafat-Ripoll, C., & Catrain-González, M. (2015). Testing progress on receptive skills in CLIL and non-CLIL contexts. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based*

language learning in multilingual educational environments (pp. 123–137). Springer.
15. Ruiz de Zarobe, Y. (2008). CLIL and foreign language learning: A longitudinal study in the Basque Country. *International CLIL Research Journal*, *1*(1), 60–73.
16. San Isidro, X. (2010). An insight into Galician CLIL: Provision and results. In D. Lasagabaster & Y. Ruiz de Zarobe (Eds.), *CLIL in Spain: Implementation, results and teacher training* (pp. 55–78). Cambridge Scholars.
17. San Isidro, X. & Lasagabaster, D. (2019). The impact of CLIL on pluriliteracy development and content learning in a rural multilingual setting: A longitudinal study. *Language Teaching Research*, *23*(5), 584–602.
18. San Isidro, X. & Lasagabaster, D. (2020). Students' and families' attitudes and motivations to language learning and CLIL: A longitudinal study. *The Language Learning Journal*, DOI: 10.1080/09571736.2020.1724185
19. Wang, L. & Kirkpatrick, A. (2013). Trilingual education in Hong Kong primary schools: a case study. *International Journal of Bilingual Education and Bilingualism*, *16(*1), 100–116.

Appendix E. Outcomes of included studies

Study	Strength	Sample size	Reported outcomes	Specific outcomes	Specific effect	Net effect
1. Amengual-Pizarro & Prieto-Arranz (2015)	Moderate	151	Attitudes	Views towards multilingualism	=	=
2. García Mayo & Villarreal Olaizola (2011)	Weak	78	L3 skills	Morphosyntax	=	=
3. Grisaleña et al. (2009)	Weak	?	Attitudes	Program satisfaction	+	+
		229	L3 skills	Four skills	+	
4. Gutiérrez Mangado & Martínez-Adrián (2018)	Weak	35	L3 skills	Morphosyntax	=	+
			L3 skills	Pragmatics	+	
5. Herraiz Martinez & Sánchez Hernández (2019)	Weak	19	L3 skills	Pragmatics	+	+
6. Lasagabaster (2009)	Moderate	277	Attitudes	Views towards multilingualism	+	+
7. Lázaro Ibarrola (2011)	Weak	26	L3 skills	Morphosyntax	+	+
8. López-Deflory & Juan-Garau (2017)	Strong	73	Attitudes	Views towards multilingualism	+	+
9. Martínez Adrián & Gutiérrez Mangado (2015)	Moderate	44	L3 skills	Morphosyntax	+/-	=
10. Merino & Lasagabaster (2018)	Moderate	285	L1 skills	Reading, writing	=	=

			L3 skills	Four skills	=	
11. Ní Dhiorbháin (2020)	Strong	6	Attitudes	Program satisfaction	+	+
12. Ollo Jiménez & Martínez-Adrián (2019)	Moderate	78	L3 skills	Global competence	+	+
				L1-based CSs	+	
13. Pérez-Vidal & Roquet (2015)	Weak	100	L3 skills	Reading, writing	+	+
				Listening	=	
14. Prieto-Arranz et al. (2015)	Moderate	87	L3 skills	Reading	+	+
				Listening	=	
15. Ruiz de Zarobe (2008)	Weak	21	L3 skills	Speaking	+	+
16. San Isidro (2010)	Weak	287	L3 skills	Four skills	+	+
17. San Isidro & Lasagabaster (2019)	Strong	44	Content	Content knowledge	+	+
			L1 skills	Four skills	+	
			L3 skills	Four skills	+	
18. San Isidro & Lasagabaster (2020)	Moderate	88	Attitudes	Views towards multilingualism	+	+
19. Wang & Kirkpatrick (2013)	Weak	144	Attitudes	Program satisfaction	+	+

Appendix F. Confounders by study

Study ID	Longitudinal?	Age	Pre-intervention score	L3 instructional hours in school	L3 instructional hours outside school	Home language(s)	Socio-economic status	CLIL subject allocation	Proportion controlled for	Final risk of bias rating
1	Y	Y	Y	N	N	Y	Y	N	4/7	Y
2	Y	Y	N	N	N	N	N	N	1/7	N
3	Y	Y	N	N	N	N	N	N	1/7	N
4	N	Y	N	Y	N	Y	Y	Y	5/7	?
5	N	Y	N	Y	N	N	N	Y	3/7	N
6	N	Y	N	N	N	N	N	N	1/7	N
7	Y	Y	N	N	N	N	N	N	1/7	N
8	N	Y	N	N	Y	Y	N	Y	4/7	?
9	N*	Y	N	Y	N	Y	Y	Y	5/7	Y
10	Y	Y	Y	N	N	N	N	N	2/7	?
12	N*	Y	N	Y	Y	N	N	Y	4/7	Y
13	Y	N	Y	Y	N	N	Y	Y	4/7	Y
14	Y	Y	N	N	Y	Y	Y	N	4/7	Y
15	Y	Y	N	N	Y	N	N	N	2/7	?
16	N	Y	N	N	N	N	N	N	1/7	N
17	Y	Y	Y	N	N	Y	N	Y	4/7	Y
18	Y	Y	Y	N	N	Y	N	Y	4/7	?
19	N	N	N	N	N	N	Y	Y	2/7	N
Y	10	16	5	5	4	7	6	9		
N	8	2	13	13	14	11	12	9		

Key: Y= yes, controlled for; N= no, not controlled for
*pseudo-longitudinal

Appendix G. Example completed data extraction form

	Item	Data
General	Date form completed	22/6/21
	Reference citation	1. Amengual-Pizarro, M., & Prieto-Arranz, J. I. (2015). Exploring affective factors in L3 learning: CLIL vs. non-CLIL. In M. Juan-Garau & J. Salazar-Noguera (Eds.), Content-based language learning in multilingual educational environments (pp. 197–220). Springer.
	Study author contact details	marian.amengual@uib.es
	Publication type	Book chapter
	Document source	Scopus
Study Overview	Research questions	1. Does learning context (CLIL vs. non-CLIL) play a role in the development of affective factors? 2. What is the impact of CLIL programmes on affective factors related to the content subject taught through English? 3. Does the participants' language profile have an effect on their interest in language learning? 4. Does learning context in combination with the learner's gender influence the development of affective factors?
	Study design	longitudinal case study
	Data type	quantitative non-randomized
	Study duration	2 school years
	Location and language of publication	Spain, English
Participants	School setting (social and educational context)	state (public) school
	Population description	five secondary schools in the Balearic Islands
	Method of recruitment into CLIL	general academic record, EFL grades, EFL placement test
	Languages spoken	Catalan
		Spanish
		English
	Age	13-14 -> 14-15

	Gender	F/M - 47/38 CLIL , 39/27 - NCLIL
	Other relevant sociodemographics	
Intervention	Languages of intruction (hours/week)	not reported
		not reported
		CLIL English 6 (3 EFL + 3 content)
	Subject allocation	not reported
	Length of exposure to CLIL pedagogy	1 year
	Non-CLIL control	yes, just EFL classes, age matched
	Number of participants	151
		85
		66
	Class grouping	participants chose from 3 different schools
	Baseline imbalances	not reported
	Attrition	170 --> 151, no reasons reported
Outcomes	Outcome type	attitude
	Outcome name	language attitudes
	Unit(s) of measure	likert scale 1 (totally agree) - 5 (totally disagree) low = positive attitude
	Time points measured	2

	Descriptive outcomes	1. "I am studying English bc it is a compulsory subject", T2 *(t*=-2.212, p<.05), CLIL more positive than NCLIL time- all students: 4 (I like English music and I want to understand it), 5 (I like watching English films and to be able to understand them), 10 (I wantt to travel abroad and English learning will help me), 11 (I want to speak English bc I want to communicate with people from diff countries) --- T2 more + > T1 ; 8 (I like the English lang but I do not like the English lessons) T2 more - > T1 (*p*<.05)
	Effect sizes	T2 (*t*=-2.212, *p*<.05) – *d*=0.37
Outcomes	Outcome type	attitude
	Outcome name	beliefs about learning English
	Unit(s) of measure	likert scale 1 (totally agree) - 5 (totally disagree) low = positive attitude
	Time points measured	2
	Descriptive outcomes	5. (I get nervous when I have to speak English) CLIL more + > NCLIL, *t*=2116; p<.05) time: 7 (I would like to get to know more English language speakers) *t*=-2.729, *p*<.05, T1-->T2 CLIL constant, NCLIL less positive than --> more positive than
	Effect sizes	*d*= -0.376644
Outcomes	Outcome type	attitude
	Outcome name	motivation towards learning English
	Unit(s) of measure	survey
	Time points measured	2

	Descriptive outcomes	key motivating factors (T1&2): marks, group work, activities, teaching method, amount of work aspect most like about Eng lang: what L2 enables me to do, how it sounds, TL people, TL culture, how it is written (no stat sig reported) reasons for learning Eng: gettting good job, being able to communicate with people around the world
	Effect sizes	N/A
	Summary	1. Mainly no stat sig differences concerning affective factors, and both groups became more positive over time (only diff - I am studying english bc compulsory-- CLIL more positive), lower anxiety in CLIL 2. no sig difference among CLIL students regarding views towards CLIL subjects, highly motivated but no difference over time 3. CLIL/NCLIL class make up very different regarding parent L1 (CLIL more catalan, nclil more spanish=ethnically diverse)

Appendix H. Example completed quality evaluation form (MMAT)

Paper: Amengual-Pizarro & Prieto-Arranz (2015)

	Methodological quality criteria	**Responses**			
		Yes	No	Can't tell	Comments
Screening questions	S1. Are there clear research questions?	x			
	S2. Do the collected data allow to address the research questions?	x			
	Further appraisal may not be feasible or appropriate when the answer is 'No' or 'Can't tell' to one or both screening questions.				
1. Qualitative	1.1. Is the qualitative approach appropriate to answer the research question?				
	1.2. Are the qualitative data collection methods adequate to address the research question?				
	1.3. Are the findings adequately derived from the data?				
	1.4. Is the interpretation of results sufficiently substantiated by data?				
	1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation?				
2. Quantitative randomized controlled trials	2.1. Is randomization appropriately performed?				
	2.2. Are the groups comparable at baseline?				
	2.3. Are there complete outcome data?				
	2.4. Are outcome assessors blinded to the intervention provided?				
	2.5 Did the participants adhere to the assigned intervention?				

3. Quantitative non-randomized	3.1. Are the participants representative of the target population?	x			*several schools, relatively even gender distribution, public schools, large sample*
	3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?	x			*used in other testing, Cronbach's alpha measuring internal consistency between .74 and .93*
	3.3. Are there complete outcome data?			x	*attrition reported in number but no reason, but fairly low proportion overall*
	3.4. Are the confounders accounted for in the design and analysis?	x			*longitudinal design to determine change over time and account for the initial differences*
	3.5. During the study period, is the intervention administered (or exposure occurred) as intended?			x	*low risk of contamination, though intervention at different schools, unclear what subjects were taught*
4. Quantitative descriptive	4.1. Is the sampling strategy relevant to address the research question?				
	4.2. Is the sample representative of the target population?				
	4.3. Are the measurements appropriate?				
	4.4. Is the risk of nonresponse bias low?				
	4.5. Is the statistical analysis appropriate to answer the research question?				
5. Mixed methods	5.1. Is there an adequate rationale for using a mixed methods design to address the research question?				
	5.2. Are the different components of the study effectively integrated to answer the research question?				
	5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?				
	5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?				
	5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?				