Impact of English Medium Instruction in Higher Education: A Multi-Level Meta-Analysis

To Cite:

Lee, H., Rose, H., Macaro, E., & Lee, J. H. (2025). Effectiveness of EMI in higher education: A multi-level meta-analysis. *System*, *133*, 103755. https://doi.org/10.1016/j.system.2025.103755

Abstract

In this study, we conducted a multi-level meta-analysis to systematically synthesise empirical evidence on the effectiveness of English Medium Instruction (EMI) in higher education for enhancing students' content learning and English proficiency. A total of 41 samples (N = 8,747) from 28 quantitative studies were analysed, including 23 samples on content learning (N = 7,659) and 18 on English performance (N = 1,088). A comparison of EMI students' post-test and pre-test scores (within-group comparison) revealed that students significantly improved on measures of content knowledge (d = 1.57, p < .001) and English proficiency (d = 0.81, p < .001), although we acknowledge that this finding does not account for external factors or establish causality. A comparison of post-test performance between EMI and non-EMI groups (between-group comparison), which aimed to assess EMI's effectiveness relative to non-EMI, showed that EMI students achieved comparable outcomes in content learning (d = 0.13, p = 0.14) and significantly outperformed non-EMI students in English development (d = 0.33, p = 0.009). Moderator analyses further revealed that studies with methodological limitations, such as selection biases for within-group comparisons or unaddressed group homogeneity for between-group comparisons, often overestimate the effectiveness of EMI for content learning.

Keywords: content learning; English medium instruction; English learning; multi-level meta-analysis

Introduction

The considerable body of research that exists on the phenomenon and practice of English Medium Instruction (EMI) now includes a substantial and growing interest in the success or 'effectiveness' of the practice of teaching academic subjects through the medium of English to students whose first language is not English (e.g., Aguilar & Muñoz, 2014; Aizawa et al., 2024; Bälter et al., 2024; García-Álvarez de Perea & Ramírez-García, 2024; Guo et al., 2018; Lei & Hu, 2014; Lin & He, 2019; Rogier, 2014; Satayev et al., 2022; Sato & Hemmi, 2022). However, we should note that interest in 'effectiveness' (the term used in this paper) has not overshadowed the continued concern that some authors have expressed concerning the possible additional (potentially negative) impact that EMI might have on a given language, on national identity or on a socio-cultural group. It is therefore of some importance to attempt to define effectiveness and subsequently evaluate evidence of whether EMI has been effective to an extent that counterbalances arguments related to its potential negative impacts.

Effectiveness in EMI higher education contexts (the focus of our study) is generally evaluated, or specifically measured, according to two dimensions: gains in academic content knowledge, and gains in English language proficiency (Rose et al., 2021). In the case of the first dimension, stakeholders would want to ascertain, through research, that content outcomes were at least as good as outcomes achieved through first language (L1) medium instruction (Macaro, 2022). However, given that, for many disciplines in higher education, very large quantities of research are published in English then it would not be unreasonable to expect that EMI might, in the long run, outperform its non-EMI counterpart. In the case of the second dimension – increase in English language proficiency – again it is not unreasonable to expect that students would improve their English more quickly and more deeply in EMI programmes than in non-EMI

equivalents. Indeed, the expectation of greater English proficiency is a cornerstone of a number of national language policies in which there is an ambition to develop much higher levels of bilingualism among the population (see, for example, Ferrer & Lin, 2024 regarding Taiwan; Rose & McKinley, 2018 regarding Japan, and Fang, 2018 regarding China).

To evaluate effectiveness of these two dimensions, however, is not as simple as it sounds. First, in terms of content learning there are a number of sub-factors that need to be taken into account: Should researchers take into account prior content learning at high school (or in a previous phase of education) and in which language that took place?; Are there disciplinary differences that might allow faster learning through EMI in one as opposed to another discipline? Do the various and differing models of EMI such as full and partial programs (see below) make a difference as to the likelihood of recognizable content learning effectiveness?

In terms of English proficiency effectiveness there are also obvious factors to be taken into account. What aspect of the English language is being measured (vocabulary; grammatical correctness; fluency; reported self-confidence), and are these different aspects likely to bias one group being measured against another? Similarly to the above, at what point is it reasonably justifiable to measure these aspects of language proficiency and how does one control for lack of similarities between groups (e.g. social status where one group may have received intensive private tuition).

So, attempting to answer the question of effectiveness of EMI is a complex one but its complexity should certainly not deter researchers from attempting to do so, but merely encourage them to exercise caution in summarizing their findings. While the factors included in this meta-analysis have been largely limited to those that have been researched by a requisite number of

studies, we trust that readers will situate our findings within this wider perspective of effectiveness.

Literature Review

Theoretical Foundations of English Medium Instruction in Higher Education

As EMI has emerged as an educational phenomenon, there have been ongoing debates over where the research field should theoretically be positioned (Macaro & Aizawa, 2024). In a recent review of research, some scholars have noted: "While it may not be accurate to say there is a complete "lack" of theory in English Medium Instruction (EMI) research, it's fair to argue that there is a need for more comprehensive, robust, and specific theoretical frameworks to guide this field of study." (Curle et al., 2024, p. 4). Nonetheless, there are certain cognitive and linguistic theories which can be applied to interpret content and language gains in an EMI educational environment.

Cognitive and Linguistics Theories Underpinning EMI Effectiveness

Relevant theories related to language and content development in EMI centre around the perceived benefits of greater levels of L2 input and interaction afforded by learning content through English as an additional or second language. As Costa and Mariotti (2023) claim, "field exposure to input is the prime mover of the teaching-learning process and… …its quality is fundamental for the development of discipline-specific knowledge with particular reference to university settings" (p. i). Macaro (2018) has previously noted that university EMI lectures are heavily characterized by large quantities of teacher input, further highlighting the applicability of input theories. Thus it is this input, alongside academic readings and other teaching materials, that is thought to primarily drive learning processes in EMI (Costa & Mariotti, 2023). According to VanPatten (2009) input and interaction theories are embedded in the notion that when learner

attention is simultaneously focused on meaning and the properties of language, learning occurs most effectively. Parallels can also be drawn with theories of bilingual education, where the *Interdependent Hypothesis* purports that cognitive and linguistic capacity in one's L1 enhances the development of capacities in the L2 in such a way that cognitive or literacy skills develop more rapidly in the second language, if already acquired in the first (Cummins, 1979). Thus, according to theories of SLA and bilingual education, English medium instruction in later stages of education should–in theory–lead to positive outcomes in language development *and* drive content learning.

However, there are other educational theories that suggest a possible negative effect of EMI on learning, one of which is *Cognitive Load Theory*. Cognitive Load Theory posits that learning processes may be limited by the cognitive load, that is the amount of information our working memory can process at any given time (Sweller, 2011). Drawing on this theory, Curle et al (2024) illustrate:

> "For instance, a student who is not proficient in English might experience a high extraneous cognitive load when trying to understand a lecture in English, leaving less cognitive resources for processing the actual academic content of the lecture. According to the Cognitive Load Theory, this could potentially lead to poorer academic outcomes for the student." (p. 4).

Similarly, SLA theories of *comprehensible input* have seen a resurgence in research applicability due to their relevance to EMI (Macaro, 2018; Costa & Mariotti, 2023). The theory of comprehensible input (e.g. Krashen, 1982) posits that the level of language must be within the realms of understandability for students, or just above their current level of proficiency, for learning to occur. Similar notions can be found in the *Threshold Hypothesis* (Cummins, 2021),

which proposes that learners require a minimum threshold level of proficiency in an L2 before they can benefit from its use as a medium of instruction in school, otherwise there may be negative effects on learning. Taken together, these linguistic and cognitive theories suggest that while EMI affords opportunities for greater content and language input for many learners, if lecture and reading content is too difficult or beyond the linguistic capabilities of students, learning outcomes may be negatively affected. Understanding how EMI affects learning outcomes, therefore, is a matter of both practical and theoretical importance to the field.

Synthetic Evidence on the Effectiveness of English Medium Instruction

In this section, we present previous efforts to synthesise empirical research on the effectiveness of EMI and their implications and limitations, on the basis of which the present meta-analysis was designed and conducted. While medium of instruction has been investigated in several previous research syntheses of CLIL (see, for example, Kaiypova et al., 2025; Lee et al., 2023; Lo & Lo, 2014), there have been fewer efforts that have explicitly and exclusively targeted EMI research.

Synthetic Evidence on the Effectiveness of EMI for Content Learning

In one of the first full-scale systematic reviews on this topic, Macaro et al. (2018) conducted an in-depth review of 83 studies on EMI conducted in higher education and identified several themes surrounding these studies, two of which are relevant to the topic of this meta-analysis. For one of these themes–the impact of EMI on content comprehension and learning, only four studies were identified, and these studies showed some contrasting results. Overall, Macaro et al. suggested that due to the small number of studies and their methodological shortcomings (e.g., validity of instruments, the issue of homogeneity of EMI and non-EMI groups), they could not draw any meaningful conclusions about the effects of EMI. Graham et

al.'s (2018) systematic review focused on EMI and Content and Language Integrated Learning (CLIL) studies that had measured student learning outcomes. Of the 25 studies identified in their review, only three were studies that examined content learning outcomes in higher education, and these showed inconsistent results, with two studies finding no significant difference and one study showing the superiority of the EMI condition over its non-EMI counterpart. Finally, Peng and Xie's (2021) meta-analysis focused exclusively on Chinese studies in higher education, and compared the effectiveness of the EMI condition with that of the Chinese-medium instruction (CMI) condition. Regarding content learning outcomes, their results revealed that the EMI group performed significantly better than their CMI counterparts, with this difference showing an overall medium effect size ($d = 0.67$, $SE = 0.134$, $p < .001$). Commenting on this rather unexpected finding, they explained that approximately 74% of their dataset was from medical disciplines, which may have influenced the overall effect size; indeed, no significant difference was found when only samples related to non-medical disciplines were considered ($d = –0.06$, $p > .05$, 95% CI [-0.18, 0.05]).

Synthetic Evidence on the Effectiveness of EMI for English Learning

The aforementioned systematic review by Macaro et al. (2018) identified 10 studies on English learning outcomes, and found that they generally showed improvement, but only in specific skills or subcomponents of skills. Macaro et al. further highlighted the need to distinguish between general and academic English proficiency when measuring the effectiveness of EMI, as it is expected that EMI may be more beneficial for the latter, suggesting that the type of English proficiency may be a potential moderator of the effectiveness of EMI. In the systematic review by Graham et al. (2018), only one study (Yang, 2015) was identified regarding language learning outcomes, and its results showed that the EMI (CLIL) condition favored

receptive skills, although this study had some methodological issues (e.g., small sample size and no comparison group involved in the study). The aforementioned meta-analysis by Peng and Xie (2021) additionally analyzed English learning achievements with 12 independent samples, and showed that the EMI group had better English learning outcomes, with a large effect size (d = 1.58, SE = 0.34, $p < .001$).

Although the systematic reviews by Macaro et al. (2018) and Graham et al. (2018) provide a detailed synthesis of existing studies on the impact of EMI on content and English learning outcomes, the results of these reviews are far from conclusive on the effectiveness of EMI on student learning outcomes in tertiary education. It is also noteworthy that more studies have been conducted on this topic since these research syntheses were conducted, pointing to the need for an updated meta-analysis that includes more recent studies. Furthermore, given the nature of Peng and Xie's (2021) meta-analysis (i.e., focusing exclusively on the Chinese context), a more comprehensive meta-analytic effort that includes other educational contexts seems necessary–a gap that the present meta-analysis aims to fill.

Key Features Influencing the Effectiveness of English Medium Instruction

According to a recent meta-analytic structural equation modelling study on the mechanisms underlying successful EMI programmes in higher education (Authors, 2025a), which analysed 50 studies (N = 15,032), three key features were identified as influencing EMI success. These features include L1-English proximity (i.e., whether the first language [L1] and English are linguistically related), the EMI context (i.e., full EMI vs. partial EMI), and institutional support (i.e., whether any institutional support was provided or not). These findings were derived from a comprehensive review of theoretical frameworks and existing literature,

with success often measured by outcomes such as students' GPAs or test results following EMI programmes.

A further meta-analysis by Authors (2025b) focused on the effectiveness of EMI at the primary level rather than higher education. Their study, based on 28 samples (N = 214,103), suggested that learners' age, EMI context, and confirmation of homogeneity were potential moderators influencing both content and language learning. In secondary education, studies by Kaiypova et al. (2025) and Lee et al. (2023), which analysed 29 samples (N = 36,905) and 44 samples (N = 7,434) respectively, identified the following potential moderators: L1-English proximity, EMI context, homogeneity confirmation, test language, the nature of the content subject, and the target linguistic dimension.

When critically reviewing these findings to assess their relevance to EMI in higher education, across diverse learner backgrounds, instructional contexts, and methodological considerations, eight key features emerge as influential to EMI effectiveness. (1) L1-English proximity would play a significant role, as learners whose first language (L1) shares linguistic or structural similarities with English may adapt more easily to EMI compared to those whose L1 is unrelated. (2) Institutional support, such as the provision of language workshops, tutoring, and faculty training, would be another pivotal factor in facilitating EMI success. (3) Aspects of the EMI context would be also critical. The extent of EMI exposure, whether a programme adopts full or partial EMI, can affect both content knowledge acquisition and English language development. (4) The duration of EMI programmes also would matter, as the outcomes of short-term (semester-long) programmes may differ from those of longer-term programmes spanning one to four years. (5) The subject matter taught through EMI would be an additional consideration, as STEM courses, for example, may pose less challenges due to their specialised

terminology (i.e., less language loaded) when compared to non-STEM disciplines. (6) The measurement of English learning outcomes also would vary, with some studies focusing on academic English skills tailored to higher education while others measure general English proficiency, which may not fully capture the language demands of EMI programmes. (7) Homogeneity confirmation would influence the reliability of EMI research findings, as studies that failed to check homogeneity in their samples would tend to include more motivated and talented participants. Similar to the purpose of homogeneity confirmation, (8) entry prerequisites, such as requiring specific levels of English proficiency before enrolment, would serve as a gatekeeping mechanism to ensure students are adequately prepared. These eight features collectively underscore the multifaceted nature of EMI in higher education and the variability in its effectiveness across institutions and contexts.

Current Study

In this study, we conducted a meta-analysis to systematically synthesise prior empirical research on the effectiveness of English Medium Instruction (EMI) programs in higher education. We adopted a multi-level meta-analytic approach to account for multiple effect sizes within studies reporting more than one measurement. This method facilitated the calculation of overall average effect sizes and the conduct of subgroup analyses to examine moderator effects while addressing the multi-level structure of the dataset (e.g., H. Lee et al., 2019; J.H. Lee et al., 2023). Specifically, effect sizes at the measurement level were nested within the sample level, enabling more precise estimates and mitigating dependence issues arising from multiple effect sizes within a single sample (Hox et al., 2010). Through this approach, we aimed to investigate the effectiveness of EMI programmes for enhancing learner's content knowledge and English proficiency. This was examined from both non-experimental perspectives (e.g., pre-test and post-

test measurements within EMI groups) and (quasi-) experimental perspectives (e.g., post-test comparisons between EMI and non-EMI groups). The following research questions guided the current study.

Research Question 1:

What is the overall average effect size of English Medium Instruction (EMI) programmes in higher education for content learning?

RQ 1.1: What is the overall average effect size for content learning within-group samples (from start to end of the programme)?

RQ 1.2: What is the overall average effect size for content learning between-group samples (EMI vs. instruction in students' first languages)?

Research Question 2:

What is the overall average effect size of English Medium Instruction (EMI) programmes in higher education for English learning?

RQ 2.1: What is the overall average effect size for English learning within-group samples (from start to end of the programme)?

RQ 2.2: What is the overall average effect size for English learning between-group samples (EMI vs. instruction in students' first languages)?

Research Question 3:

What are the key features that influence the impact of EMI programmes in higher education?

Research Question 4:

How do the identified features specifically influence the impact of EMI programmes in higher education?

RQ 4.1: How do these features impact content learning in within-group and between-group samples?

RQ 4.2: How do these features impact English learning in within-group and between-group samples?

Method

Literature Search

To construct the dataset for our meta-analysis, the first and fourth authors collaboratively searched the literature for empirical studies on the effectiveness of EMI in higher education. First, we conducted keyword searches for relevant databases, such as ERIC (EBSCO), ProQuest Social Science Premium Collection, Scopus, and Web of Science Core Collection, using the following keyword combination of: [EMI OR CLIL] AND [English] AND [“higher education”] AND [randomized OR “control group” OR post-test OR quantitative OR cross-section OR matched OR “more effective” OR non-EMI OR non-CLIL OR “statistical difference” OR experimental]. The keyword combination for the meta-analysis was applied specifically to the abstract field of each study to ensure a precise and relevant selection of research. The first element, [EMI OR CLIL]**,** identifies studies focused on English Medium Instruction (EMI) or Content and Language Integrated Learning (CLIL). The second element, [English], confirms that the target language for EMI or CLIL is English. The third element, [“higher education”], ensures the studies are situated within higher education contexts. Finally, the fourth element, which includes terms such as [randomized, “control group,” post-test, quantitative, cross-section, matched, “more effective,” non-EMI, non-CLIL, “statistical difference,” experimental], filters for quantitative studies using experimental, comparative, or statistical methods.

Second, in addition to the database search, we conducted a backward search by reviewing the reference lists of systematic reviews and meta-analyses on this topic to identify additional empirical studies. Additionally, we carried out a forward search using Google Scholar's "Cited by" function to locate empirical studies that cited these systematic reviews and meta-analyses after their publication.

Third, we compiled the studies identified through the database, forward, and backward searches, removed duplicates, and conducted a preliminary screening by reviewing their titles and abstracts to exclude irrelevant studies before proceeding to a full review.

Lastly, we performed a comprehensive review of the remaining studies to assess their eligibility based on the following four inclusion criteria: (1) the study must be written in English, (2) it must focus on EMI contexts in higher education, (3) it must aim to evaluate the effectiveness of EMI programmes for content learning, English learning, or both, and (4) it must report statistical data necessary for calculating an effect size (i.e., Cohen's *d*) for the effectiveness of EMI programmes, such as comparing EMI students' pre-test and post-test results (within-group comparison), comparing post-test performance between EMI and non-EMI groups (between-group comparison), or both. When a study did not fully provide statistical data, we contacted its authors to request the missing data.

Dataset Construction, Calculation of Effect Sizes and Coding of Moderator Variables

After finalising the list of empirical studies to be included in the meta-analysis, the first and fourth authors collaboratively reviewed whether a study included more than one independent sample. If so, each sample was treated as a separate study.

We then calculated effect sizes in the form of Cohen's *d,* by extracting the relevant statistical data for each sample. If a sample had multiple measurements, we computed separate

effect sizes for each. For studies providing descriptive statistics, such as means, standard deviations, and sample sizes, we used these values to calculate effect sizes. When inferential statistics were provided, such as the results of ANCOVA with pre-test scores as covariates or regression coefficients controlling for other factors, we used these estimates to calculate more precise effect sizes. For these calculations, we used a web-based effect-size calculator (Wilson, 2023).

After calculating the effect sizes, the first and fourth authors collaboratively coded moderator variables to assess whether each sample provided sufficient information on key features influencing the effectiveness of EMI. (1) L1-English proximity was coded based on whether the learner's L1 and English are linguistically related, following the rationale proposed by Lee et al. (2023) and Lee and Lee (2024), using a web-based tool for quantifying genetic proximity (Beaufils, 2024). (2) Institutional support was coded based on whether language workshops, tutoring, and/or faculty training were provided. (3) EMI context was coded to distinguish between full EMI programmes, where all courses at the institution are taught in English, and partial EMI, where only selected courses are delivered in English. (4) Duration of EMI programmes was coded according to whether post-tests were conducted after short-term programmes (semester-long) or longer programmes spanning one to four years. (5) Subject matter was coded to identify whether the programme focused on STEM or non-STEM disciplines. (6) Measurement of English learning outcomes was coded based on whether the study focused on academic English skills or general English proficiency. (7) Homogeneity confirmation was coded based on whether within-group samples controlled for selection bias in EMI participants or whether between-group samples confirmed homogeneity between EMI and

non-EMI groups. (8) Entry prerequisites were coded based on whether any prerequisite courses or specific levels of English proficiency were required for enrolment.

Data Analysis: Computation of Overall Average Effect Size and Moderator Analysis

Using the calculated effect sizes and coded moderator variables, we computed the overall average effect sizes of EMI programmes for content learning and English learning in both within-group and between-group samples. We then conducted moderator analyses to assess the extent to which the identified features influenced the effectiveness of EMI programmes for content learning and English learning in both within-group and between-group samples. We used STATA version 16 as the main statistical software for these data analyses. For interpreting the magnitude of Cohen's *d*, we adhere to the widely recognized standards of 0.2, 0.5, and 0.8, representing small, medium, and large effect sizes, respectively.

First, to calculate the overall average effect sizes, we conducted a multi-level regression analysis, treating the computed effect sizes as the dependent variable and using their standard errors as the Level 1 variance, representing the variability of individual effect sizes within studies. In this model, no independent variables were included, so the intercept represents the overall average effect size. This approach is a variation of standard multi-level regression analysis, distinguished by its calculation of only the Level 2 variance, which captures variability between studies, while the Level 1 variance is predetermined by the standard errors of the effect sizes. As a result, it is referred to as a variance-known model (Hox et al., 2010). By incorporating the computed standard errors—reflecting the sample sizes of the samples—this method accounts for the hierarchical nature of the dataset (e.g., multiple effect sizes from the same sample) and weights each effect size accordingly. This ensures that more precise estimates, characterised by smaller variances or larger sample sizes, have a greater influence on the overall effect size.

When computing the overall average effect sizes, we included forest plots to provide visual cues for the reader's intuitive understanding of the distribution, magnitude, and variability of the effect sizes, as well as potential patterns in the data. Additionally, to assess potential biases, we evaluated small-study effects—commonly referred to as publication bias in meta-analyses—using funnel plots and the regression-based Egger's test in STATA software, along with outlier and influence diagnostics from the *metafor* package in R. Our analysis revealed that none of the four groups—content learning (within-group and between-group) and English learning (within-group and between-group)—exhibited significant signs of publication bias when both criteria were considered simultaneously. While some groups showed significant signs of bias according to one criterion, this was not consistently supported by the other. Due to space limitations, the results and interpretation of the publication bias analysis are included in Appendix 1.

Second, to conduct moderator analyses, we extended the multi-level regression model by including moderator variables as independent variables in the equation. In this setup, the computed effect sizes remained the dependent variable, and the standard errors of the effect size estimates continued to serve as the Level 1 variance. Moderator variables, representing key features of EMI across diverse learner backgrounds, instructional contexts, and methodological considerations, were added to the model to examine their influence on the effectiveness of EMI. In this way, the analysis identifies which factors significantly influence the effectiveness of EMI programmes and quantifies the extent of their impact. This approach also allows us to explain some of the Level 2 variance (between-study variability) by linking it to specific key features, providing deeper insights into the conditions under which EMI is most effective. In this study, while we initially had 41 samples (N = 8,747), dividing them into four groups—content learning

(within-group and between-group) and English learning (within-group and between-group)—left insufficient sample sizes for multiple regression in the moderator analysis. As a result, we conducted moderator analyses using subgroup analysis, dividing samples by categorical moderator values. This approach yields less nuanced findings compared to multiple regression (meta-regression), which can isolate moderation effects while controlling for other variables, offering more precise and reliable estimates. The complete dataset including study title, calculated effect sizes, and coded values for moderator variables, are included in Appendix 2.

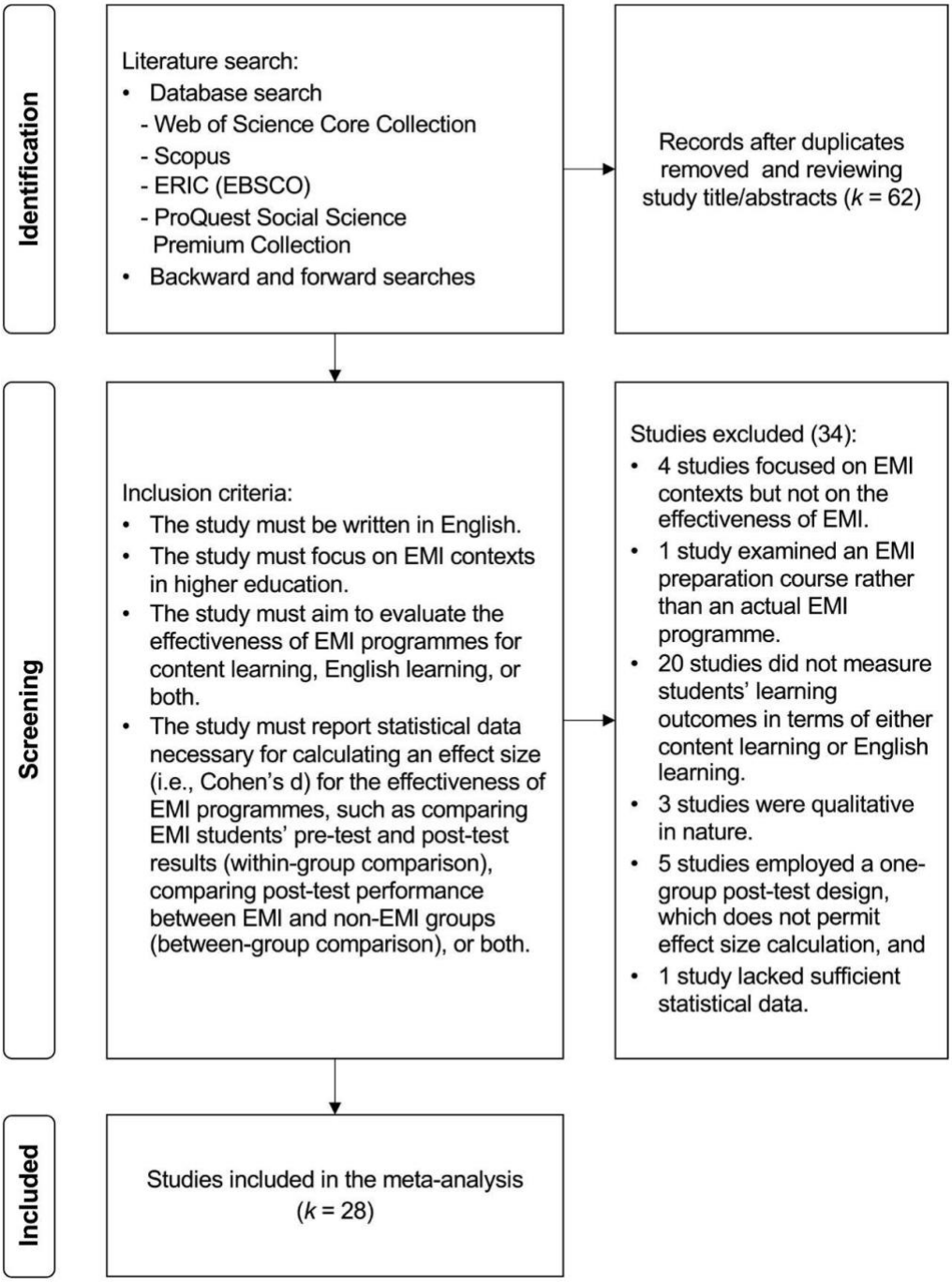
Results

Literature Search

Based on our database searches, we identified 60 studies from the *Web of Science Core Collection*, 6 from *Scopus*, 20 from *ERIC (EBSCO)*, and 20 from the *ProQuest Social Science Premium Collection*. An additional 26 studies were identified through backward and forward searches. After removing duplicates and irrelevant studies during the title and abstract screening, 62 studies remained for a full-text review. Applying our four inclusion criteria, we excluded 34 studies for the following reasons: (1) 4 studies focused on EMI contexts but not on the effectiveness of EMI, (2) 1 study examined an EMI preparation course rather than an actual EMI programme, (3) 20 studies did not measure students' learning outcomes in terms of either content learning or English learning, (4) 3 studies were qualitative in nature, (5) 5 studies employed a one-group post-test design, which does not permit effect size calculation, and (6) 1 study lacked sufficient statistical data. For the final case, we contacted the authors, who confirmed that the relevant data was unavailable. Consequently, 28 studies were included in the final dataset. The complete list of studies is in Appendix 3. Figure 1 illustrates the literature search procedure using a PRISMA flowchart.

Figure 1

PRISMA Flowchart for the Literature Search Procedure



Dataset Construction

Among the 28 studies, we identified a total of 41 independent samples, as 7 studies included multiple samples (20 in total), while the remaining 21 studies contributed one sample each. First, Aizawa et al. (2023) employed a pre-post design with a control group, featuring a between-group comparison of EMI and non-EMI groups and an additional within-group EMI analysis, contributing 2 independent samples. Second, Satayev et al. (2022b) included two EMI groups that measured both content learning and English learning. As these outcomes were treated separately, the study provided 4 independent samples. Third, Yang et al. (2019) conducted cohort comparisons between EMI and non-EMI groups for the academic years 2015 and 2016, resulting in 2 independent samples. Fourth, del Campo et al. (2016) offered cohort comparisons for EMI and non-EMI groups across three academic years (2009-2010, 2010-2011, and 2011-2012), yielding 3 independent samples. Fifth, Thi et al. (2023) examined EMI and non-EMI groups for both content learning and English learning, contributing 2 independent samples. Sixth, Guo et al. (2018) employed a pre-post design with a control group to assess both content learning and English learning. So, the study included two between-group comparison of EMI and non-EMI groups for content and English learning and had two within-group comparisons for the EMI group across pre- and post-tests for content and English learning. As content learning and English learning were treated as separate outcomes, the study contributed a total of 4 independent samples. Lastly, Salamanca & Montoya (2018) used a pre-post design with a control group, incorporating a between-group comparison of EMI and non-EMI groups and an additional within-group EMI analysis, contributing 2 independent samples. The details of the number of samples in each study and how they were coded are described in Table 1.

Table 1

Overview of 41 Independent Samples from 28 Studies

Study	Sample	N	Content Learning	English Learning	Within-Group	Between-Group	Title
1	1	39	x	o	o	x	Osam & Korun (2016)
2	2	61	o	x	o	x	Joe & Lee (2013)
3	3	194	o	x	x	o	Lin & Lei (2021)
4	4	41	x	o	o	x	Satayev et al. (2022a)[1]
5	5	58	x	o	o	x	Rogier (2012)
6	6	104	x	o	o	x	Coşgun & Hasırcı (2017)
7	7	46	o	x	o	x	Aizawa et al. (2023)[1]
	8	46	o	x	x	o	Aizawa et al. (2023)[2]
8	9	63	x	o	o	x	Aguilar & Muñoz (2014)
9	10	12	o	x	o	x	Satayev et al. (2022b)[1]
	11	12	o	x	o	x	Satayev et al. (2022b)[2]
	12	12	x	o	o	x	Satayev et al. (2022b)[3]
	13	12	x	o	o	x	Satayev et al. (2022b)[4]
10	14	29	x	o	o	x	Yang (2015)
11	15	53	x	o	o	x	Li (2018)
12	16	206	o	x	x	o	Arco-Tirado et al (2018)
13	17	256	o	x	x	o	Yang et al (2019)[1]
	18	247	o	x	x	o	Yang et al (2019)[2]
14	19	45	o	x	x	o	del Campo et al. (2016)[1]
	20	66	o	x	x	o	del Campo et al. (2016)[2]
	21	98	o	x	x	o	del Campo et al. (2016)[3]
15	22	29	x	o	o	x	Thi et al. (2023)[1]
	23	29	x	o	x	o	Thi et al. (2023)[2]
	24	78	o	x	x	o	Thi et al. (2023)[3]
16	25	498	o	x	x	o	Lin & He (2019)
17	26	39	o	x	o	x	Guo et al. (2018)[1]
	27	39	o	x	x	o	Guo et al. (2018)[2]
	28	39	x	o	o	x	Guo et al. (2018)[3]
	29	39	x	o	x	o	Guo et al. (2018)[4]
18	30	424	o	x	x	o	del Campo et al. (2023)
19	31	136	x	o	x	o	Lei & Hu (2014)
20	32	88	x	o	o	x	Yufrizal & Hasan (2017)
21	33	53	o	x	x	o	Ortín et al (2016)
22	34	4151	o	x	x	o	Civan & Coskun (2016)

23	35	21	o	x	o	x	Badrie & Abir (2018)
24	36	63	x	o	o	x	Salamanca & Montoya (2018)[1]
	37	63	x	o	x	o	Salamanca & Montoya (2018)[2]
25	38	815	o	x	x	o	Bälter et al (2024)
26	39	122	o	x	x	o	de Perea & Ramírez-García (2024)
27	40	130	o	x	o	x	Kim & Kim (2022)
28	41	191	x	o	o	x	Sato & Hemmi (2022)

Calculation of Effect Sizes

For the 41 identified samples, we calculated effect sizes (i.e., Cohen's d), resulting in a total of 51 effect sizes. Thirty-three samples contributed one effect size each, while eight samples had multiple effect sizes, yielding 18 additional effect sizes in total. Specifically, the following studies had multiple measurements: Lin and Lei (2021) measured three outcomes (i.e., assignment, participation, and final exam for content learning); Satayev et al. (2022a) measured two outcomes (i.e., listening and grammar for English learning); Aguilar and Muñoz (2014) assessed two outcomes (i.e., listening and reading for English learning); Yang (2015) examined two outcomes (i.e., listening and reading for English learning); Li (2018) measured three outcomes (i.e., vocabulary, reading comprehension, and morphological awareness for English learning); Yufrizal and Hasan (2017) assessed two outcomes (i.e., English oral performance and English competence for English learning); de Perea and Ramírez-García (2024) measured two courses (i.e., Introduction to Accounting and Intermediate Financial Accounting for content learning); and Sato and Hemmi (2022) measured two outcomes (i.e., speaking and writing for English learning). For most samples, we manually computed Cohen's *d* using the reported descriptive statistics, such as means, standard deviations, and sample sizes. However, we also used various estimates for effect size calculation, including beta coefficients from regression

analyses for Lin and Lei (2021), Arco-Tirado et al. (2018), Lei and Hu (2014), and Civan and Coskun (2016); t-test results for Yang (2015), del Campo et al. (2023), and Salamanca and Montoya (2018); and ANCOVA results for Guo et al. (2018).

Coding of Moderator Variables

The coding results for the moderator variables, addressing the third research question, are shown in Table 2. For content learning in within-group samples, the moderators "L1-English Related," "Longer than Semester," and "Entry Prerequisite" lacked values for certain categories, making data analysis for these variables infeasible. The "Academic English" variable, coded exclusively for English learning, was not applicable in this context. In contrast, for content learning in between-group samples, all seven moderators had values for each category, except for the "Academic English" variable. For English learning in within-group samples, all eight moderators had values for each category, allowing for comprehensive analysis. However, for English learning in between-group samples, the moderators "Full EMI," "Homogeneity Check," and "Entry Prerequisite" lacked values for certain categories, preventing moderator analysis for these variables. These findings highlight the variability in data availability across target domains and study designs, affecting the feasibility of certain moderator analyses.

Table 2

Number of Categorical Values for the Eight Moderator Variables

Moderator	Content Learning				English Learning			
	Within-Group (n = 7, k = 7)		Between-Group (n = 19, k = 16)		Within-Group (n = 21, k = 14)		Between-Group (n = 4, k = 4)	
	True	False	True	False	True	False	True	False
L1-English Related	0	7	8	8	2	12	1	3
Institutional Support	1	6	5	11	4	10	2	2
Full EMI	3	4	8	8	6	8	0	4

Longer than Semester	0	7	8	8	3	11	1	3
STEM Course	4	3	6	10	6	8	2	2
Academic English	NA	NA	NA	NA	2	12	2	2
Homogeneity Check	2	5	12	4	3	11	4	0
Entry Prerequisite	0	7	2	14	1	13	0	4

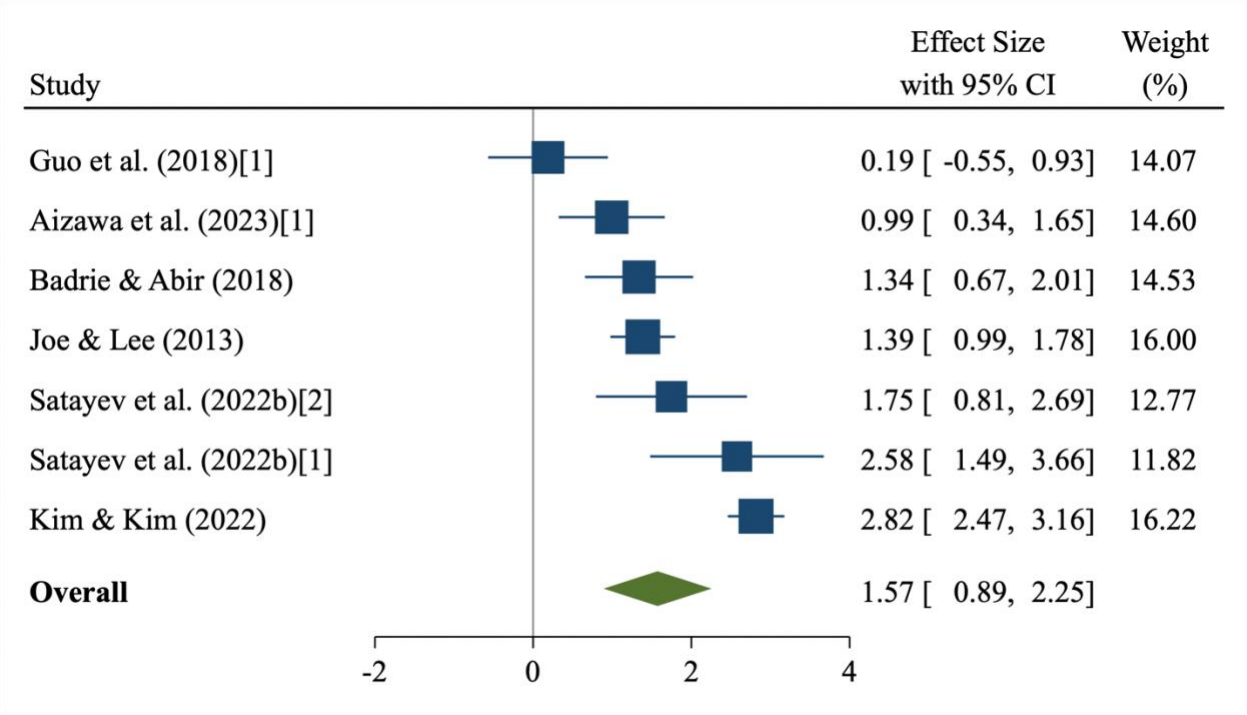
Note. The number of categorical values for the moderator variables is determined at the sample level.

Data Analysis 1: Computation of Overall Average Effect Size

To address the first and second research questions, we included forest plots (Figures 2~5) and calculated overall average effect sizes for four samples across target domains and study designs. The forest plots display effect sizes at the sample level, with measurement-level effect sizes averaged for clarity. These plots, accompanied by 95% confidence intervals, provide a visual summary of the overall trends in the effectiveness of EMI in enhancing students' content and English learning.

Figure 2

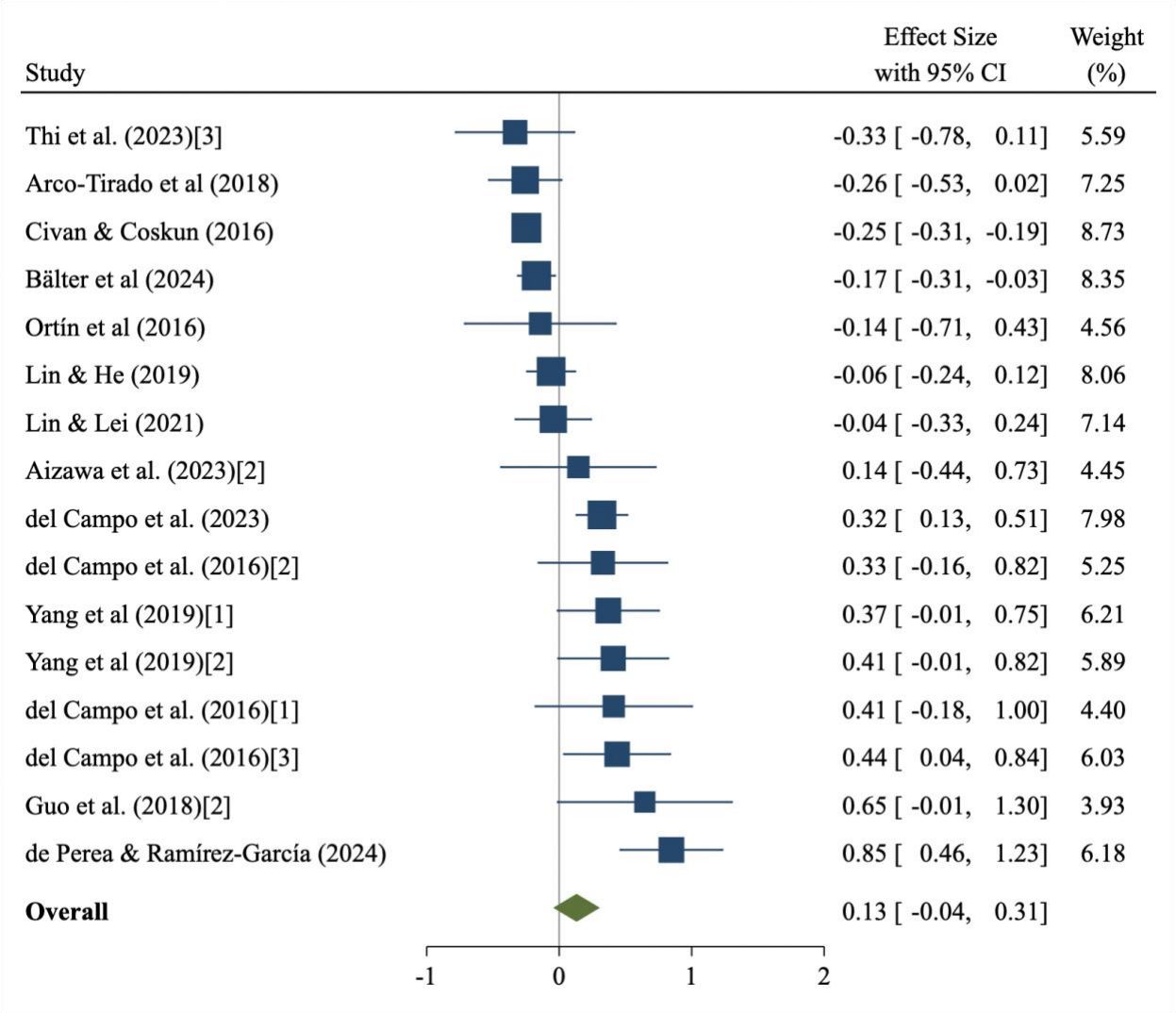
Forest Plot of Effect Sizes at the Sample Level for Content Learning in Within-Group Samples



For content learning in within-group samples (see Figure 2), all but one study, Guo et al. (2018)[1], demonstrated significant positive effect sizes. Guo et al. (2018) reported a non-significant average effect size ($d = 0.19$, $p > .05$, 95% CI [-0.55, 0.93]). Overall, the EMI programs included in this meta-analysis appear to have a statistically significant positive impact on students' content learning. The computation of the overall average effect size yielded a significant large effect size ($d = 1.57$, $p < .001$, 95% CI [0.90, 2.25]). This indicates substantial gains over time in students' content knowledge in EMI contexts. However, as within-group effect sizes do not establish causality, these gains cannot be directly attributed to EMI, as other factors may have contributed. Nonetheless, they do indicate that across studies, students enrolled in EMI programmes do appear to successfully acquire content knowledge.

Figure 3

Forest Plot of Effect Sizes at the Sample Level for Content Learning in Between-Group Samples

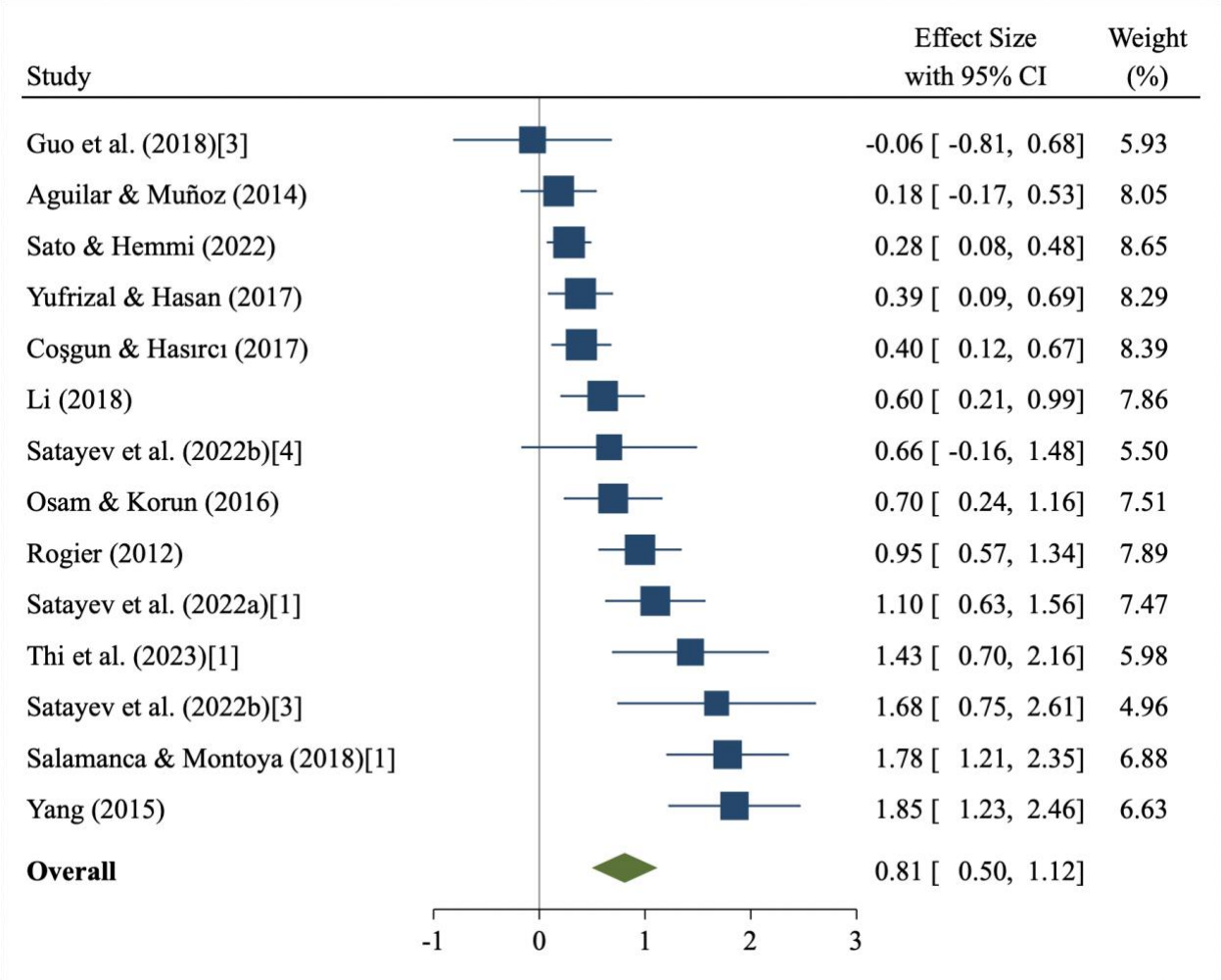


For content learning in between-group samples (see Figure 3), seven samples had negative average effect sizes, while nine samples had positive average effect sizes. However, when considering their 95% confidence intervals, only five samples demonstrated statistically significant results. Among these, two had negative average effect sizes, and three had positive average effect sizes. Overall, the EMI programs included in this meta-analysis appear to have a comparable pedagogical impact on students' content learning when compared to non-EMI

programs. The computed overall average effect size was $d = 0.13$ ($p > .05$, 95% CI [-0.04, 0.31]), indicating a non-significant marginal effect size. This suggests that, despite some variability across studies, EMI on the whole does not harm students' content learning and may allow them to perform at levels comparable to those studying in their first language (L1).

Figure 4

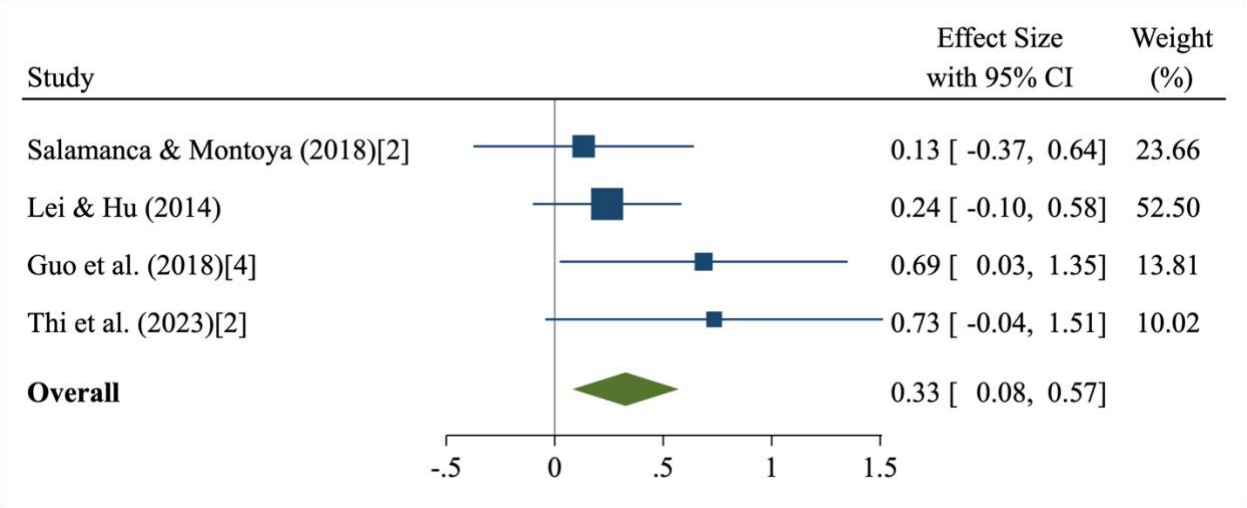
Forest Plot of Effect Sizes at the Sample Level for English Learning in Within-Group Samples



For English learning in within-group samples (see Figure 4), all but three studies—Guo et al. (2018)[3], Aguilar & Muñoz (2014), and Satayev et al. (2022b)[4]—demonstrated significant positive effect sizes, while these remaining three showed non-significant average effect sizes. Overall, the EMI programs included in this meta-analysis appear to have a statistically significant positive impact on students' English learning. The computation of the overall average effect size yielded a significant large effect size ($d = 0.81$, $p < .001$, 95% CI [0.50, 1.12]). This suggests meaningful improvements in students' English proficiency over time in EMI contexts. However, within-group designs lack causal inference, and these observed improvements may be influenced by external factors, such as increased English exposure outside of EMI instruction. Such factors are outside the remit of this meta-analysis.

Figure 5

Forest Plot of Effect Sizes at the Sample Level for English Learning in Between-Group Samples



For English learning in between-group samples (see Figure 5), all four samples demonstrated positive average effect sizes. However, three of these were not statistically significant, as their 95% confidence intervals include zero. Overall, the EMI programs analyzed in this meta-analysis appear to have a modest pedagogical advantage in enhancing students' English learning compared to non-EMI programs. The computed overall average effect size was $d = 0.33$ ($p = 0.009$, 95% CI [0.08, 0.57]), representing a significant small effect size. The limited sample size ($n = 4$, $k = 4$) weakens the reliability of this estimate, making it unsuitable for drawing generalizable conclusions about the impact of EMI on English learning.

Data Analysis 2: Analysis of Moderation Effects

To address the final research question, Table 3 presents the results of the analyses of moderation effects. For content learning, the analysis revealed that "Full EMI" had significant moderator effects in within-group samples, while "Homogeneity Check" demonstrated significant moderation effects in both within-group and between-group samples. However, no significant effects were observed for the other moderators in either within-group or between-group samples. In contrast, for English learning, none of the moderators exhibited significant moderation effects in either within-group or between-group samples.

Table 3

Results of Moderator Analysis

Moderator	Content Learning				English Learning			
	Within-Group ($n = 7$, $k = 7$)		Between-Group ($n = 19$, $k = 16$)		Within-Group ($n = 21$, $k = 14$)		Between-Group ($n = 4$, $k = 4$)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
L1-English Related	-	-	0.14	0.18	0.13	0.46	-0.30	0.36
Institutional Support	-0.22	1.04	0.21	0.19	0.44	0.33	-0.50	0.29
Full EMI	1.46**	0.42	0.11	0.18	0.10	0.33	-	-

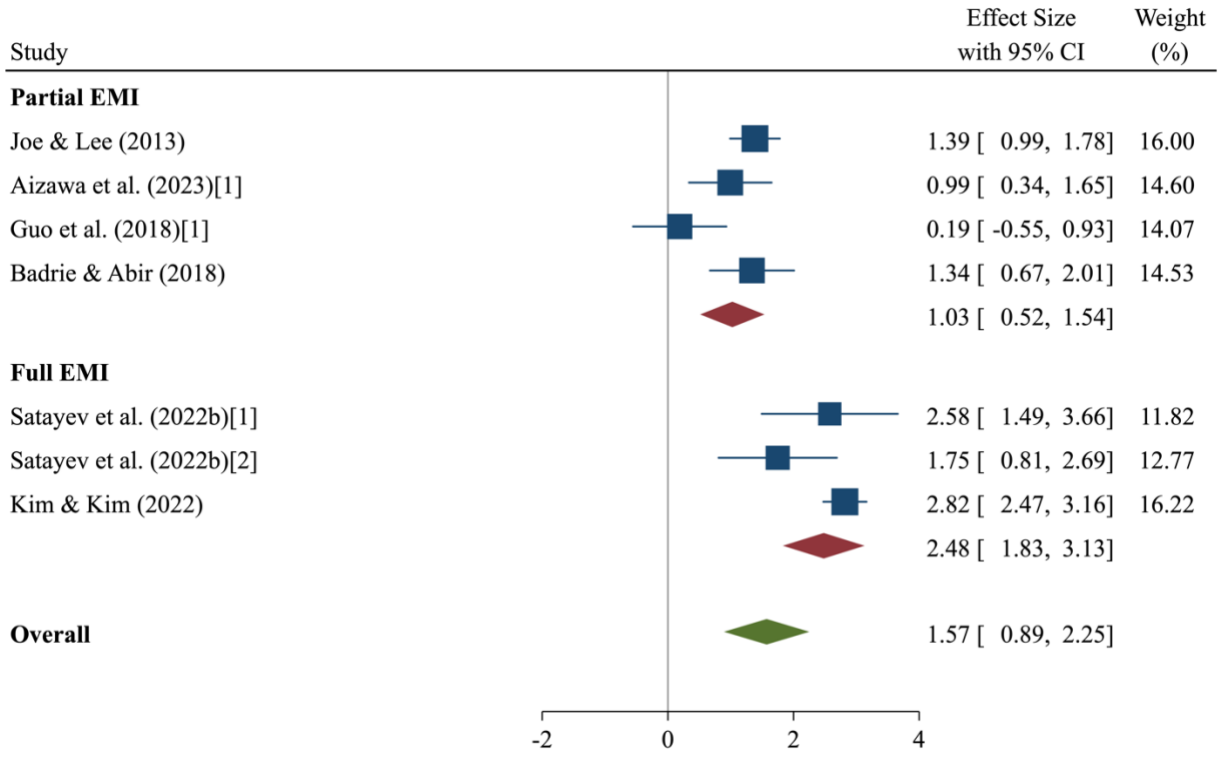
Longer than Semester	-	-	0.11	0.18	0.26	0.39	-0.21	0.34
STEM Course	0.13	0.76	-0.14	0.19	0.53	0.30	-0.04	0.35
Academic English	NA	NA	NA	NA	-0.14	0.50	0.50	0.29
Homogeneity Check	-1.36*	0.60	-0.38*	0.19	0.35	0.40	-	-
Entry Prerequisite	-	-	0.29	0.27	0.31	0.62	-	-

** $p < .01$, * $p < .05$

First, as shown in Figure 6, for content learning in within-group samples, we found a subset of studies from full EMI contexts had a significantly larger overall impact ($d = 2.48$, $p < .001$, 95% CI [1.83, 3.13]) than another subset of studies from partial EMI contexts ($d = 1.03$, $p < .001$, 95% CI [0.52, 1.54]).

Figure 6

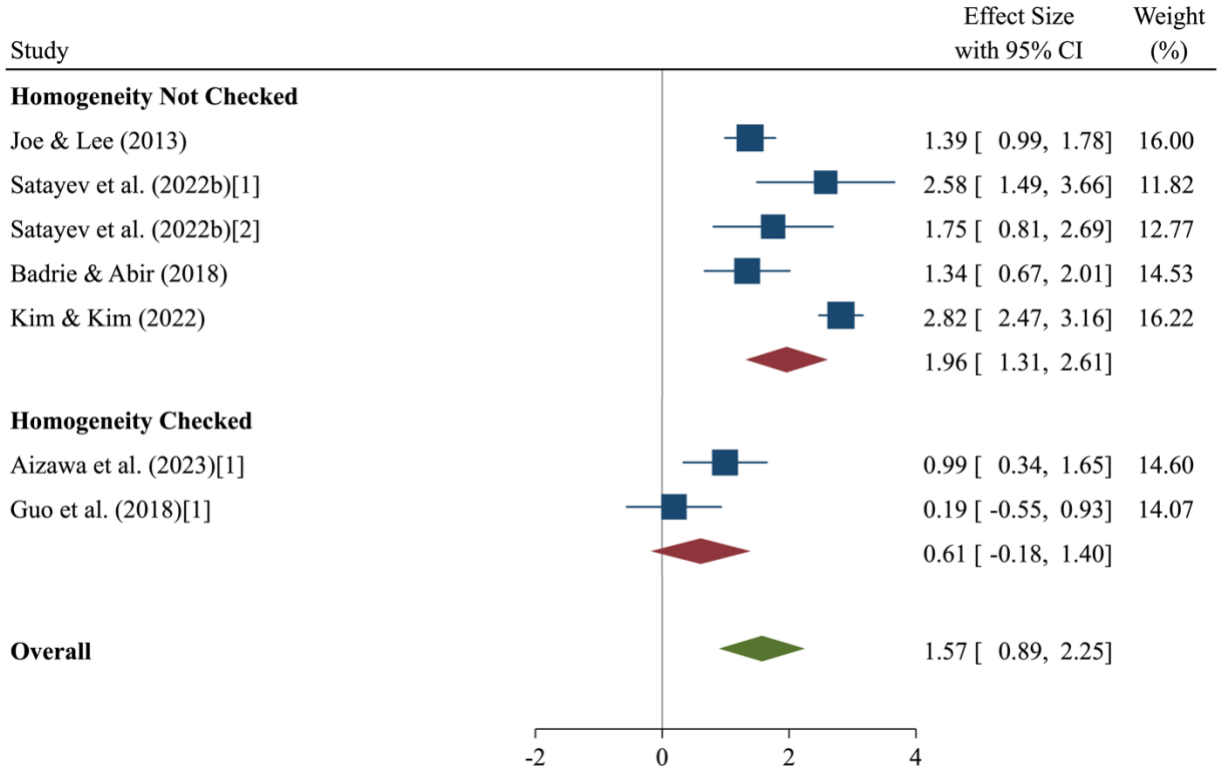
Ful EMI vs. Partial EMI for Content Learning in Within-Group Samples



Second, as shown in Figure 7, for content learning in within-group samples, we found that studies that did not control for potential selection biases—such as the possibility that more talented and motivated students self-select into EMI programmes—reported inflated effect sizes ($d = 1.96$, $p < .001$, 95% CI [1.31, 2.61]), further overestimating EMI's impact.

Figure 7

Homogeneity Checked vs. Unchecked for Content Learning in Within-Group Samples

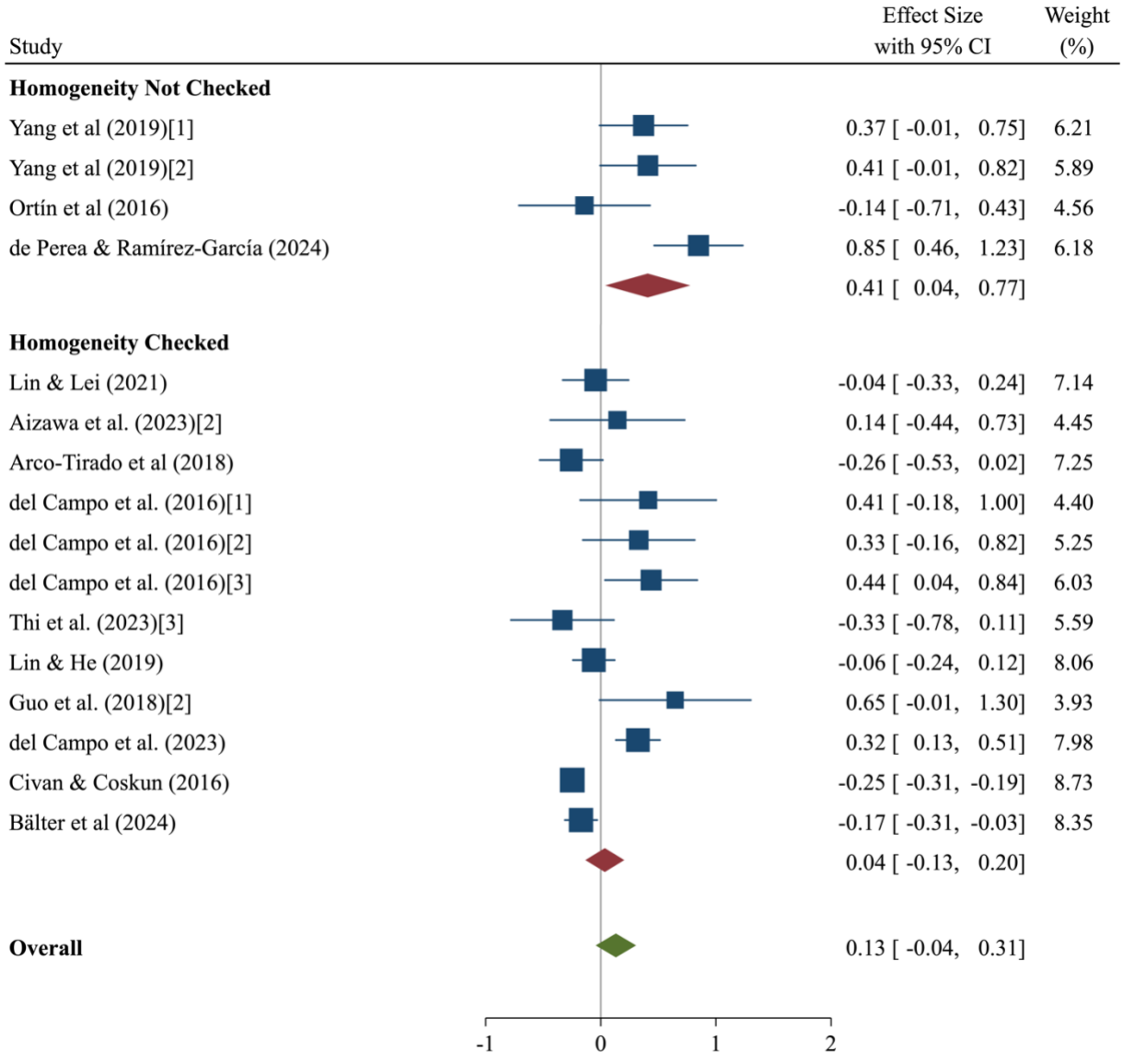


Lastly, as shown in Figure 8, for content learning in between-group samples, we found that studies that did not account for homogeneity ($d = 0.41$, $p > 0.05$, 95% CI [0.04, 0.77])

showed an apparent advantage of EMI, which may reflect pre-existing differences rather than the instructional method itself, emphasizing the need for cautious interpretation.

Figure 8

Homogeneity Checked vs. Not Chekced for Content Learning in Between-Group Samples



Discussion

Achieving Dual Goals of English Medium Instruction in Higher Education

While causal conclusions cannot be definitively drawn from within-group samples, the evidence across a range of studies suggests that EMI successfully fulfills its dual goals: primarily enhancing learners' content knowledge and secondarily fostering English skill development (Rose et al., 2021). This is evident in the large overall average effect sizes observed for both content learning ($d = 1.57$, $p < .001$; 95% CI [0.89, 2.25]) and English learning ($d = 0.81$, $p < .001$, 95% CI [0.50, 1.12]).

The substantial gains in content knowledge aligns with Cummins' framework of bilingual education (1979, 2000, 2021). Specifically, the developmental interdependence hypothesis and the threshold hypothesis propose that achieving a certain level of proficiency in both L1 and L2 is essential for students to benefit from bilingual programmes, such as EMI. As students in higher education generally possess stronger L1 and English proficiency compared to those at secondary or lower levels, the observed gains in content knowledge are consistent with these hypotheses. Furthermore, the large overall effect size is particularly encouraging, given the cognitive and linguistic demands of higher education content (Chin & Li, 2021). On the whole, the samples in this meta-analysis do not indicate that EMI exerts a cognitive load on learners that negatively impacts their overall content knowledge and linguistic development. However, this finding should be interpreted cautiously, as methodological limitations inherent in EMI studies—discussed in detail in the *Methodological Considerations for English Medium Instruction in Higher Education* section—may impact the reliability of these results.

With regard to English proficiency development, several factors likely contribute to the positive outcomes. The increased exposure to English input in EMI programmes and greater

opportunities for authentic interaction in English are key elements (e.g., Lee et al., 2023; Martínez Agudo, 2022; Muñoz, 2015). Unlike traditional EFL lessons, the input and interaction in EMI contexts are more closely tied to real-world communication, as students use English as a tool for learning subject content (Gallardo-del-Puerto et al., 2020). This authenticity in language use may further enhance students' English proficiency development.

Low Costs Concerning Content Learning and Benefits for Enhancing English Proficiency

A critical consideration in evaluating the effectiveness of EMI is balance between gains in English proficiency and potential trade-offs in content learning outcomes (Macaro, 2018). This evaluation hinges on two key hypotheses: (1) EMI does not negatively affect content learning compared to non-EMI instruction where academic subjects are taught in L2, consistent with a cost-benefit perspective (Macaro, 2022); and (2) EMI leads to superior English learning outcomes compared to its non-EMI counterpart. Addressing these hypotheses requires studies involving between-group comparisons (Cohen et al., 2018), which include non-EMI groups as controls.

The first hypothesis—that EMI does not detrimentally impact content learning—was supported, as no significant cost to academic content learning was observed. In fact, students appear to perform at comparable levels (d = 0.13, p = .14, 95% CI [-0.04, 0.31]). This finding aligns with a recent meta-analysis on EMI-CLIL effects on secondary students' content learning, which similarly reported no significant differences between EMI-CLIL and mainstream groups (Kaiypova et al., 2025). However, this finding does not suggest that the EMI conditions and non-EMI conditions are identical for students in terms of their learning experience. Previous research has shown that EMI students often must work harder in order to succeed in learning content (Aizawa et al., 2023). Other studies show EMI peers rely on the use of adaptive learning

strategies, particularly listening-related strategies, to process English input effectively (e.g., Zhou & Thompson, 2023). Additionally, intrinsic motivation and self-efficacy may enhance students' ability to deeply engage with and process content, enabling them to match the content knowledge acquisition of non-EMI groups (Zhou et al., 2023).

The second hypothesis—that EMI improves English proficiency compared to L1 instruction—was also supported, though the overall average effect size was small ($d = 0.33$, $p = .009$, 95% CI [0.01, 0.57]) and derived from a limited sample ($n = 4$, $k = 4$). This positive outcome likely stems from the extensive amount of specialised English input present in EMI programmes (e.g., Lee et al., 2023; Martínez Agudo, 2022; Muñoz, 2015), which is a notion supported in previous CLIL research (Lo & Murphy, 2014). However, the small number of between-group studies focusing on English learning ($k = 4$) limits the reliability of this conclusion. Further, research with larger samples is needed to confirm these findings.

Pedagogical Benefits of 'Full' English Medium Instruction in Higher Education

One of the significant moderators identified was the distinction between full and partial EMI contexts for content learning outcomes. From within-group samples, studies conducted in full EMI contexts demonstrated a significantly larger overall impact on learners' content learning ($d = 2.48$, $p < .001$, 95% CI [1.83, 3.13]) compared to those in partial EMI contexts (d = 1.03, p < .001, 95% CI [0.52, 1.54]). A possible explanation for this result lies in the enhanced opportunities for students in full EMI contexts to acquire both general academic English vocabulary (Coxhead, 2024) and subject-specific vocabulary, along with improved communicative competence (Jablonkai, 2021). These linguistic gains likely enable students to better comprehend academic instruction and, consequently, grasp content knowledge more effectively.

This finding highlights the potential benefits of equipping prospective EMI students with essential linguistic tools before they enter full EMI programmes. For instance, preparatory programmes at EMI institutions could focus on teaching the *Academic Spoken Word List* (Dang et al., 2017), which covers over 90% of academic speech, alongside discipline-specific vocabulary. Providing these resources may help bridge linguistic gaps and enhance students' ability to thrive in full EMI settings.

Methodological Considerations for English Medium Instruction in Higher Education

Our findings have identified a significant discrepancy across various studies suggesting a potential overestimation of EMI's impact on content learning. The observed benefits reported by some studies may partly reflect pre-existing differences among students rather than the instructional method itself. This issue arises in both between-group comparisons (e.g., EMI vs. non-EMI) and within-group analyses (e.g., EMI studies using pre-test and post-test measurements).

First, methodological flaws were evident in between-group studies that fail to assess group homogeneity. That is, studies that did not account for such homogeneity often reported an apparent advantage for EMI (mean difference = 0.38, $p < .05$, 95% CI [0.00, 0.75]). However, this advantage may primarily reflect pre-existing differences rather than the instructional method itself, underscoring the need for cautious interpretation. The recurring issue of group homogeneity in EMI research (and similar instructional approaches) has been noted extensively in the literature (e.g., Bruton, 2011a, 2011b; Graham et al., 2018; Lo & Lo, 2014). Specifically, it has been suggested that experimental groups in these studies often consist of students with higher motivation and aptitude. This pattern is consistent with findings from prior meta-analyses of secondary-level students (Kaiypova et al., 2025; Lo & Lo, 2014), demonstrating that this

methodological limitation is not confined to higher education contexts. When restricting analyses to studies that confirmed comparable baseline characteristics between EMI and non-EMI groups, the results suggest that EMI does not negatively affect content learning ($d = 0.04$, $p = .68$, 95% CI [-0.13, 0.20]). These findings, derived from more controlled samples, provide stronger and more reliable evidence that EMI enables learners to achieve performance levels comparable to instruction in their first language (L1).

Second, within-group analyses exhibit a similar pattern. Studies that failed to control for potential selection biases—such as the tendency for more talented and motivated students self-select into EMI programmes—reported inflated effect sizes (mean difference = 1.36, $p < .05$, 95% CI [0.19, 2.53]). In contrast, studies that ensured comparable characteristics among EMI participants revealed more moderate and realistic differences between pre-test and post-test outcomes ($d = 0.61$, $p > .05$, 95% CI [-0.18, 1.40]), representing a medium-sized effect. These findings further support the conclusion that EMI's impact on content learning is comparable to L1 instruction when methodological rigor is applied. However, the overall average within-group effect size for content learning—without accounting for selection bias—was notably larger ($d = 1.57$, $p < .001$; 95% CI [0.89, 2.25]). This highlights a risk of uncritically accepting evidence of EMI's effectiveness from studies that embody less rigorous designs.

Conclusion

In this study, we conducted a multi-level meta-analysis to synthesize quantitative findings on the effectiveness of EMI in enhancing college students' content and English learning. Our results indicate that EMI improves English learning and has a comparable impact on content learning relative to instruction in students' first language. These findings are supported by the methodological rigor of our meta-analytic approach, which incorporates both within- and

between-group designs. Notably, the comparable impact of EMI on content learning holds even when controlling for whether studies ensured homogeneity between EMI and non-EMI groups—a key concern in prior synthetic reviews of EMI at other educational levels (e.g., Graham et al., 2018; Kaiypova et al., 2025; Lee et al., 2023; Lo & Lo, 2014). These results further validate the view that EMI can serve as an effective pedagogical approach (Macaro et al., 2018). However, the non-comparability of EMI and non-EMI groups in terms of learner characteristics remains a significant threat to the internal validity of findings from between-subject samples across educational contexts. To address this limitation, we call for more rigorously designed empirical studies on EMI. Future research should prioritize ensuring comparability between EMI and non-EMI groups through advanced methodological approaches. For within-group designs, researchers should incorporate strategies to control for selection bias, such as accounting for the possibility that more talented, more motivated or more advantaged through social class students may self-select into EMI programs.

Some limitations of the present meta-analysis warrant attention. First, datasets for content learning and English learning were derived from different samples, as most studies assessed only one domain; only three studies evaluated both. Similarly, only four studies included both within-group and between-group designs. If more empirical studies include both content learning and English learning outcomes, as well as within-group and between-group designs, future meta-analyses could produce overall effect sizes with lower standard errors, leading to greater precision. Second, the limited number of samples constrained our ability to calculate overall average effect sizes and conduct moderator analyses effectively. For example, the small number of effect sizes for content learning in within-group designs and English learning in between-group designs may not adequately represent broader trends among learners or the complexities of

EMI implementation. Additionally, due to the limited sample size, some moderator variables lacked sufficient categorical diversity, rendering several moderators ineligible for analysis. Even when moderator analyses were conducted, inferential statistics based on only two or three effect sizes were unlikely to yield findings that could be considered generalizable.

Overall, this study provides valuable insights into the impact of EMI on content and English learning in higher education while identifying critical areas for future research. We encourage EMI researchers to address these gaps by employing improved research designs and larger, more diverse samples. Such efforts will refine our understanding of EMI's effectiveness across educational contexts and contribute to the development of more effective and pedagogically sound EMI practices.

References

Authors. (2025a).

Authors. (2025b).

Beaufils, V. (2024). *eLinguistics.net. Quantifying the genetic proximity between languages*. http://www.elinguistics.net/

Bruton, A. (2011a). Are the differences between CLIL and non-CLIL groups in Andalusia due to CLIL? A reply to Lorenzo, Casal and Moore (2010). *Applied Linguistics*, *32*(2), 236–241. https://doi.org/10.1093/applin/amr007

Bruton, A. (2011b). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, *39*(4), 523–532. https://doi.org/10.1016/j.system.2011.08.002

Chin, J. S., & Li, N. (2021). Exploring the language and pedagogical models suitable for EMI in Chinese-speaking higher education contexts. In L. I. W. Su, H. Cheung, & J. R. Wu (Eds.), *Rethinking EMI: Multidisciplinary perspectives from Chinese speaking regions* (pp. 1–20). Routledge. https://doi.org/10.4324/9780429352362-1

Cimermanová, I. (2020). Meta-analysis of studies on the acquisition of receptive skills and vocabulary in CLIL. *Journal of Language and Cultural Education*, *8*(1), 30–52. https://doi.org/10.2478/jolace-2020-0003

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th Ed.). Routledge.

Costa, F., & Mariotti, C. (Eds.). (2023). *Input in English-medium instruction*. Routledge.

Coxhead, A. (2024). Current issues and future research on teaching and learning of academic vocabulary in EMI contexts. *International Journal of TESOL Studies, 6*(2), 105–108. https://doi.org/10.58304/ijts.20240207

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251. https://doi.org/10.3102/00346543049002222

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Multilingual Matters.

Cummins, J. (2021). *Rethinking the education of multilingual learners*. Multilingual Matters.

Curle, S., Rose, H., & Yuksel, D. (2024). English medium instruction in emerging contexts: An editorial introduction to the special issue. *System*, 122, 103262. https://doi.org/10.1016/j.system.2024.103262

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning, 67*(4), 959–997. https://doi.org/10.1111/lang.12253

Fang, F. G. (2018). Review of English as a medium of instruction in Chinese universities today: Current trends and future directions: New language policies to promote multilingualism and language support for EMI will be needed in Chinese tertiary contexts. *English Today*, 34(1), 32-37. https://doi.org/10.1017/S0266078417000360

Ferrer, A., & Lin, T. B. (2024). Official bilingualism in a multilingual nation: A study of the 2030 bilingual nation policy in Taiwan. Journal of Multilingual and Multicultural Development, 45(2), 551-563. https://doi.org/10.1080/01434632.2021.1909054

Jablonkai, R. (2021). Corpus linguistic methods in EMI research: A missed opportunity? In J. Pun & S. Curle. (Eds.) *Research methods in English Medium Instruction* (pp. 92–106). Taylor & Francis.

Gallardo-del-Puerto, F., Basterrechea, M., & Martínez-Adri´an, M. (2020). Target language proficiency and reported use of compensatory strategies by young CLIL learners.

International Journal of Applied Linguistics, 30(1), 3–18. https://doi.org/10.1111/ijal.12252

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

Hox, J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Kaiypova, F., Lee, H., Lo, Y. Y., & Lee, J. H. (2025). Effects of content and language integrated learning on secondary-level students' content learning: A meta-analysis. *System*, *129*, 103580. https://doi.org/10.1016/j.system.2024.103580

Lee, H., Jung, G., & Lee, J. H. (2022). Simple view of second language reading: A meta-analytic structural equation modeling approach. *Scientific Studies of Reading*, *26*(6), 585-603. https://doi.org/10.1080/10888438.2022.2087526

Lee, H., & Lee, J. H. (2024). Extending the simple view of reading in second and foreign language learning: A meta-analytic structural equation modeling approach. *Review of Educational Research*, *94*(4), 467-500. https://doi.org/10.3102/00346543231186605

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. H., Lee, H., & Lo, Y. Y. (2023). Effects of EMI-CLIL on secondary-level students' English learning: A multilevel meta-analysis. *Studies in Second Language Learning and Teaching, 13*(2), 317–345. https://doi.org/10.14746/ssllt.38277

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., & 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

Macaro, E. (2018). *English Medium Instruction*. Oxford University Press.

Macaro, E. & Aizawa, I. (2024). Who owns English medium instruction? *Journal of Multilingual and Multicultural Development*, 45(10), 4037–4050. https://doi.org/10.1080/01434632.2022.2136187

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching, 51*(1), 36–76. https://doi.org/10.1017/S0261444817000350

Martínez Agudo, J. (2022). Do CLIL programmes help to balance out gender differences in content and language achievement? *Language, Culture and Curriculum, 35*(2), 119–133. https://doi.org/10.1080/07908318.2021.1942033

Muñoz, C. (2015). Time and timing in CLIL: A comparative approach to language gains. In M. Juan–Garau & J. Salazar–Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 87–102). Springer. https://doi.org/10.1007/978-3-319-11496-5_6

Peng, J. E., & Xie, X. (2021). English-medium instruction as a pedagogical strategy for the sustainable development of EFL learners in the Chinese context: A meta-analysis of its effectiveness. *Sustainability*, *13*(10), 5637. https://doi.org/10.3390/su13105637

Rose, H. & McKinley, J. (2018). Japan's English-medium instruction initiatives and the globalization of higher education. Higher Education, 73(1): 111–129. https://doi.org/10.1007/s10734-017-0125-1

Rose, H., Curle, S., Aizawa, I., & Thompson, G. (2020). What drives success in English medium taught courses? The interplay between language proficiency, academic skills, and motivation. *Studies in Higher Education*, *45*(11), 2149–2161. https://doi.org/10.1080/03075079.2019.1590690

Sweller, J. (2011). Cognitive load theory. *Psychology of Learning and Motivation,* 55, 37-76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

VanPatten, B. (2009). Processing matters in input enhancement. In T. Piske & M. Young-Scholten (Eds.), *Input matters in SLA* (pp. 47–61). Multilingual Matters.

Wilson, D. B. (2023). *Practical meta-analysis effect size calculator* (Version 2023.11.27). https://www.campbellcollaboration.org/calculator/

Yang, W. (2015). Content and language integrated learning next in Asia: evidence of learners' achievement in CLIL education from a Taiwan tertiary degree programme. *International Journal of Bilingual Education and Bilingualism, 18*(4), 361–382. https://doi.org/10.1080/13670050.2014.904840

Zhou, S., Fung, D., & Thomas, N. (2023). Towards deeper learning in EMI lectures: The role of English proficiency and motivation in students' deep processing of content knowledge.

Journal of Multilingual and Multicultural Development. https://doi.org/10.1080/01434632.2023.2248078

Zhou, S., & Thompson, G. (2023). A longitudinal study on students' self-regulated listening during transition to an English-medium transnational university in China. *Studies in Second Language Learning and Teaching*, *13*(2), 427–450. https://doi.org/10.14746/ssllt.38281