Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 13 (2). 2023. 317-345
https://doi.org/10.14746/ssllt.38277
http://pressto.amu.edu.pl/index.php/ssllt

Effects of EMI-CLIL on secondary-level students' English learning: A multilevel meta-analysis

Jang Ho Lee
Chung-Ang University, Seoul, Republic of Korea
https://orcid.org/0000-0003-2767-3881
jangholee@cau.ac.kr

Hansol Lee 🖂
Korea Military Academy, Seoul, Republic of Korea
https://orcid.org/0000-0002-6912-7128
hansol@kma.ac.kr

Yuen Yi Lo
The University of Hong Kong, China
https://orcid.org/0000-0002-0850-5447
yuenyilo@hku.hk

Abstract

This meta-analysis synthesized the effects of the English medium instruction and content and language integrated learning (EMI-CLIL) approach on secondary-level students' English learning. The dataset included 44 samples (*N* = 7,434) from 38 primary studies. The results revealed EMI-CLIL's overall effectiveness for the development of English competence compared to the mainstream condition in the short term (*d* = 0.73, *SE* = 0.06, 95% CI [0.61, 0.86]) and longer term (*d* = 1.01, *SE* = 0.06, 95% CI [0.88, 1.15]). Additionally, we found that EMI-CLIL's overall effectiveness was influenced by several moderator variables. Its effectiveness was significantly: (1) higher for learners whose first language (L1) was linguistically related to English; (2) lower for primary studies which confirmed the homogeneity of the EMI-CLIL and comparison groups; (3) lower when studies targeted the productive (rather than receptive or overall) dimension of English learning; and (4) higher when outcome measures focused on vocabulary. Implications for pedagogy and future research are discussed.

Keywords: content and language integrated learning; English as a foreign language; English medium instruction; multilevel meta-analysis; secondary level

1. Introduction

This meta-analysis synthesized the effects of using English as the medium of instruction in content subjects on students' English development. English medium instruction (EMI) refers to a pedagogical approach to teaching content subjects (e.g., mathematics, science, history, geography; other than English) in English in contexts where English is not the majority language (Macaro, 2018). EMI can be traced back to the use of English as the medium of instruction at pre-tertiary levels in some post-colonial contexts in which English, as a former colonial language and current global lingua franca, is maintained as the medium of instruction for its social status and economic value (Evans, 2017). EMI has also been implemented in higher education institutions in non-Anglophone contexts in attempts to attract more international students and/or enhance local students' English proficiency, hence their competitiveness in the globalized world (Rose et al., 2021). The underlying assumption of using EMI is that, by providing language exposure, it would facilitate students' English learning while teaching content subjects. However, the evidence of such "two-for-one" benefits for English development remains inconclusive (see the summaries of recent systematic reviews by Goris et al., 2019 and Graham et al., 2018). Hence, a more rigorous statistical analysis of previous studies' findings is needed. The present study responds to this need through a meta-analysis of relevant primary studies.

In addition to EMI studies, we included research studies under the label *content and language integrated learning* (CLIL), a similar pedagogical approach associated to a greater extent with Europe. An early definition of CLIL by Marsh (2002) states that it is "a generic umbrella term which would encompass any activity in which a foreign language is used as a tool in the learning of a non-language subject in which both language and the subject have a joint particular role" (p. 58). The first part of this definition, that is, the use of a foreign language (mostly, English) as a tool and the involvement of non-language subjects, closely resembles that of EMI, whereas the second part highlights the "joint role" of both language and content learning. Although such a joint role has been reinforced in other CLIL definitions (e.g., Coyle et al., 2010; Morton & Llinares, 2017), whether and how the integration of content and language learning can be achieved in practice remains unclear (Dalton-Puffer, 2013). As the target language (e.g., English) is often not widely used beyond formal instruction contexts, CLIL students also learn it as an additional language in designated language lessons (Dalton-Puffer et al., 2014). However, the same practice is also common in EMI programs, particularly those implemented at the secondary level.

English learning is a desired outcome of both EMI and CLIL. The reviewed literature did not reveal any substantial difference between these approaches in terms of their practice (i.e., students learning content knowledge through English while learning English as an additional language in parallel) or teacher and student profiles (i.e., most teachers are non-English speakers trained to be content specialists; students share their first language, or L1, as the majority language). Accordingly, we believe that EMI and CLIL share a "functionally equivalent" context (Rose et al., 2021, p. 1) and can be included together in the current meta-analysis.

2. Literature review

2.1. EMI-CLIL's effectiveness for English learning

The popularity of the EMI-CLIL approach partly derives from the search for more effective or innovative foreign language teaching approaches (Pérez-Cañado, 2016). With the paradigm shift to communicative language teaching (CLT; Nunan, 2011), which emphasizes meaningful language use resembling students' language use outside the classroom, the potential of teaching content subjects through English for facilitating English learning has been recognized. In addition to timetabled English language lessons, adopting the EMI-CLIL approach in some or all content subjects increases exposure to English input and English-use opportunities. The nature of the communication is also assumed to differ from that in English language lessons; discussions based on content knowledge (e.g., historical events, science experiments, social issues) may provide more meaningful contexts for purposeful communication (Lyster & Ruiz de Zarobe, 2018). Therefore, in view of second language acquisition (SLA) theories, such as the input (Krashen, 1982), interaction (Long, 1996), and output (Swain, 1995) hypotheses, the EMI-CLIL approach appears to provide favorable conditions for English learning. Learning content subjects in English may also increase students' English-learning motivation, which is indispensable for academic success (Genesee & Lindholm-Leary, 2013).

Lo and Lo's (2014) meta-analysis of EMI secondary education in Hong Kong confirmed the aforementioned benefits. They compared academic achievement, first language and second language (L2 English) development, and affective variables between students studying with EMI and Chinese medium instruction (CMI). Consolidating the results of 10 studies, they revealed that EMI students outperformed their CMI peers in English proficiency with a moderate effect size. The meta-analysis also revealed some potential moderators that affected the intergroup comparison. Of particular interest to the current study are their students' initial abilities and type of outcome measures. It was found that when students' initial abilities were not controlled, EMI students seemed to outperform CMI students in content subjects, yet the opposite was true when students' initial abilities were considered. Regarding outcome measures, it was found that the EMI group performed better than their CMI counterpart in studies employing standardized measurements (e.g., high-stakes examinations developed by the authorities), but not for self-designed tests. This moderating impact of outcome measure type may have come from the fact that self-designed tests were generally more geared toward measuring certain knowledge addressed during the research period, which in turn may have favored CMI condition associated with instruction in the L1.

2.2. Skepticism about EMI-CLIL's effectiveness for English learning

Despite the aforementioned theoretical support and Lo and Lo's (2014) meta-analysis, the EMI-CLIL approach is not universally praised (Bruton, 2013); the criticisms are summarized as follows: first, some studies have demonstrated that the "assumed" favorable conditions for English learning in EMI-CLIL may be absent in some classrooms. For instance, Lo and Macaro (2012) showed that the quality and quantity of teacher-student interactions were rather limited in EMI lessons, leaving students few opportunities to negotiate or interact with teachers and peers in English. An et al. (2019) and Hu and Gao (2021) revealed a lack of language scaffolding (operationalized as language-oriented or language-related practices) in EMI lessons in mainland China and Hong Kong respectively, implying that comprehensible input or content and language integrated teaching may not be available in practice.

Second, some researchers have identified methodological flaws in relevant studies, including heterogeneity between experimental and comparison groups, lack of robust statistical analyses of significant intergroup differences, and lack of control over confounding variables such as exposure to English outside classrooms (Pérez-Cañado, 2012). In particular, EMI-CLIL research has largely been criticized for selection bias, such that EMI-CLIL groups consist of students with higher socio-economic status, better academic ability, and stronger motivation (e.g., Broca, 2016; Bruton, 2013). Hence, EMI-CLIL groups and their comparison counterparts (i.e., mainstream EFL groups) may not be homogeneous.

Some reviews have also cast doubt on EMI-CLIL's effectiveness for English learning. For example, Graham et al.'s (2018) systematic review included 25 EMI-

CLIL studies published between 2008 and 2018 that had examined either language or content learning outcomes. Regarding the former, the authors reported that previous studies revealed mixed findings, with some studies showing the superiority of the EMI-CLIL condition over the comparison condition and others presenting no significant difference. They concluded that extant literature does not provide convincing evidence regarding EMI-CLIL's effectiveness for language learning. In an in-depth review of 21 European studies, Goris et al. (2019) focused on longitudinal studies examining the effects of CLIL on English skills and knowledge. They revealed that a considerable proportion of longitudinal studies reported null effects. They further noted that longitudinal research on this issue has only started to flourish recently and called for more longitudinal CLIL studies. Finally, in their report for the Education Endowment Foundation, Murphy et al. (2020) concluded that EMI-CLIL programs may be more effective than mainstream ones (but mainly for vocabulary knowledge and receptive skills). They qualified their conclusion by suggesting that the observed superiority of EMI-CLIL may have arisen from the combined effects of such instruction and other confounding variables (e.g., additional exposure to the target language input for the EMI-CLIL group).

Overall, the aforementioned skepticism and theoretical support for EMI-CLIL's effectiveness demand a more systematic approach to synthesizing the results of relevant primary studies; a range of moderators that may play important roles in EMI-CLIL's effectiveness must also be identified.

2.3. Potential moderator variables

Before delving into the full meta-analysis, we first review potential moderator variables influencing EMI-CLIL's effectiveness. Based on the above-mentioned literature, we identified six noteworthy moderators: (1) L1-English relation, (2) intensity of EMI-CLIL program, (3) homogeneity confirmation, (4) target linguistic dimension, (5) vocabulary targeted, and (6) language test type.

2.3.1. L1-English relation

Most EMI-CLIL studies have been conducted in European countries (e.g., the Netherlands, Germany, Sweden, Belgium) and in the Asia-Pacific region (e.g., Hong Kong and South Korea). Given such diverse contexts and learners' L1s, the relationship between learners' L1 and the target language (i.e., English) has been hypothesized to influence EMI-CLIL's effectiveness. Jeon and Yamashita (2014) adopted a similar rationale and approach to examining the role of L1-L2 distance in mediating the relationships between L2 reading comprehension and other components. Furthermore, Lo and Lo (2014) speculated that the typological difference between Chinese and English may explain the diverse results of studies conducted in Hong Kong and other educational contexts.

2.3.2. EMI-CLIL program intensity

Most EMI-CLIL programs in relevant studies (e.g., Dallinger et al., 2016; Gierlinger & Wagner, 2016; Pérez Cañado & Lancaster, 2017) have taught one to three content subjects through English. However, in a small number of EMI-CLIL contexts, more than 50% of weekly instructional hours were implemented in English (some EMI programs in Hong Kong: Lin & Morrison, 2010; Lo & Murphy, 2010; Salili & Lai, 2003; CLIL in the Netherlands: Goris et al., 2013; Verspoor et al., 2015; an international South Korean high school: Lee, 2020). Given this variation in exposure to English inputs as well as a recent finding in a Spanish CLIL project on the superiority of a more intensive CLIL course compared to a less intensive one (Merino & Lasagabaster, 2018), it was hypothesized that EMI-CLIL program intensity may be associated with the development of English competence (see Murphy et al., 2020, for a similar discussion).

2.3.3. Homogeneity confirmation

As mentioned above, EMI-CLIL research has been severely criticized for unfairly comparing mainstream (i.e., traditional EFL instruction) and EMI-CLIL groups, owing to baseline intergroup differences (e.g., Bruton, 2013; Goris et al., 2019; Graham et al., 2018; Macaro, 2018). Such a recurrent methodological flaw undermines any conclusive statement about EMI-CLIL's effectiveness for English learning. While the initial advantage of EMI-CLIL groups persists in much EMI-CLIL research, some recent studies have started to resolve this issue through diverse approaches (see *Coding Scheme*, section 4.5., for examples). However, whether the homogeneity of the mainstream and EMI-CLIL groups is associated with the (reported) degree of developments of English competence through the EMI-CLIL approach remains unconfirmed. Accordingly, we included homogeneity confirmation as a moderator variable, in line with Lo and Lo's (2014) meta-analysis.

2.3.4. Target linguistic dimensions

Dalton-Puffer's (2008) much-cited summary of CLIL outcomes in Europe pointed to the differential effects of CLIL intervention on a range of target linguistic aspects. She suggested that the CLIL approach may be more beneficial for receptive skills (i.e., reading and listening), vocabulary, and fluency as part of oral competence than writing, syntactic knowledge, and pronunciation. In a similar effort to summarize the findings of previous EMI-CLIL studies on language learning outcomes, Graham et al. (2018) grouped a range of examined target linguistic aspects into overall English proficiency, receptive skills, and productive skills. Their analysis revealed that, in general, studies in each category showed mixed findings regarding EMI-CLIL's effectiveness. Like Dalton-Puffer's (2008) and Graham et al.'s (2018) reviews, which are geared towards providing pedagogical implications for EMI-CLIL practitioners, the present meta-analysis also aims to offer evidence-based suggestions regarding the differential effects of EMI-CLIL on different linguistic aspects but based on a rigorous statistical approach.

2.3.5. Vocabulary targeted

It has been suggested that the EMI-CLIL approach exposes learners to a wider range of English vocabulary than its mainstream counterpart (Dalton-Puffer, 2007; Macaro, 2018); vocabulary is generally the sole aspect of English knowledge dealt with explicitly in EMI-CLIL lessons (An et al., 2019). Accordingly, much EMI-CLIL research has examined English vocabulary as the target linguistic knowledge via diverse lexical measurements (e.g., Canga Alonso & Arribas García, 2015; Gierlinger & Wagner, 2016; Goris et al., 2013; Hendrikx & Van Goethem, 2020; Lo & Murphy, 2010; Martínez Agudo, 2020; Olsson, 2015). Regarding the findings of studies on vocabulary, Dalton-Puffer (2011) summarized that "they concur that CLIL students' receptive and productive lexicon is larger overall, contains more words from lower frequency bands, has a wider stylistic range, and is used more appropriately" (p. 186). Given vocabulary's status in EMI-CLIL research (Ruiz de Zarobe, 2011), we included the moderator variable "vocabulary targeted;" this assessed EMI-CLIL's effects on the development of vocabulary knowledge.

2.3.6. Language test type

EMI-CLIL research has employed different types of language tests to measure language learning outcomes. Some studies have adopted validated tests like the standardized Key English Test (Cambridge ESOL, 2008), employed by Merino and Lasagabaster (2018), and the Vocabulary Levels Test developed by Schmitt et al. (2001), which has been widely employed in EMI-CLIL research targeting vocabulary developments (e.g., Bayram et al., 2019; Castellano-Risco et al., 2020). Other studies have employed self-designed tests, conveniently and purposefully developed by the researchers, sometimes in collaboration with the teachers in the target context. As aforementioned, Lo and Lo's (2014) meta-analysis found "type of outcome measures" to be a significant moderator. Hence, we included language test type as a moderator.

3. The present study

The number of recently published review articles on EMI-CLIL's effects on English competence (e.g., Goris et al., 2019; Graham et al., 2018; Murphy et al., 2020) attests to the flourishing interest in this area. While these reviews have their own objectives and are of significant value for summarizing primary studies' findings, they have some limitations owing to their descriptive and qualitative nature. Another limitation is their rather subjective interpretation of the roles of potential moderators, which can be more systematically controlled in a meta-analysis. Accordingly, we seek to address these limitations and evaluate EMI-CLIL's effectiveness with a rigorous, systematic, methodological meta-analysis.

This meta-analysis focuses on secondary-level learners for the following reasons. First, theoretically, the language development trajectory of students at different key stages of education (particularly between primary and secondary/tertiary levels) is deemed to vary (see Johnson & Swain, 1994, for discussion about the language and conceptual development of learners who start acquiring content knowledge through an L2 at different ages). Therefore, focusing on learners at a particular key stage eliminates one potential confounding variable. Second, secondary level was found to be the most widely studied education level in EMI-CLIL research (Macaro, 2018), offering a sufficient number of samples and effect sizes based on which EMI-CLIL's effectiveness could be calculated meta-analytically. Our research questions (RQs) are:

1. To what extent does the EMI-CLIL approach lead to higher levels of English competence for secondary-level learners than its mainstream counterpart?
2. To what extent are the identified moderators related to EMI-CLIL's effect on the development of secondary-level learners' English competence?

4. Method

4.1. Literature search and inclusion criteria

We began by conducting a literature search and identified studies for review based on the following inclusion criteria: they were required to (1) be written in English and published between 2001 and 2021; (2) be conducted in the English as a foreign language (EFL) context (including Hong Kong, for which several EMI studies have been conducted); (3) target secondary education students; (4) be geared toward teaching content subjects (other than English) in English; (5) measure outcomes related to English proficiency or linguistic knowledge; (6) include EMI-CLIL and comparison groups; and (7) report descriptive statistics to enable effect size calculation. Figure 1 illustrates our literature search steps according to the PRISMA flow diagram.

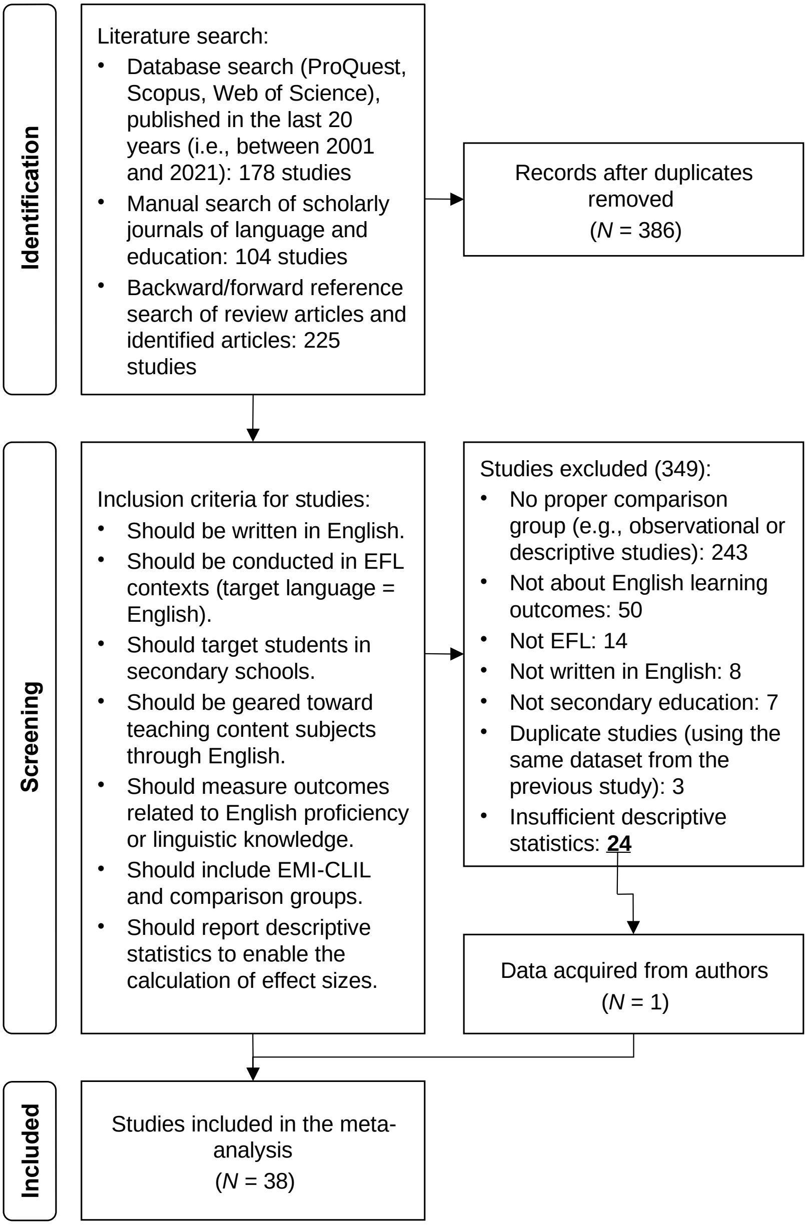

Identification
Literature search:
• Database search (ProQuest, Scopus, Web of Science), published in the last 20 years (i.e., between 2001 and 2021): 178 studies
• Manual search of scholarly journals of language and education: 104 studies
• Backward/forward reference search of review articles and identified articles: 225 studies
Records after duplicates removed
(N = 386)
Screening
Inclusion criteria for studies:
• Should be written in English.
• Should be conducted in EFL contexts (target language = English).
• Should target students in secondary schools.
• Should be geared toward teaching content subjects through English.
• Should measure outcomes related to English proficiency or linguistic knowledge.
• Should include EMI-CLIL and comparison groups.
• Should report descriptive statistics to enable the calculation of effect sizes.
Studies excluded (349):
• No proper comparison group (e.g., observational or descriptive studies): 243
• Not about English learning outcomes: 50
• Not EFL: 14
• Not written in English: 8
• Not secondary education: 7
• Duplicate studies (using the same dataset from the previous study): 3
• Insufficient descriptive statistics: 24
Data acquired from authors
(N = 1)
Included
Studies included in the meta-analysis
(N = 38)

Figure 1 PRISMA flow diagram (Page et al., 2021)

First, we conducted keyword searches in databases (ProQuest, Scopus, and Web of Science) with the following keywords: [EMI OR "English medium instruction" OR CLIL OR "content and language integrated learning"] AND [outcome OR development OR effect OR non-EMI OR non-CLIL OR mainstream]. This keyword combination was derived from a series of discussions among the authors aiming to identify primary studies relevant to our seven criteria. For this step, we identified 178 studies. Second, we manually searched the relevant journals (e.g., *International CLIL Research Journal, International Journal of English Studies, Journal of Immersion and Content-Based Language Education, The Language Learning Journal, System*), and identified 104 studies. Lastly, we checked the reference lists of the identified primary studies and previous systematic reviews and a meta-analysis (Goris et al., 2019; Graham et al., 2018; Lo & Lo, 2014) to conduct forward and backward searches, and identified further 225 studies. Overall, after removing duplicates, we identified 386 studies.

We then reviewed these 386 studies to confirm if they successfully met our seven criteria. We had to exclude 349 studies for the following reasons: (1) 243 did not have proper comparison groups (e.g., observational or descriptive studies); (2) 50 were not about English learning outcomes; (3) 14 were not EFL studies; (4) 8 were not written in English; (5) 7 were not conducted in secondary education contexts; (6) 3 were duplicate studies (using the same dataset as a previous study); and (7) 24 did not report sufficient descriptive statistics for effect size calculations. For the last criterion, we contacted the authors, and one study supplied the necessary data (Rumlich, 2017). Finally, 38 studies remained for our meta-analysis.

4.2. Dataset construction

Subsequently, we began dataset construction by examining 38 primary studies. First, we found that seven studies included multiple independent samples; we decided to treat them as separate studies. For example, Goris et al. (2013) had participants from three countries (the Netherlands, Germany, and Italy), which we treated independently. Similarly, Merino and Lasagabaster (2018) had participants from two distant communities in their country. Hamidavi et al. (2016), Lahuerta (2020), Lo and Murphy (2010), Salili and Lai (2003), and Verspoor et al. (2015) investigated students in two different grades, genders, or cohorts simultaneously, giving two independent samples in their studies. As a result, we could identify 8 additional samples, meaning that there were 46 samples (38 + 8) among 38 primary studies. Conversely, we found that two studies (Martínez Agudo, 2019, 2020) analyzed the same sample and that another two studies (Lahuerta, 2017, 2020) came from the same sample; thus, we counted these four studies as two samples. In summary, after considering each study's data structure, we identified 44 (38 + 8 - 2) samples (N = 7,434) from 38 primary studies.

4.3. Effect size calculation

To quantitatively synthesize previous findings on EMI-CLIL's effects on English competence, we calculated treatment effect sizes – differences in learning outcomes between the EMI-CLIL condition and its mainstream counterpart – across the collected studies. We chose to use unbiased Cohen's *d*, also known as Hedges' *g*, as our effect size unit. As this type of effect size can be calculated by – roughly speaking – dividing mean differences (measured by a certain scale) by a pooled standard deviation (measured by the said scale), the computed effect sizes are scale-free. As for interpreting effect sizes, we endorsed Plonsky and Oswald's (2014) "t-shirt size" benchmarks, where 0.4, 0.7, and 1.0 are considered small, medium, and large effect size guidelines, respectively. The equations for computing effect sizes in a unit of unbiased Cohen's *d*, which involves calculating a Cohen's *d* effect size and a correction factor (J), can be found in Hedges (1981).

When computing effect sizes for each independent sample, we found that some utilized multiple numbers of measurements. To keep these effect sizes in the dataset and avoid any statistical dependence issue (i.e., when effect sizes from the same sample are mutually dependent), we used a multilevel meta-analysis approach, which is described in the data analysis plan subsection. We computed a total of 192 posttest and 9 delayed posttest effect sizes. The delayed posttest effect sizes related to four studies (Lin & Morrison, 2010; Martínez Agudo, 2019, 2020; Pérez Cañado & Lancaster, 2017), which measured their participants' long-term learning outcomes.

4.4. Outlier diagnostics and publication bias[1]

To systematically detect outliers and influential cases among the calculated posttest effect sizes ($N = 192$; $k = 43$), we first standardized our effect sizes into *z*-scores and excluded one sample (Gutiérrez Ortiz, 2020) with absolute effect size values greater than 3.29 (for details of this approach, see Assink & Wibbelink, 2016; Tabachnik & Fidell, 2013). Then, we used Viechtbauer's (2010) *metafor* package (version 2.4-0) in *R* software (version 4.0.3), following Lee and Lee's (2022) approach. As shown in Figure 2, this package provides a filled circle to indicate an outlier based on studentized deleted residuals, which can be calculated by dividing the residuals of effect sizes by their standard errors, along with other mathematical measures, such as DFFITS values, Cook's distances, and COVRATIO values (for definitions and equations of each measure see Viechtbauer & Cheung, 2010). Accordingly, we excluded another sample (Goris et al., 2013[3]) from the dataset.

[1] Due to its small sample size ($N = 9$), we did not follow this approach for our delayed posttest effect sizes.

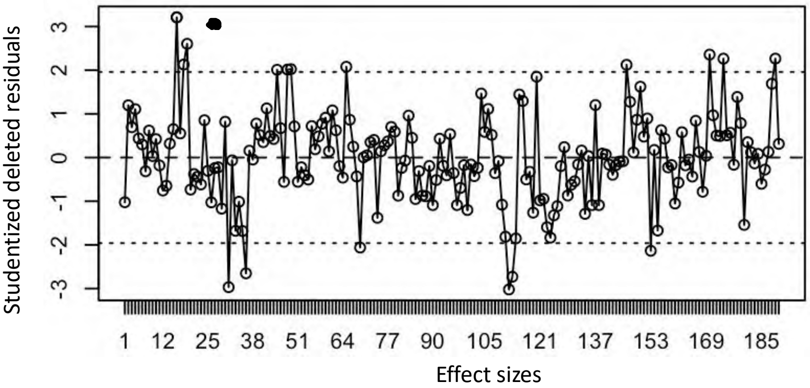


Figure 2 Plot of the studentized deleted residuals for 190 posttest effect sizes. (Filled circles are suggested outliers)

Subsequently, we conducted Egger's bias regression test and computed a funnel plot (see Figure 3) to check if the revised dataset still indicated small-study effects; there was no statistically significant sign of bias (z = 1.54, p = .122). Additionally, the funnel plot showed an overall symmetrical pattern of the effect sizes based on the computed funnel-shaped diagram (though many effect sizes were located outside it). Finally, in the revised dataset we had a total of 184 posttest effect sizes from 41 samples (N = 6,654).

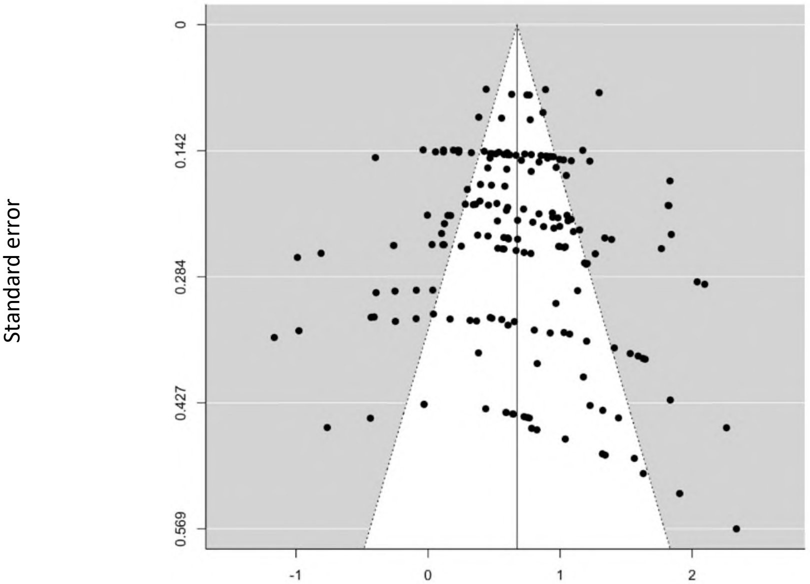


Figure 3 Funnel plot for 184 posttest effect sizes

4.5. Coding scheme

4.5.1. L1-English relation

For this moderator variable, we divided the studies into L1-L2 (English) related and not related using Beaufils and Tomin's (2020) genetic proximity calculator, which provides language relatedness scores (0~100). Scores larger (smaller) than 78 indicate unrelated (related) languages. For 136 effect sizes out of 184 (74%), students' L1s were related to English; all were from European countries (Austria, the Netherlands, Germany, Italy, Spain, Sweden, Belgium). The remaining 48 effect sizes (26%) came from studies conducted in Turkey, South Korea, and Hong Kong.

4.5.2. EMI-CLIL program intensity

For this moderator variable, we categorized the selected studies into more intensive (intensity+; 56 of 184; 30%) and less intensive (intensity–; 126 of 184; 68%) EMI-CLIL programs when they provided more or less than 50% of weekly instructional hours in English, respectively. The former included CLIL programs in the Netherlands (Goris et al., 2013; Verspoor et al., 2015), and an international high school in South Korea (Lee, 2020), and EMI programs in Hong Kong (Lin & Morrison, 2010; Lo & Murphy, 2010; Salili & Lai, 2003). The latter included most EMI-CLIL programs in which one to three content subjects were taught in English. Two samples (Hamidavi et al., 2016[1],[2]) did not report relevant information.

4.5.3. Homogeneity confirmation

Studies were coded as "confirmed" (36 of 184; 20%) or "not confirmed" (148 of 184; 80%), depending on whether they confirmed the homogeneity of the experimental (i.e., EMI-CLIL) and comparison (i.e., mainstream) groups before measuring English learning outcomes. We only coded studies as "confirmed" when they: (1) checked the baseline differences between the groups through a pretest[2] related to the target linguistic aspects or general English proficiency (e.g., Pérez Cañado & Lancaster, 2017; Prieto-Arranz et al., 2015; Rallo Fabra & Jacob, 2015) and reported no statistically significant difference; (2) employed some statistical adjustments such as propensity score matching (Feddermann et al., 2021) and elimination of outliers (Pérez Cañado & Lancaster, 2017; Prieto-Arranz et al., 2015); or (3) adopted a probability sampling method (Martínez

[2] We conducted an independent *t*-test based on the descriptive statistics related to the pretest when the author(s) of the selected studies did not explicitly mention intergroup homogeneity.

Agudo, 2019). Studies were coded as “not confirmed,” when they failed to test significant intergroup differences via a pretest or when they checked baseline differences through measurements (e.g., English learning motivation) other than a pretest of the target linguistic aspects and English proficiency.

4.5.4. Target linguistic dimensions

In view of Graham et al.’s (2018) and Goris et al.’s (2019) systematic reviews as well as the number of effect sizes in our dataset, we grouped the target linguistic aspects into “receptive,” “productive,” and “overall proficiency.” “Receptive” (50 of 184; 27%) included listening (e.g., Dallinger et al., 2016; Lasagabaster, 2008) and reading skills (e.g., Bayram et al., 2019; Martínez Agudo, 2020), and linguistic knowledge (i.e., vocabulary and grammar) measured receptivity (Lasagabaster, 2008; Martínez Agudo, 2020). For “productive” (119 of 184; 65%), any measurements related to speaking and writing (e.g., Lahuerta, 2017; Pérez Cañado & Lancaster, 2017; Rallo Fabra & Juan-Garau, 2011) were included, along with linguistic knowledge measured productively (e.g., Gutiérrez-Mangado & Martínez-Adrián, 2018; Lo & Murphy, 2010). “Overall proficiency” (15 of 184; 8%) included any measurements described as measuring participants’ overall proficiency (e.g., Merino & Lasagabaster, 2018; Verspoor et al., 2015).

4.5.5. Vocabulary targeted

Studies were coded as “vocabulary” for this moderator variable if they measured participants’ overall English vocabulary knowledge (e.g., Castellano-Risco et al., 2020; Gierlinger & Wagner, 2016) or specific aspects of productive vocabulary competence through speaking or writing tasks, with, e.g., lexical complexity (Lee, 2020), lexical diversity (Van Mensel et al., 2020), and lexical error ratio (Lahuerta, 2020). Studies that measured participants’ knowledge about English idioms (Goris et al., 2013; Hendrikx & Van Goethem, 2020) were also categorized as vocabulary-related (51 of 184; 28%). The rest were coded as “others” (133 of 184; 72%).

4.5.6. Language test type

Language test types were coded as “validated” or “self-designed,” depending on their characteristics and descriptions. As mentioned above, tests developed by national agencies, professional assessment organizations, or applied linguistics specialists, which have been validated through several studies, were coded as “validated” (100 of 184; 54%). The rest, which were generally developed by study authors without reference to other literature, were coded as “self-designed” (84 of 184; 46%).

4.6. Data analysis plan

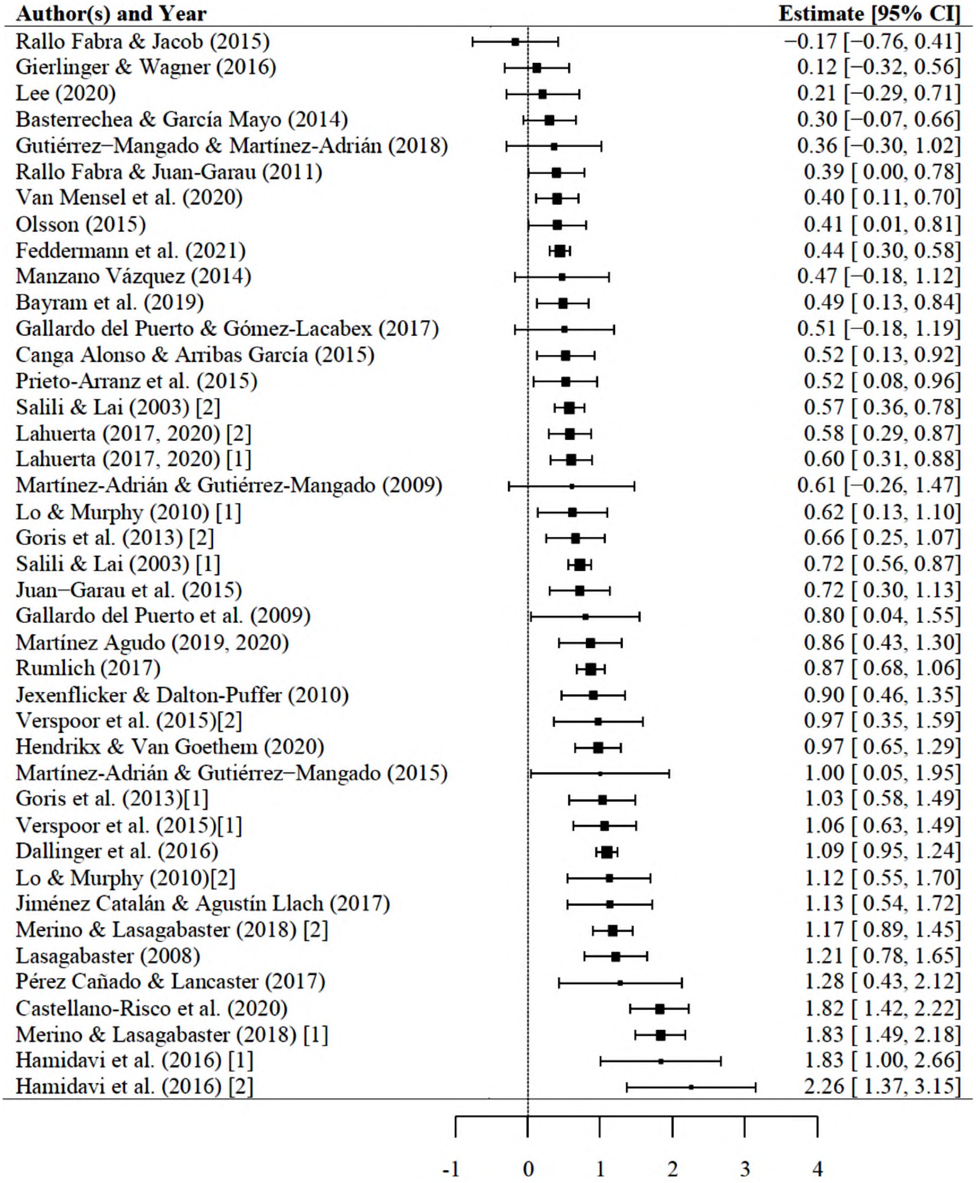
We used the *metafor* (version 2.4-0) package (Viechtbauer, 2010) in *R* (version 4.0.3) as a meta-analysis tool. As aforementioned, we computed effect sizes in accordance with different measurements, and samples could have more than one effect size. When computing an average effect size based on these effect sizes, their data structure membership should be considered in addition to their sampling errors (i.e., standard errors of effect sizes). To this end, the *metafor* package provides a multilevel approach using a random-effects model (Fernández-Castilla et al., 2020) to include intra-sample effect size variance taking the multilevel data structure into account. The same approach was used for moderator analyses to provide accurate statistical estimates. Specifically, we first employed simple meta-regression for each moderator variable before conducting a multiple meta-regression to check if previous results were changed after controlling for other variables for higher precision (see Lee et al., 2019 for a similar approach).

5. Results

5.1. EMI-CLIL's overall effectiveness for English competence development

Figure 4 represents a forest plot for posttest effect sizes for 41 samples. Due to space limitations, we could not plot all 184 effect sizes; Figure 4's estimates should be read only for understanding a general trend of the computed effect sizes across the included primary studies. Overall, although about eight studies seemed to include zeros in the 95% confidence intervals of their effect sizes (indicating comparable learning outcomes between EMI-CLIL and mainstream conditions), we found that most studies had positive effect sizes, indicating that the EMI-CLIL condition led to higher levels of English competence for secondary-level learners than its mainstream counterpart.

Subsequently, the multilevel meta-analysis of 184 posttest effect sizes calculated from 41 independent samples (N = 6,654) revealed that EMI-CLIL's overall effectiveness compared to the mainstream condition was 0.73 (SE = 0.06, z = 11.66, p < .001, 95% CI [0.61, 0.86]). Further, we found that the delayed posttest results based on a total of eight effect sizes from three independent samples (N = 676) indicated that the overall mean effect size was 1.01 (SE = 0.06, z = 17.55, p < .001, 95% CI [0.88, 1.15]). It should be noted that mean effect sizes for the short- and longer-term interventions should not be statistically compared as they were based on different datasets; thus, interpretations about the relative importance of these two findings (e.g., the effectiveness was greater longer term) should be avoided.



Author(s) and Year	Estimate [95% CI]
Rallo Fabra & Jacob (2015)	−0.17 [−0.76, 0.41]
Gierlinger & Wagner (2016)	0.12 [−0.32, 0.56]
Lee (2020)	0.21 [−0.29, 0.71]
Basterrechea & García Mayo (2014)	0.30 [−0.07, 0.66]
Gutiérrez−Mangado & Martínez-Adrián (2018)	0.36 [−0.30, 1.02]
Rallo Fabra & Juan-Garau (2011)	0.39 [0.00, 0.78]
Van Mensel et al. (2020)	0.40 [0.11, 0.70]
Olsson (2015)	0.41 [0.01, 0.81]
Feddermann et al. (2021)	0.44 [0.30, 0.58]
Manzano Vázquez (2014)	0.47 [−0.18, 1.12]
Bayram et al. (2019)	0.49 [0.13, 0.84]
Gallardo del Puerto & Gómez-Lacabex (2017)	0.51 [−0.18, 1.19]
Canga Alonso & Arribas García (2015)	0.52 [0.13, 0.92]
Prieto-Arranz et al. (2015)	0.52 [0.08, 0.96]
Salili & Lai (2003) [2]	0.57 [0.36, 0.78]
Lahuerta (2017, 2020) [2]	0.58 [0.29, 0.87]
Lahuerta (2017, 2020) [1]	0.60 [0.31, 0.88]
Martínez-Adrián & Gutiérrez-Mangado (2009)	0.61 [−0.26, 1.47]
Lo & Murphy (2010) [1]	0.62 [0.13, 1.10]
Goris et al. (2013) [2]	0.66 [0.25, 1.07]
Salili & Lai (2003) [1]	0.72 [0.56, 0.87]
Juan−Garau et al. (2015)	0.72 [0.30, 1.13]
Gallardo del Puerto et al. (2009)	0.80 [0.04, 1.55]
Martínez Agudo (2019, 2020)	0.86 [0.43, 1.30]
Rumlich (2017)	0.87 [0.68, 1.06]
Jexenflicker & Dalton-Puffer (2010)	0.90 [0.46, 1.35]
Verspoor et al. (2015)[2]	0.97 [0.35, 1.59]
Hendrikx & Van Goethem (2020)	0.97 [0.65, 1.29]
Martínez-Adrián & Gutiérrez−Mangado (2015)	1.00 [0.05, 1.95]
Goris et al. (2013)[1]	1.03 [0.58, 1.49]
Verspoor et al. (2015)[1]	1.06 [0.63, 1.49]
Dallinger et al. (2016)	1.09 [0.95, 1.24]
Lo & Murphy (2010)[2]	1.12 [0.55, 1.70]
Jiménez Catalán & Agustín Llach (2017)	1.13 [0.54, 1.72]
Merino & Lasagabaster (2018) [2]	1.17 [0.89, 1.45]
Lasagabaster (2008)	1.21 [0.78, 1.65]
Pérez Cañado & Lancaster (2017)	1.28 [0.43, 2.12]
Castellano-Risco et al. (2020)	1.82 [1.42, 2.22]
Merino & Lasagabaster (2018) [1]	1.83 [1.49, 2.18]
Hamidavi et al. (2016) [1]	1.83 [1.00, 2.66]
Hamidavi et al. (2016) [2]	2.26 [1.37, 3.15]

Figure 4 Forest plot for posttest effect sizes at the sample level (*k* = 41) (The dotted vertical line in the plot indicates zero.)

5.2. Moderator analyses

To investigate how the potential moderators were related to EMI-CLIL's overall effectiveness, we conducted moderator analyses. Note that moderator analyses were conducted in two phases for greater precision and that delayed posttest results were excluded due to their limited sample sizes (*N* = 8, *k* = 3).

In the first phase, a series of simple meta-regression analyses were employed to check how each moderator variable was related to EMI-CLIL's effectiveness without considering any relationships among moderators. Only one moderator, "Target linguistic dimensions," had statistically significant contributions (*p* = .01) to the overall effectiveness (see Table 1). Specifically, we found that learners in studies where the productive aspect of English was the main focus of evaluation showed lower English learning gains than in studies focusing on overall English skills or receptive aspects of English (β = -0.26, *SE* = 0.10, *z* = -2.64, *p* = .01, 95% CI [-0.46, -0.07]).

Table 1 Simple meta-regression for each moderator variable

Moderator variable	Category	# of ES	Posttest (*N* = 184; *k* = 41)				
			Est.	*SE*	*z*	*p*	95% CI
L1–English relation	Related	136	-0.01	0.15	-0.05	0.96	-0.31 ~ 0.30
	Not related	48			(reference)		
EMI-CLIL program intensity	≥ 50%	56	0.05	0.14	0.37	0.71	-0.22 ~ 0.32
	< 50%	126			(reference)		
Homogeneity confirmation	Not confirmed	148	0.08	0.14	0.55	0.58	-0.19 ~ 0.34
	Confirmed	36			(reference)		
Language test type	Self-developed	84	-0.20	0.11	-1.86	0.06	-0.42 ~ 0.01
	Validated	100			(reference)		
Target linguistic dimensions	Productive	119	-0.26**	0.10	-2.64	0.01	-0.46 ~ -0.07
	Others (overall, receptive)	65			(reference)		
Vocabulary-targeted	Yes	51	0.14	0.10	1.37	0.17	-0.06 ~ 0.35
	No	133			(reference)		

Note. Reference level of each moderator variable is marked "(reference)"; ** *p* < .01.

Table 2 Multiple meta-regression with all moderator variables

Moderator variables	Category	Posttest (*N* = 178; *k* = 39)				
		Est.	*SE*	*z*	*p*	95% CI
L1–English relation	Related	0.47*	0.19	2.49	0.01	0.10 ~ 0.84
	Not related			(reference)		
EMI-CLIL program intensity	≥ 50%	0.30	0.16	1.88	0.06	-0.01 ~ 0.62
	< 50%			(reference)		
Homogeneity confirmation	Not confirmed	0.29*	0.13	2.25	0.03	0.04 ~ 0.55
	Confirmed			(reference)		
Language test type	Self-developed	-0.17	0.10	-1.58	0.12	-0.37 ~ 0.04
	Validated			(reference)		
Target linguistic dimensions	Productive	-0.34**	0.11	-3.12	0.00	-0.55 ~ -0.12
	Others (overall, receptive)			(reference)		
Vocabulary-targeted	Yes	0.25*	0.10	2.49	0.01	0.05 ~ 0.45
	No			(reference)		

Note. Reference level of each moderator variable is marked "(reference)." As a total of six effect sizes from three samples were omitted due to missing values across the included moderators, the results were based on 178 effect sizes from 38 samples, after controlling for the length of EMI-CLILs. Additionally, learners' countries were statistically controlled for; * *p* < .05, ** *p* < .01

In the second phase of moderator analyses, we conducted a multiple meta-regression with all variables included as independent variables in the equation in order to compute more precise coefficients after controlling for other moderators. Four moderator variables, “L1-English relation,” “Homogeneity confirmation,” “Target linguistic dimensions,” and “Vocabulary-targeted,” reached statistical significance (*p* < .05), but “EMI-CLIL program intensity” and “Language test type” did not show any statistically significant moderating effect (*p* > .05; see Table 2).

The results showed that EMI-CLIL’s effectiveness was (1) larger where learners’ L1s were linguistically related to English (*β* = 0.47, *SE* = 0.19, *z* = 2.49, *p* = .01, 95% CI [0.10, 0.84]); (2) larger in studies where baseline differences between EMI-CLIL and mainstream conditions were not confirmed (*β* = 0.29, *SE* = 0.13, *z* = 2.25, *p* = .03, 95% CI [0.04, 0.55]); (3) smaller in studies where the productive aspect of English was the main evaluation focus than in studies focusing on overall English skills or receptive aspects of English (*β* = -0.34, *SE* = 0.11, *z* = -3.12, *p* = .00, 95% CI [-0.55, -0.12]); and (4) larger when vocabulary learning outcomes were targeted rather than other aspects of English knowledge and skills (*β* = 0.25, *SE* = 0.10, *z* = 2.49, *p* = .01, 95% CI [0.05, 0.45]).

6. Discussion

6.1. EMI-CLIL’s overall effectiveness for English learning

Our results, based on 184 posttest effect sizes from 41 samples (*N* = 6,654) and 8 delayed posttest effect sizes from 3 samples (*N* = 676), revealed that the overall effect sizes of EMI-CLIL were 0.73 (short term) and 1.01 (longer term). This represents a medium-sized overall effect according to Plonsky and Oswald’s (2014) field-specific benchmark of effect size for intergroup comparison. Therefore, we suggest EMI-CLIL is moderately beneficial for students’ English learning in EFL contexts compared to other types of L2 interventions in general. The present meta-analysis, by and large, corresponds to that of Lo and Lo (2014) regarding EMI-CLIL’s overall benefits for English learning.

We believe that this positive finding resulted from a complex combination of multiple components related to EMI-CLIL pedagogy. First, one contributing factor may be EMI-CLIL’s meaningful contexts for purposeful communication (Dallinger et al., 2016; Lorenzo et al., 2010; Lyster & Ruiz de Zarobe, 2018), through which learners may receive English input, interact with others in English, and produce output, all of which are key components of SLA (Krashen, 1982; Long, 1996; Swain, 1995). Second, learning content subjects in English may have served as a motivating factor (Genesee & Lindholm-Leary, 2013), as some samples in the included studies were preparing for higher education in English-speaking countries. Above all, it should be noted that EMI-CLIL groups received, on average, hundreds of additional hours of English instruction more than mainstream EFL groups, as reported by some included studies (e.g., Castellano-Risco et al., 2020; Jiménez Catalán & Agustín Llach, 2017; Martínez-Adrián & Gutiérrez-Mangado, 2009). This must be a powerful variable that could account for the superiority of the EMI-CLIL approach (Macaro, 2018).

6.2. The roles of moderator variables

The findings from the moderator analyses revealed that (1) EMI-CLIL's overall effectiveness can be significantly influenced by two learner factors (i.e., learners' L1 and baseline intergroup differences) and that (2) certain linguistic aspects were particularly sensitive to EMI-CLIL. These findings are worth discussing in depth to provide evidence-based pedagogical directions for classroom teachers in EMI-CLIL contexts.

6.2.1. Influential learner-related moderator variables

EMI-CLIL's overall effectiveness was significantly influenced by learners' L1 and baseline intergroup differences. Regarding the former, the L1-English relation turned out to be a significant moderator. This result indicates learners whose L1s were more closely related to English (mostly European EFL learners) showed better English learning outcomes than those whose L1s were less related (Asian EFL learners). The potential advantage for the former may derive from structural L1-L2 similarities, which have been suggested to facilitate L2 learning (e.g., Lado, 1957; Stockwell et al., 1965). The debate about positive and negative transfer is beyond the scope of this paper, but one potential implication of the role played by language typology is that EMI-CLIL teachers may need to provide more language scaffolding when students' L1 is less closely related to English; this could involve raising students' metalinguistic awareness (e.g., morphology, grammar, sentence structure).

Additionally, we found that baseline intergroup differences significantly influenced EMI-CLIL's overall effectiveness. We included homogeneity confirmation to address a methodological issue in EMI-CLIL research, namely selection bias; it has been claimed that EMI-CLIL groups are more motivated and have higher levels of English proficiency than their mainstream counterparts prior to EMI-CLIL (e.g., Bruton, 2013; Macaro, 2018). The results showed that studies which did not confirm homogeneity between EMI-CLIL and mainstream groups revealed larger effect sizes. That is, when there are baseline intergroup differences, the advantaged EMI-CLIL group shows even more positive learning outcomes. Thus, our result corroborates claims that EMI-CLIL research suffers from selection bias. Echoing Lo and Lo's (2014) results, we highlight the need to identify more comparable comparison groups (e.g., high aptitude students; Verspoor et al., 2015) and check intergroup homogeneity (or at least conduct statistical adjustments).

6.2.2. Linguistic aspects sensitive to EMI-CLIL

In addition to some learner-related moderators, we found that certain linguistic aspects were particularly sensitive to the EMI-CLIL approach. For example, the moderator analyses revealed that vocabulary is a particular beneficiary, which accords with Dalton-Puffer's (2008) summary of previous CLIL research findings. Such benefits of EMI-CLIL to English vocabulary learning may result from wider vocabulary exposure, including subject-specific academic vocabulary and low-frequency words (Dalton-Puffer, 2007; Macaro, 2018).

Furthermore, we found that the EMI-CLIL condition was less favorable for the development of productive skills, compared to their receptive counterparts and overall proficiency. Dallinger et al. (2016) suggested a possible explanation: as producing output "is encouraged but usually not forced, their productive skills (speaking, writing) might benefit to a smaller extent" (p. 24). This is supported by studies revealing limited opportunities for teacher and peer L2 interactions in EMI-CLIL classrooms (e.g., Lo & Macaro, 2012). One implication for teacher education and pedagogy is that EMI-CLIL teachers, most of whom are content subject specialists, should be informed about and try to provide favorable L2 learning conditions (e.g., opportunities to interact and use the target L2).

6.3. Limitations and future research

We address this study's limitations here in view of their implications for future EMI-CLIL studies and meta-analyses. First, we could not include some potentially important moderators due to a lack of detailed descriptions in the methodology sections of the selected studies. Some of these potential moderators, the roles of which have not generally been discussed in previous studies, include (1) the quality of English input in EMI-CLIL lessons (Van Mensel et al., 2020), (2) EMI-CLIL program quality control systems (Verspoor et al., 2015), and (3) (non-)nativeness of instructors (Gallardo del Puerto & Gómez-Lacabex, 2017). Thus, future EMI-CLIL studies should provide more detailed descriptions of their EMI-CLIL contexts and directly examine the aforementioned moderator variables.

Second, the present meta-analysis focused only on English learning outcomes and did not include studies examining content learning outcomes. However, both types of outcomes should be analyzed to provide a more comprehensive view of EMI-CLIL's effectiveness (e.g., Graham et al., 2018; Lo & Lo, 2014);

this would also align with Macaro's (2018) "cost-benefit" perspective. It remains unclear as to how much the development of English competence "costs" in terms of its effect on content learning. Thus, future meta-analyses should widen their search scope and include studies on both language and content learning outcomes. Notably, the costs and benefits of EMI-CLIL may not be restricted to student learning outcomes; other aspects such as academic and career prospects and international mobility may be considered, notwithstanding the complications of capturing these intangible benefits through meta-analyses.

7. Conclusion

Despite these limitations, the present study significantly contributes to EMI-CLIL research by effectively synthesizing the findings of relevant studies conducted in the last two decades with a rigorous multilevel meta-analytic approach. Our perusal of the selected studies revealed that careful consideration of potential confounding variables has only recently emerged, indicating that there is plenty of room for improvements in the methodological designs of EMI-CLIL studies targeting secondary education. Nevertheless, our findings lend further weight to the argument that researching EMI-CLIL is worthwhile, given its potential contributions to English competence development among secondary-level learners, as well as its potential to serve as an alternative to traditional EFL instruction.

References[3]

An, J., Macaro, E., & Childs, A. (2019). Language focused episodes by monolingual teachers in English medium instruction science lessons. *Journal of Immersion and Content-Based Language Education, 7*(2), 166-191. https://doi.org/10.1075/jicb.18019.an

Assink, M., & Wibbelink, C. J. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology, 12*(3), 154-174. https://doi.org/10.20982/tqmp.12.3.p154

*Basterrechea, M., & García Mayo, M. del P. (2014). Dictogloss and the production of the English third person "-s" by CLIL and mainstream EFL learners: A comparative study. *International Journal of English Studies*, *14*(2), 77-98. https://doi.org/10.6018/j.177321

*Bayram, D., Öztürk, R. Ö., & Atay, D. (2019). Reading comprehension and vocabulary size of CLIL and non-CLIL students: A comparative study. *Language Teaching and Educational Research*, *2*(2), 101-113. https://doi.org/10.35207/later.639337

Beaufils, V., & Tomin, J. (2020, October 30). Stochastic approach to worldwide language classification: The signals and the noise towards long-range exploration. *SocArXiv Papers*. https://doi.org/10.31235/osf.io/5swba

Broca, Á. (2016). CLIL and non-CLIL: Differences from the outset. *ELT Journal, 70*(3), 320-331. https://doi.org/10.1093/elt/ccw011

Bruton, A. (2013). CLIL: Some of the reasons why . . . and why not. *System, 41*(3), 587-597. https://doi.org/10.1016/j.system.2013.07.001

Cambridge ESOL. (2008). *Official examination papers from University of Cambridge ESOL examinations*. Cambridge University Press.

*Canga Alonso, A., & Arribas García, M. (2015). The benefits of CLIL instruction in Spanish students' productive vocabulary knowledge. *Encuentro*, *24*, 15-31. Retrieved from http://www3.uah.es/encuentrojournal/index.php/encuentro/issue/viewIssue/12/24

*Castellano-Risco, I., Alejo-González, R., & Piquer-Píriz, A. M. (2020). The development of receptive vocabulary in CLIL vs EFL: Is the learning context the main variable? *System*, *91*, 102263. https://doi.org/10.1016/j.system.2020.102263

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

*Dallinger, S., Jonkmann, K., Hollm, J., & Fiege, C. (2016). The effect of content and language integrated learning on students' English and history competences:

[3] References marked with an asterisk indicate studies included in the meta-analysis.

Killing two birds with one stone? *Learning and Instruction*, *41*, 23-31. https://doi.org/10.1016/j.learninstruc.2015.09.003

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms*. John Benjamins.

Dalton-Puffer, C. (2008). Outcomes and processes in content and language integrated learning (CLIL): Current research from Europe. In W. Delanoy & L. Volkmann (Eds.), *Future perspectives for English language teaching* (pp. 139-157). Carl Winter.

Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, *31*, 182-204. https://doi.org/10.1017/S0267190511000092

Dalton-Puffer, C. (2013). A construct of cognitive discourse functions for conceptualising content-language integration in CLIL and multilingual education. *European Journal of Applied Linguistics, 1*(2), 216-253. https://doi.org/10.1515/eujal-2013-0011

Dalton-Puffer, C., Llinares, A., Lorenzo, F., & Nikula, T. (2014). "*You can stand under my umbrella*": Immersion, CLIL and bilingual education. A response to Cenoz, Genesee & Gorter (2013). *Applied Linguistics, 35*(2), 213-218. https://doi.org/10.1093/applin/amu010

Evans, S. (2017). Language policy in Hong Kong education: A historical overview. *European Journal of Language Policy, 9*(1), 67-84. https://doi.org/10.3828/ejlp.2017.5

*Feddermann, M., Möller, J., & Baumert, J. (2021). Effects of CLIL on second language learning: Disentangling selection, preparation, and CLIL-effects. *Learning and Instruction*, *74*, 101459. https://doi.org/10.1016/j.learninstruc.2021.101459

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods, 52*, 2031-2052. https://doi.org/10.3758/s13428-020-01373-9

*Gallardo del Puerto, F., & Gómez-Lacabex, E. (2017). Oral production outcomes in CLIL: An attempt to manage amount of exposure. *European Journal of Applied Linguistics*, *5*(1), 31-54. https://doi.org/10.1515/eujal-2015-0035

*Gallardo del Puerto, F., Gómez-Lacabex, E., & García Lecumberri, M. (2009). Testing the effectiveness of content and language integrated learning (CLIL) in foreign language contexts: The assessment of English pronunciation. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (Eds.), *Content and language integrated learning: Evidence from research in Europe* (pp. 63-80). Multilingual Matters. https://doi.org/10.21832/9781847691675-007

Genesee, F., & Lindholm-Leary, K. (2013). Two case studies of content-based language education. *Journal of Immersion and Content-Based Language Education, 1*(1), 3-33. https://doi.org/10.1075/jicb.1.1.02gen

*Gierlinger, E., & Wagner, T. (2016). The more the merrier: Revisiting CLIL-based vocabulary growth in secondary education. *LACLIL*, *9*(1), 37-63. https://doi.org/10.5294/laclil.2016.9.1.3

*Goris, J., Denessen, E., & Verhoeven, L. (2013). Effects of the content and language integrated learning approach to EFL teaching: A comparative study. *Written Language & Literacy*, *16*(2), 186-207. https://doi.org/10.1075/wll.16.2.03gor

Goris, J., Denessen, E., & Verhoeven, L. (2019). Effects of content and language integrated learning in Europe: A systematic review of longitudinal experimental studies. *European Educational Research Journal*, *18*(6), 675-698. https://doi.org/10.1177/1474904119872426

Graham, K. M., Choi, Y., Davoodi, A., Razmeh, S., & Dixon, L. Q. (2018). Language and content outcomes of CLIL and EMI: A systematic review. *LACLIL*, *11*(1), 19-37. https://doi.org/10.5294/laclil.2018.11.1.2

*Gutiérrez-Mangado, M. J., & Martínez-Adrián, M. (2018). CLIL at the linguistic interfaces. *Journal of Immersion and Content-Based Language Education*, *6*(1), 85-112. https://doi.org/10.1075/jicb.17002.gut

*Gutiérrez Ortiz, M. (2020). Assessing the development of second language syntax in content and language integrated learning. *Revista De Lenguas Para Fines Específicos*, *26*(2), 111-130. Retrieved from https://ojsspdc.ulpgc.es/ojs/index.php/LFE/article/view/1333

*Hamidavi, N., Shekaramiz, M., & Gorjian, B. (2016). The effect of CLIL method on teaching reading comprehension to junior high school students. *Modern Journal of Language Teaching Methods*, *6*(9), 64-73. Retrieved from http://mjltm.org/article-1-65-en.html

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128. https://doi.org/10.3102/10769986006002107

*Hendrikx, I., & Van Goethem, K. (2020). Receptive knowledge of intensifying adjectival compounds: Belgian French-speaking CLIL and non-CLIL learners of Dutch and English. *International Journal of Bilingual Education and Bilingualism*. Advance Publication Online. https://doi.org/10.1080/13670050.2020.1720592

Hu, J., & Gao, X. (2021). Understanding subject teachers' language-related pedagogical practices in content and language integrated learning classrooms. *Language Awareness, 30*(1), 42-61. https://doi.org/10.1080/09658416.2020.1768265

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160-212. https://doi.org/10.1111/lang.12034

*Jexenflicker, S., & Dalton-Puffer, C. (2010). The CLIL differential: Comparing the writing of CLIL and non-CLIL students in higher colleges of technology. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 169-189). John Benjamins.

*Jiménez Catalán, R. M., & Agustín Llach, M. P. (2017). CLIL or time? Lexical profiles of CLIL and non-CLIL EFL learners. *System*, *66*, 87-99. https://doi.org/10.1016/j.system.2017.03.016

Johnson, R. K., & Swain, M. (1994). From core to content: Bridging the L2 proficiency gap in late immersion. *Language and Education, 8*(4), 211-229. https://doi.org/10.1080/09500789409541392

*Juan-Garau, M., Prieto-Arranz, J. I., & Salazar-Noguera, J. (2015). Lexico-grammatical development in secondary education CLIL learners. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 179-195). Springer. https://doi.org/10.1007/978-3-319-11496-5_11

Krashen, S. (1982). *Principles and practice in second language acquisition*. Prentice Hall.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.

*Lahuerta, A. (2017). Syntactic complexity in secondary-level English writing: Differences among writers enrolled on bilingual and non-bilingual programmes. *Porta Linguarum*, *28*, 67-80. Retrieved from https://www.ugr.es/~portalin/articulos/PL_numero28/5%20Lahuerta.pdf

*Lahuerta, A. (2020). Analysis of accuracy in the writing of EFL students enrolled on CLIL and non-CLIL programmes: The impact of grade and gender. *The Language Learning Journal*, *48*(2), 121-132. https://doi.org/10.1080/09571736.2017.1303745

*Lasagabaster, D. (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal*, *1*, 30-41. https://doi.org/10.2174/1874913500801010030

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, *40*(5), 721-753. https://doi.org/10.1093/applin/amy012

*Lee, J. (2020). Assessing the effects of CLIL on Korean high school students' writing. *Linguistic Research*, *37*, 89-112. https://doi.org/10.17250/khisli.37..202009.004

Lee, J. H., & Lee, H. (2022). Teachers' verbal lexical explanation for L2 vocabulary learning: A meta-analysis. *Language Learning*, *72*(2), 576-612. https://doi.org/10.1111/lang.12493

*Lin, L. H. F., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes*, *9*(4), 255-266. https://doi.org/10.1016/j.jeap.2010.09.002

Lo, Y. Y., & Lo, E. S. C. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research*, *84*(1), 47-73. https://doi.org/10.3102/0034654313499615

Lo, Y. Y., & Macaro, E. (2012). The medium of instruction and classroom interaction: Evidence from Hong Kong secondary schools. *International Journal of Bilingual Education and Bilingualism, 15*(1), 29-52. https://doi.org/10.1080/13670050.2011.588307

*Lo, Y. Y., & Murphy, V. A. (2010). Vocabulary knowledge and growth in immersion and regular language-learning programmes in Hong Kong. *Language and Education*, *24*(3), 215-238. https://doi.org/10.1080/09500780903576125

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). Academic Press.

Lorenzo, F., Casal, S., & Moore, P. (2010). The effects of content and language integrated learning in European education: Key findings from the Andalusian bilingual sections evaluation project. *Applied Linguistics*, *31*(3), 418-442. https://doi.org/10.1093/applin/amp041

Lyster, R., & Ruiz de Zarobe, Y. (2018). Introduction: instructional practices and teacher development in CLIL and immersion school settings. *International Journal of Bilingual Education and Bilingualism, 21*(3), 273-274. https://doi.org/10.1080/13670050.2017.1383353

Macaro, E. (2018). *English medium instruction*. Oxford University Press.

*Manzano Vázquez, B. (2014). Lexical transfer in the written production of a CLIL group and a non-CLIL group. *International Journal of English Studies*, *14*(2), 57-76. https://doi.org/10.6018/j.166251

Marsh, D. (Ed.) (2002). *CLIL/EMILE. The European dimension. Actions, trends, and foresight potential*. University of Jyväskylä.

*Martínez-Adrián, M., & Gutiérrez-Mangado, M. J. (2009). The acquisition of English syntax by CLIL learners in the Basque country. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (Eds.), *Content and language integrated learning: Evidence from research in Europe* (pp. 176-196). Multilingual Matters.

*Martínez-Adrián, M., & Gutiérrez-Mangado, M. J. (2015). L1 use, lexical richness, accuracy and complexity in CLIL and non-CLIL learners of English. *Atlantis, Journal of the Spanish Association for Anglo-American Studies*, *37*(2), 175-197. https://www.atlantisjournal.org/index.php/atlantis/article/view/273

*Martínez Agudo, J. (2019). Which instructional programme (EFL or CLIL) results in better oral communicative competence? Updated empirical evidence from a monolingual context. *Linguistics and Education*, *51*, 69-78. https://doi.org/10.1016/j.linged.2019.04.008

*Martínez Agudo, J. (2020). The impact of CLIL on English language competence in a monolingual context: A longitudinal perspective. *The Language Learning Journal*, *48*(1), 36-47. https://doi.org/10.1080/09571736.2019.1610030

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

Morton, T., & Llinares, A. (2017). Content and language integrated learning: Type of programme or pedagogical model? In A. Llinares & T. Morton (Eds.), *Applied linguistics perspectives on CLIL* (pp. 1-16). John Benjamins.

Murphy, V. A., Arndt, H., Briggs J. G., Chalmers, H., Macaro, E., Rose, H., Vanderplank, R., & Woore, R. (2020). *Foreign language learning and its impact on wider academic outcomes: A rapid evidence assessment*. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Publications/Foreign_language_learning_and_its_impact _on_wider_academic_outcomes_-_A_rapid_evidence_assessment.pdf

Nunan, D. (2011). *Teaching English to young learners*. Anaheim University Press.

*Olsson, E. (2015). Progress in English academic vocabulary use in writing among CLIL and non-CLIL students in Sweden. *Moderna Språk*, *109*(2), 51-74. https://ojs.ub.gu.se/index.php/modernasprak/article/view/3261

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*(71), 1-9. https://doi.org/10.1136/bmj.n71

Pérez-Cañado, M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism, 15*(3), 315-341. https://doi.org/10.1080/13670050.2011.630064

Pérez-Cañado, M. L. (2016). From the CLIL craze to the CLIL conundrum: Addressing the current CLIL controversy. *Bellaterra Journal of Teaching & Learning Language & Literature, 9*(1), 9-31. https://doi.org/10.5565/rev/jtl3.667

*Pérez Cañado, M. L., & Lancaster, N. K. (2017). The effects of CLIL on oral comprehension and production: A longitudinal case study. *Language, Culture and Curriculum*, *30*(3), 300-316. https://doi.org/10.1080/07908318.2017.1338717

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/10.1111/lang.12079

*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. https://doi.org/10.1007/978-3-319-11496-5_8

*Rallo Fabra, L., & Jacob, K. (2015). Does CLIL enhance oral skills? Fluency and pronunciation errors by Spanish-Catalan learners of English. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 163-177). Springer. https://doi.org/10.1007/978-3-319-11496-5_10

*Rallo Fabra, L., & Juan-Garau, M. (2011). Assessing EFL pronunciation in a semi-immersion setting: The effects of CLIL instruction on Spanish-Catalan learners perceived comprehensibility and accentedness. *Poznań Studies in Contemporary Linguistics*, *47*(1), 96-108. https://doi.org/10.2478/psicl-2011-0008

Rose, H., Macaro, E., Sahan, K., Aizawa, I., Zhou, S., & Wei, M. (2021). Defining English medium instruction: Striving for comparative equivalence. *Language Teaching*. https://doi.org/10.1017/S0261444821000483

Ruiz de Zarobe, Y. (2011). Which language competencies benefit from CLIL? An insight into applied linguistics research. In Y. Ruiz de Zarobe, J. M. Sierra, & F. Gallardo del Puerto (Eds.), *Content and foreign language integrated learning: Contributions to multilingualism in European contexts* (pp. 129-153). Peter Lang.

*Rumlich, D. (2017). CLIL theory and empirical reality: Two sides of the same coin? *Journal of Immersion and Content-Based Language Education, 5*(1), 110-134. https://doi.org/10.1075/jicb.5.1.05rum

*Salili, F., & Lai, M. K. (2003). Learning and motivation of Chinese students in Hong Kong: A longitudinal study of contextual influences on students' achievement orientation and performance. *Psychology in the Schools*, *40*(1), 51-70. https://doi.org/10.1002/pits.10069

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Stockwell, R., Bowen, J., & Martin, J. (1965). *The grammatical structures of English and Spanish*. University of Chicago Press.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125-144). Oxford University Press.

Tabachnik, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Allyn and Bacon.

*Van Mensel, L., Bulon, A., Hendrikx, I., Meunier, F., & Van Goethem, K. (2020). Effects of input on L2 writing in English and Dutch: CLIL and non-CLIL learners in French-speaking Belgium. *Journal of Immersion and Content-Based Language Education*, *8*(2), 173-199. https://doi.org/10.1075/jicb.18034.van

*Verspoor, M., de Bot, K., & Xu, X. (2015). The effects of English bilingual education in the Netherlands. *Journal of Immersion and Content-Based Language Education*, *3*(1), 4-27. https://doi.org/10.1075/jicb.3.1.01ver

Viechtbauer, W. (2010). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software, 36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112-125. https://doi.org/10.1002/jrsm.11