AUTHOR	Peter Launonen
TITLE	Examining L1 ability in a Finnish secondary education context: A comparison of CLIL and non-CLIL students' oral production
YEAR	2025
DOI	https://doi.org/10.1075/jicb.24016.lau
VERSION	Author's accepted manuscript
CITATION	Peter Launonen 2025, Examining L1 ability in a Finnish secondary education context: A comparison of CLIL and non-CLIL students' oral production. Journal of Immersion and Content-Based Language Education, https://doi.org/10.1075/jicb.24016.lau

Examining L1 ability in a Finnish secondary education context: a comparison of CLIL and non-CLIL students' oral production

Abstract

This article examines the relationship between L1 ability and CLIL in a lower secondary education context in Finland. In the study, 31 participants responded to a verbal fluency task, a picture-naming task and a survey comprising L1 self-evaluation questions. The results were analysed quantitatively in light of additional variables such as socioeconomic status and L1 grades. The CLIL students outperformed the non-CLIL students in the verbal fluency task, whereas the results of the picture-naming task were not suggestive of group-level differences. Additionally, the CLIL students' picture-naming task results were less reflective of within-group differences than those of the non-CLIL students. This study sheds light on the relationship between CLIL and L1 ability, while highlighting the possible role of CLIL in mitigating the impact of individual differences on L1 ability.

Keywords: CLIL, L1, socioeconomic status, psycholinguistics, verbal fluency, picture-naming

1 Introduction

Content and language integrated learning (CLIL) is a form of bilingual education in which a target language is used to teach content subjects (e.g., science) with the aim of developing students' language skills simultaneously (Coyle et al., 2010). Despite the term CLIL being closely associated with bilingual education in Europe (Nikula & Mård-Miettinen, 2014), there are also implementations in other regions, such as Asia (Itoi, 2024) and Oceania (Turner, 2013), that are recognised as CLIL. As a form bilingual education, CLIL promotes the development of students into multilingual speakers rather than aiming to supplant their L1 with another language, such as English (Villabona & Cenoz, 2022). The approach has been particularly popular in the European Union (EU), where there have been implementations focusing not only on English as the CLIL language but also CLIL classes whose aim is to promote the learning of other regional or national languages. There have been implementations in contexts that have been categorised as monolingual (e.g., Pérez Cañado, 2018), as well as others that are considered multilingual (e.g., San Isidro & Lasagabaster, 2019). Moreover, the use of CLIL to promote language learning in various European countries supports the EU's aims of promoting multilingualism as part of a European identity (European Commission, 1995) and as part of one's professional competences (Council of the

European Union, 2018). In secondary education specifically, using CLIL to enable the simultaneous learning of language and content is also reflective of the EU's recommendation that all citizens should be able to speak at least two languages in addition to their native language by the end of their compulsory education (European Parliament & Council of the European Union, 1998). To sum up, as an important element of the EU's current pedagogical framework, CLIL plays a role in shaping the continually evolving linguistic landscape of a multilingual EU.

In previous CLIL research, considerable emphasis has been placed on investigating L2 learning (e.g., Pérez Cañado, 2012), content learning (e.g., Fernández-Sanjurjo et al., 2019) or other benefits, such as CLIL's possible role in promoting equality (Lorenzo et al., 2021). However, several researchers have pointed out that limited attention has been given to exploring and analysing the various factors that may affect the L1 ability of students in CLIL contexts (Cenoz et al., 2014; Pérez Cañado, 2018; San Isidro & Lasagabaster, 2019). Given that the EU's multilingual objectives are, arguably, ambitious, as well as the fact that students need to forego some of their typical L1 learning and exposure in order to be educated (even partially) in a second or third language, it is worth considering whether this could have any inadvertent negative impact on students' ongoing L1 acquisition and learning. Apart from the impact of CLIL education, the role of socioeconomic status in L1 ability is also worth scrutinising, along with students' grades in L1 as a subject and their L1 self-evaluations, as they may reveal group-level discrepancies in academic performance, actual ability and perceived ability. Thus, the objective of this study is to examine the effect of CLIL on L1 ability while also considering socioeconomic status, grades in L1 as a subject, and L1 self-evaluations.

As an area of research that reflects both L1 and L2 acquisition, along with language cognition in multilinguals, the relationship between CLIL and L1 ability can be studied not only in terms of language pedagogy, but also from the perspectives of language acquisition and psycholinguistics. Additionally, given that bilingual education, such as CLIL, can play a role in helping to shape the developing bilingualism of its participants, research from the field of bilingualism is utilised to place the present study within a broader theoretical framework. In order to explore and elucidate the relationship between participation in CLIL education and L1 ability in a Finnish context, this article is guided by the following research questions:

RQ1: How do CLIL and non-CLIL students differ in terms of their L1 performance in:

a. a verbal fluency task?
b. a picture-naming task?

RQ2: To what extent are the results in RQ1 consistent with individual differences in:

a. socioeconomic status?
b. grades in L1 as a subject?
c. L1 self-evaluations?

In the following sections, the theoretical framework related to this study will be covered, after which the participants and context will be described, along with the methods that have been chosen for data collection and analysis. Next, the results emerging from the data will be presented and discussed in the context of the teaching method (i.e., CLIL or non-CLIL) and socioeconomic status, grades in L1 as a subject and L1 self-evaluations. Finally, the main conclusions and implications emanating from this study will be outlined and explored.

2 CLIL, bilingualism, and L1: Bilingual advantages and costs in CLIL students

Existing research concerning the relationship between CLIL and L1 ability has resulted in the emergence of contradictory conclusions. For example, in some studies (e.g., Merisuo-Storm, 2007; Pavón Vázquez, 2018, Rascón Moreno, 2018), CLIL students have been found to be stronger in L1 than non-CLIL students, whereas in others (e.g., Holmberg, 2019; Lim Falk, 2019), CLIL students' L1 has been found to be weaker than that of non-CLIL students. Moreover, there are also studies (e.g., Ohlsson, 2021) in which no significant difference has been found between such groups. This inconsistency appears to reflect differences in contextual factors, such as the type of CLIL implementation (e.g., total or partial CLIL), the overarching pedagogical or linguistic context (e.g., Finland, Spain, etc.), the setting (e.g., urban or rural) as well as, for instance, the possible influence of selection into CLIL (Bruton, 2011), which may also play a larger role in one context than another.

The fact that studies focusing on the relationship between CLIL and L1 ability have produced inconclusive outcomes is mirrored by the fact that research in bilingualism has also resulted in two contrasting perspectives, often construed as a bilingual advantage and a bilingual cost. This is relevant given the contribution that CLIL, as a form of bilingual education, makes to its participants' simultaneous L1 and L2 acquisition. Regarding the advantages, bilinguals

have previously been found to be stronger in various domains, such as cognitive abilities (Białecka et al., 2023), narrative skills (Haman et al., 2017), the ability to adapt flexibly to various situations (Bialystok et al., 2012) and, for example, language learning (Antoniou et al., 2015). Simultaneously, there is a body of research illustrating the potential cost that bilingualism may entail, particularly insofar as L1 ability is concerned (e.g., Faroqi-Shah et al., 2021; Haman et al., 2017; Sadat et al., 2016). Furthermore, when comparing results of individuals or groups, it is not always clear whether the results are reflective of a bilingual advantage, a cost, both or neither. It is quite possible that an individual's language ability is reflective of both an advantage and a cost simultaneously. Given that such parallel discrepancies exist in research in both CLIL and bilingualism in general, this study aims to draw from both of these fields in order to understand to what extent CLIL students' L1 Finnish ability is reflective of bilingual advantages or costs (or both or neither).

3 Ability to access and retrieve lexis in bilinguals

One area in which a bilingual cost can be investigated is in participants' ability to access and retrieve items from their mental lexicon, which can be approximated by way of a verbal fluency task (e.g., Rojczyk, 2018) and a picture-naming task (e.g., Faroqi-Shah et al., 2021). This phenomenon refers to a speaker's ability to access and retrieve lexis in response to various prompts or stimuli which could, in the case of bilingual speakers, be affected by a lower level of language use in each language or, for example, by transfer between languages (Faroqi-Shah et al., 2021). As is the case in bilingual education, previous research in this area has also produced mixed results, such as weaker L1 oral production in bilinguals (Baus et al., 2013) and no difference between bilingual and non-bilingual groups (e.g., Paap et al., 2017; Tabari, 2021). In terms of picture naming specifically, bilinguals have been found to retrieve words more slowly than monolinguals, particularly when it comes to low-frequency words (Gollan et al., 2008; Ivanova & Costa, 2008). Regarding frequency, it seems reasonable to assume that bilinguals may struggle to retrieve such words given that they probably retrieve them less often in each language than monolinguals do in their L1. Conversely, it has been argued that bilinguals may develop cognitive control mechanisms which strengthen their executive ability (Blumenfeld & Marian, 2011; Shao et al., 2014), which in turn has been found to contribute to lower picture-naming response times (RT) than monolinguals (Gollan & Goldrick, 2016). Such findings add weight to the notion that both bilingual advantages and

costs may, in some cases, be present simultaneously, even if they affect different aspects of a speaker's L1 ability.

4 Factors related to L1 ability in bilingual education

Aside from bi- or multilingualism itself, there are other factors that should be considered in order to form a broad understanding of L1 ability in CLIL students. For instance, research on L1 acquisition has consistently found higher socioeconomic status (SES) to have a positive relationship with the development of language skills (Fernald et al., 2013; Lecheile et al., 2020; Levie et al., 2017). Similarly, in CLIL research, higher SES has previously been associated with stronger results in terms of L1 ability (e.g., Peréz Cañado, 2018), but there is also evidence that its effect on L1 ability may be stronger in non-CLIL students than in CLIL students (Lorenzo et al., 2021). This means that CLIL may potentially contribute to mitigating the impact that low or high SES would otherwise have on L1 development, thereby promoting more equality. This also highlights the need to assess the suitability of bilingual education for low SES students – or for students who are otherwise considered to be at-risk – which is an area in which more research is needed in different contexts (Genesee & Fortune, 2014).

Apart from SES, differences in students' individual ability are an important consideration as they have been associated with differences in educational achievement (Deary et al., 2007), including attainment in L1 as a subject, which has been found to be shaped by both environmental factors and, for example, learners' individual academic ability (Kovas et al., 2007) and general cognitive ability (Petrill et al., 2006). In addition to SES and individual differences, examining students' self-evaluations may provide further insight into their abilities, including their L1. For instance, in a study involving over 1,000 middle-school students, Lauermann et al. (2020) found that students' self-concept and intelligence were comparable predictors of performance in L1. Such evaluations of self-concept may function as a useful metric against which actual results can be compared, as they have been found to reflect not only academic achievement but also test performance related to specific areas of linguistic ability (Arens & Jansen, 2016).

5 Methods

5.1 Participants

This study included 31 participants who were lower secondary school students from three groups: one CLIL group (n = 12) and one non-CLIL group (n = 6) from the same school, as well as one non-CLIL group from a separate school (n = 13). The non-CLIL group from a separate school (NC2) was enlisted due to the small sample size of the first non-CLIL group (NC1). Where possible, the results for the two non-CLIL groups were analysed separately because they are from schools in different regions in Finland. All students were in the ninth grade and were either 14 or 15 years old. This CLIL context is reflective of a large-scale implementation, meaning that at least 25% of the instruction is in English in all subjects other than languages (e.g. Finnish as a mother tongue, Swedish, etc.). In the present context, students can participate in CLIL from grade 1 to grade 9, after which there is an optional English-medium International Baccalaureate (IB) programme, thus allowing students to finish their studies in English, if desired. The non-CLIL groups participated in mainstream education, in which all subjects are taught in Finnish, except for language subjects, such as English and Swedish. It is also worth noting that, although exposure to and use of English in Finland varies among individuals, English-language television programmes and online content are generally not dubbed into Finnish, which undoubtedly helps promote L2 English acquisition. Additionally, English is widely used as a lingua franca with non-Finns. Therefore, CLIL's role in promoting the development of students' English skills in Finland is not entirely comparable to its role in some other CLIL contexts.

Random sampling would not have been feasible in this research, which is why the decision was made to seek suitable intact groups. Importantly, given the RQs of the present study, only participants whose L1 was deemed to be Finnish were included in the subsequent analysis and discussion. Isolating those who could meet this criterion was complicated due to many participants being multilingual as a result of belonging to immigrant communities or having lived abroad. Therefore, those who listed Finnish as their dominant language were naturally included. Also, those who are Finnish citizens, claimed to have learned Finnish first and did not report having resided in a country where another language was spoken for more than two years, were also included. However, even if a student met these requirements, s/he was also excluded if s/he was taking Finnish as a second language instead of the equivalent subject intended for L1 Finnish speakers. Students who provided only part of the data or completed one or some of the tests were also excluded. Finally, any participants who reported

having a diagnosed learning difficulty were removed in order to increase the homogeneity within and across groups.

5.2 Data collection and analysis

The data collection for this study comprised three sources: a survey, a semantic verbal fluency task (VFT) and a picture-naming task (PNT). Although the main purpose of the survey was to elicit students' self-evaluations with respect to L1 ability, it was also used to gather data on participants' linguistic backgrounds, using questions from the LEAP-Q questionnaire (Marian et al., 2007), and also SES, for which parents' educational level was taken as a proxy, as in related research (Launonen et al., 2024, Pérez Cañado, 2018). Although other components of SES could be considered as well, the relationship between family background (including parents' education) and academic achievement has been shown to persist across countries regardless of national income (Baker et al., 2002). Students' SES was recorded as the average value (one to seven) that corresponds to the level of education – ranging from compulsory education (1) to a doctoral degree (7) – for each of the participant's parents. In cases where a child only listed one parent, the SES was the value that corresponded to the level of education for that parent. The mean SES values for CLIL, NC1 and NC2 were 4.08, 4.17 and 3.85, respectively.

With respect to their linguistic backgrounds, participants were asked to list languages they knew in order of acquisition, dominance and average weekly use, along with any time in environments (e.g., a country, a home or a school) where a language other than Finnish was spoken. Students' self-evaluations of L1 for three constructs (global L1 ability, L1 speaking and L1 vocabulary) were gathered in the form of Likert statements, with three statements per construct. The three items included for the construct concerning global L1 ability (L1 global) and the three items for L1 speaking ability were from a previously validated set of statements used in research on language self-concept (Arens & Jansen, 2016). In addition, given that lexis plays a considerable role in lexical access and retrieval, three novel items specifically related to self-evaluation of L1 vocabulary were also included. For each statement, participants were asked to choose a number between one and five as per the following: 1 – I totally disagree; 2 – I disagree; 3 – I am unsure; 4 – I agree; and 5 – I totally agree. The reliability of all items for these three constructs was calculated using SPSS version 28. In order to improve reliability, all survey respondents (n = 50) were included in the reliability

calculation (see section 6.1, Table 3), regardless of whether the participants were deemed to meet the L1 criterion for this research or whether they completed all the related tasks. The purpose of analysing participants' responses to these self-assessment prompts (see Table 1) is to provide insight into whether differences in L1 self-evaluations reflect any possible differences in L1 ability [RQ2 (c)]. Group differences between self-assessment scores were analysed via a Mann-Whitney U test because the data were not found to be normally distributed.

Table 1. Self-evaluation statements of L1 ability

Construct	Self-evaluation statement (English version)
L1 global	I am good at Finnish
	I learn things quickly in Finnish
	Finnish is easy for me
L1 speaking	Talking in Finnish is easy for me
	I always find the right words even in conversations about difficult topics in Finnish
	During discussions in class, I am always able to express myself comprehensibly in Finnish
L1 vocabulary	I know a lot of Finnish words about a range of topics
	I learn new Finnish words quickly
	I have a good vocabulary in Finnish

In the survey, the researcher also asked participants for permission to obtain their grades for Finnish language and literature *(Suomen kieli ja kirjallisuus* in Finnish), and English as a subject. The grades for English were collected as confirmation of group-level differences in English proficiency in the absence of any other data on participants' English skills. The L1 Finnish grades were used to provide further insight into individual differences among students. In Finland, assessment of Finnish language and literature is based on 17 different learning outcomes from four broad criteria: functioning in interactive situations, interpreting texts, producing texts and understanding language, literature and culture (Finnish National Agency for Education, n.d.). Students are graded holistically by their teachers on the basis of their learning across these outcomes, rather than via standardised tests or, for example, end-of-year exams. Therefore, it seems that grades received for Finnish language and literature in the Finnish school system reflect a complicated amalgamation of individual characteristics. Indeed, previous research on academic achievement has led to a greater awareness of the

relationship between school grades and not only students' intelligence or cognitive abilities (e.g., Kuncel & Hezlett, 2010; Roth et al., 2015) but also their conscientiousness (e.g., Cucina et al., 2016; Friedrich & Schütz, 2023). Although grades are ultimately subject to any potential biases or human error on the part of the teacher, they can be used to shed light on individual differences between students, which is the reason for their use in the present study. In order to provide insight into RQ2 (b), group differences between L1 grades were analysed using a Mann-Whitney U test because the data were not normally distributed, while correlations between L1 grades and results related to RQ1 were also calculated.

In the VFT, participants were given 30 seconds to name as many animals as possible. The semantic task, which is based on a category, was chosen as it appears to be more closely associated with language ability than the letter task (Wauters & Marquardt, 2018). The instructions were given in Finnish and participants were asked to avoid repetitions of words. As in previous research (Baus et al., 2013), superordinate category names (e.g., bird) were included as long as subordinate examples (e.g., pigeon, hawk) were not produced. In cases where both a superordinate category name and subordinate names were produced, the former was excluded from analysis. In order to compare differences in frequency of lexis across groups, all 453 valid names of animals produced by all participants were combined to form a pool of responses, which was used to calculate the frequency values of all valid items. [For consistency, the item *haikala* was grouped under *hai* as they are alternative forms of the same concept (*shark* in English), with the latter being the standard term.] This approach was chosen as the researcher was not able to locate an existing corpus that contained all the lexical items that emerged during the task. Given that all participants produced at least ten responses, the mean frequency value was calculated for the first ten responses for all participants from each group. In addition, the mean frequency value was calculated for all remaining responses for all participants in each group. In order to shed light on RQ1 (a), independent samples *t* tests were utilised to compare not only the total number of valid responses produced among groups but also the mean frequency scores of those items. Additionally, correlation analyses were conducted to assess the relationship between results for both the VFT and the PNT and various variables of interest, namely: L1 grades, SES and self-assessments. Given that the data were not normally distributed, Spearman's rho (r_s) was utilised to determine the strength and direction of the relationships between individual variables and the results for RQ1.

In the PNT, participants were shown and asked to name 10 images from a validated set of pictures developed by Snodgrass and Vanderwart (1980), which was updated and colorised by Rossion and Pourtois (2004). The test was conducted in a quiet room at the school, where students were shown the images on a computer using DMDX software, which recorded their (audio) responses. Instructions were given in Finnish, but students were not trained on the words nor given any clues beforehand. Images were shown, consecutively, for 3000 milliseconds. Before each image, a cross was shown on the screen for 3000 milliseconds. Additional images were not included due to time restrictions inherent to the data collection. Instead of having a greater number of images and fewer participants, a preference was made for fewer images and a greater number of participants in order to analyse effects within and across groups. However, an emphasis was made to choose everyday words that adolescents would be likely to know in their L1. In addition, it should be noted that all items chosen are indeed nouns, which is a word class that has been previously associated with a bilingual cost (Faroqi-Shah et al., 2021). For all items, the word frequency, word length in syllables (word length) and age of acquisition (AoA) were ascertained and are displayed in Table 2. AoA is of interest in the present study as it may reveal a failure to acquire words that would be expected to be acquired by a certain age. The correlations between these variables and the PNT results were then examined. In cases where a student produced a synonym (e.g., *solmio* instead of *kravatti*), the response was accepted as valid. For such items, the most commonly produced valid item was the one used for the analysis of frequency, word length and AoA. Audacity version 3.3.3 was used to pinpoint the moment – to the nearest millisecond – when the participant started producing their actual response, thus ignoring any fillers or other sounds.

Table 2. Picture-based variables

Picture	Frequency	Word length	AoA
kravatti [(neck)tie]	-7.15	3	4.7
kilpikonna (turtle)	-5.64	4	3
ruuvimeisseli (screwdriver)	-6.54	5	5.9
nappi (button)	-4.51	2	3.1
joutsen (swan)	-5.43	2	3.8
luuta (broom)	-5.48	2	4
pihdit (pliers)	-6.35	2	5.8
pyykkipoika (clothes peg)	-6.78	4	3.9
haitari (accordion)	-6.26	3	5.8

rusetti [bow (tie)]	-6.37	3	2.6

To calculate frequency, all items were first located in an extensive corpus of everyday Finnish language called *Suomi24* (City Digital Group, 2021). The raw frequency value of each word was divided by the total words in the corpus and converted to a log frequency value. Regarding AoA, at the time of conducting this study, the researcher was not aware of the existence of Finnish norms for this set of pictures. [Although the images by Rossion and Pourtois (2004) have previously been validated for Finnish in a study by Torrance et al. (2018), that study concerned written picture naming instead of oral picture naming, and it did not include AoA.]. Nevertheless, research has previously found that the acquisition of words in different languages shows a considerable amount of comparability; for instance, Łuniewska et al. (2016) found such consistency across 25 languages spanning five language families. In addition, Nishimoto et al. (2005) found statistically significant correlations between data produced using Snodgrass and Vanderwart's (1980) picture set in five separate studies focusing on different languages. Therefore, in the absence of validated norms for Finnish, it was assumed that using AoA data from previous research in a different language would serve as a reasonable proxy in this study. The AoA data from Nishimoto et al. (2005) were chosen as they contain figures for all the pictures utilised in the present study. As the subsequent data were found to be non-normally distributed, Spearman's rho (r_s) was utilised to calculate correlations between variables and provide further insight into RQ1 (b).

6 Results

In this section, the results of this study are presented in subsections. First, the data on the relationship between L1 self-evaluations and academic achievement are presented. Next, group-level differences between CLIL and non-CLIL students in the VFT and PNT are analysed (RQ1). In the final subsection, the results of the correlation analyses including the VFT and PNT results along with SES, L1 grades and L1 self-evaluations (RQ2) are presented.

6.1 The relationship between L1 self-evaluations and academic achievement

As depicted in Table 3, the reliability analysis of self-evaluation statements by construct showed that both *L1 global* and the *L1 vocabulary* had strong reliability based on their corresponding Cronbach's alpha (α) scores of .81 and .82 respectively, while *L1 speaking* had moderate reliability (.65). These *L1 global* and *L1 speaking* reliability scores are comparable to those obtained in previous related research using the same sets of three statements (Arens & Jansen, 2016). As per Table 3, the mean responses for all constructs were higher in NC1 than in the CLIL group, whereas in NC2 only the *L1 global* mean scores were higher than in the CLIL group. Mann-Whitney test results indicated that none of the differences between groups were statistically significant.

Table 3. Reliability and means of L1 self-evaluations

Construct	Reliability (α)	Mean scores		
		CLIL (n = 12)	NC1 (n= 6)	NC2 (n = 13)
L1 global	.81	4.69	5.00	4.79
L1 speaking	.65	4.44	4.83	4.41
L1 vocabulary	.82	4.64	4.83	4.34

With respect to group differences on the basis of academic achievement, the means and standard deviations of participants' grades for both L1 Finnish and L2 English are displayed in Table 4. The CLIL group had the highest mean grades both for L1 and L2, along with lower standard deviations compared to both non-CLIL groups. The Mann-Whitney U tests indicated that the difference in L1 grades between CLIL and non-CLIL was significant only insofar as NC2 was concerned (p = .019), whereas for L2 grades the difference was significant with respect to both NC1 (p = .032) and NC2 (p = <.001).

Table 4. Participants' grades in L1 Finnish and L2 English

Group	L1 grade (Finnish)		L2 grade (English)	
	Mean	Standard deviation	Mean	Standard deviation
CLIL (n = 12)	8.58	0.515	9.75	0.452

NC1 (*n* = 6)	8.33	1.211	9	0.632
NC2 (*n* = 13)	7.77	0.832	7.54	1.127

6.2 Performance in the verbal fluency task

The mean total number of accepted responses for the VFT for all three groups is depicted in Table 5. Given that these data were found to be normally distributed, an independent samples *t* test was used to assess differences between the CLIL group and the two non-CLIL groups. Levene's test for equality of variances reiterated that proceeding with the independent samples *t* test was acceptable; however, given the small sample sizes, the resampling method of bootstrapping was used as it has emerged as a robust approach when analysing empirical data from small samples (Lindstromberg, 2016; Wilcox, 2012). The bootstrapped independent samples *t* tests were conducted in SPSS version 28, with the test between CLIL and NC1 being performed on 999 samples and the test between CLIL and NC2 being performed on 1,000 samples. Given the fact that multiple *t* tests have been conducted, the results have been interpreted using the Bonferroni correction, which involves dividing the significance level (.05) by the number of tests (2), resulting in a new significance level (.025). As indicated in Table 6, the non-bootstrapped results between the CLIL group and both NC1 and NC2 were found to be significant, whereas the bootstrapped results for both tests are not considered significant in light of the Bonferroni correction.

Table 5. Descriptive statistics for VFT

Group	Min	Max	Mean	Standard deviation
CLIL (*n* = 12)	13	23	16.42	2.937
NC1 (*n* = 6)	10	17	12.67	3.077
NC2 (*n* = 13)	13	18	13.85	2.267

Table 6. Analysis of group differences for VFT totals

Relationship	Levene's test	*t*	df	Two-sided *p*	Confidence intervals	
					Lower	Upper

	Equality of variances					
CLIL & NC1	.646	-2.515	16	.023	-6.910	-0.590
CLIL & NC1 (bootstrapped)	N/A	N/A	N/A	.032	-6.625	-0.643
CLIL & NC2	.619	-2,461	23	.022	-4.731	-0.410
CLIL & NC2 (bootstrapped)	N/A	N/A	N/A	.033	-4.735	-0.536

The mean frequency values for VFT responses for each group are depicted in Table 7. As indicated earlier, the frequency values were calculated separately for the following:

- the first ten responses
- the remaining responses after the first ten
- all responses combined.

The CLIL group had a lower mean than that of both non-CLIL groups across *All responses* as well as *Remaining responses*. In *First ten responses*, NC2 had a slightly lower mean than the CLIL group, whereas NC2 had the highest mean. In order to assess between-group differences, the results for the CLIL group were compared to the results from the two non-CLIL groups using the Mann-Whitney U test because these data were non-normally distributed. As indicated in Table 8, the difference between the CLIL value and the non-CLIL values was significant for *Remaining responses*. There was also a significant difference between CLIL and NC1 with respect to *All responses*.

Table 7. Mean value of frequencies of valid VFT responses

Group	First ten responses	Remaining responses	All responses
CLIL	0.0248 (n = 120)	0.0108 (n = 77)	0.0193 (n = 197)
NC1	0.0270 (n = 60)	0.0183 (n = 16)	0.0252 (n = 76)
NC2	0.0244 (n = 130)	0.0144 (n = 50)	0.0216 (n = 180)

Table 8. Analysis of group differences for VFT frequencies

Relationship	Two-sided *p*		
	First ten responses	Remaining responses	All responses
CLIL & NC1	.671	.012	.010
CLIL & NC2	.401	.022	.274

6.3 Performance in the picture-naming task

The mean PNT total for the CLIL group was lower than that of NC1 but higher than that of NC2, whereas the mean PNT RT for the CLIL group was lower than that of both non-CLIL groups. Given that neither the PNT totals nor the RT were normally distributed, Mann-Whitney U tests were performed to compare the results across the three groups. No statistically significant differences were found between groups for either the PNT totals or the RT. However, the results in Table 9 show that for both PNT totals and RT there was less variance in the CLIL group based on the standard deviation figures. Additionally, no CLIL participants produced valid responses for all ten pictures, whereas both non-CLIL groups had at least one such case.

Table 9. PNT results per group

Group	PNT totals				PNT RT			
	Mean	Min	Max	Standard deviation	Mean	Min	Max	Standard deviation
CLIL	7.5 (*n* = 12)	5	9	1.168	1.222 (*n* = 90)	0.727	2.555	0.392
NC1	7.67 (*n* = 6)	5	10	1.862	1.230 (*n* = 46)	0.710	2.758	0.474
NC2	7.46 (*n* = 13)	2	10	1.984	1.260 (*n* = 97)	0.743	2.723	0.449

6.4 Picture-based analysis of naming task results

In order to enable the inclusion of picture-based variables, the focus in the analysis of PNT results was then shifted from the participants to the pictures, whereby the picture-based mean valid responses and mean RT for all three groups were calculated and included in a correlation analysis, using Spearman's rho (r_s), together with word frequency (frequency), word length in syllables (word length) and age of acquisition (AoA).

Table 10. Correlations of picture-based results and variables

		Frequency	Word length	AoA
Frequency	r_s	.000	-.636	-.365
	p	-	.048	.300
Word length	r_s	-.636	1.000	.124
	p	.048	-	.732
AoA	r_s	-.365	.124	1.000
	p	.300	.732	-
CLIL totals	r_s	.326	.141	-.289
	p	.358	.697	.417
CLIL RT	r_s	-.479	.159	.547
	p	.162	.661	.102
NC1 totals	r_s	-.032	.627	-.373
	p	.929	.052	.289
NC1 RT	r_s	-.370	-.172	.353
	p	.293	.635	.318
NC2 totals	r_s	.097	.286	-.566
	p	.789	.423	.088
NC2 RT	r_s	-.770	.140	.316
	p	.009	.700	.374

The results, depicted in Table 10, show that the CLIL totals and RT had a moderate correlation with word frequency, whereas in the non-CLIL groups this relationship was only apparent in terms of RT. [In fact, NC2 had a strong correlation (-.770) and a significant (.009) *p* value.] Both word length and AoA correlated more strongly with NC1 totals and NC2 totals than with CLIL totals. The CLIL RT correlated strongly with AoA, whereas NC1 RT and

NC2 RT had a moderate correlation. However, none of these results was found to be statistically significant.

6.5 Correlation analysis for both tasks

Spearman's rho (r_s) correlation figures and *p* values were calculated for both tasks, indicating the direction, strength and significance of any relationships across groups. The subsequent calculations concerning the VFT totals, PNT totals and PNT RT together with the other variables of interest (SES, L1 grades and L1 self-evaluations), are depicted in Table 11. Given that NC1 self-evaluations did not contain, in some cases, enough variation to allow for calculation, coupled with the fact that NC1 only had six participants, NC1 and NC2 were combined for the purpose of this analysis. Combining these two groups does not impair the results because the teaching method (i.e., mainstream education) was the same in both groups and the results for the PNT and VFT were comparable. As indicated in Table 10, in the combined non-CLIL cohort there was a strong negative (-.593) correlation between PNT RT and PNT total, along with a significant (.008) *p* value. Similarly, for the combined non-CLIL cohort, strong correlations existed between L1 grades and both the PNT RT (-.670) and PNT total (.654) figures, with a *p* value of .002 in both cases. Regarding the relationship between SES and VFT totals, the data are suggestive of a moderate (.447) correlation for CLIL (albeit not significant) and a strong (.518) correlation for the combined non-CLIL cohort with a significant (.023) *p* value.

Table 11. Correlations of PNT and VFT

		CLIL			NC1 & NC2		
		VFT total	PNT total	PNT RT	VFT total	PNT total	PNT RT
VFT total	r_s	1.000	.250	-.071	1.000	.397	.119
	p	-	.433	.827	-	.092	.626
PNT total	r_s	.250	1.000	.172	.397	1.000	-.593
	p	.433	-	.593	.092	-	.008
PNT RT	r_s	-.071	.172	1.000	.119	-.593	1.000
	p	.827	.593	-	.626	.008	-
L1 grade	r_s	-.198	.333	-.073	.253	.654	-.670
	p	.537	.290	.821	.297	.002	.002
SES	r_s	.447	.011	.046	.518	.657	-.302

	p	.145	.972	.886	.023	.002	.209
L1 global (self-evaluation)	r_s	-.112	-.087	.067	-.147	.341	-.344
	p	.728	.788	.837	.548	.153	.149
L1 speaking (self-evaluation)	r_s	.192	-.142	.243	-.074	.120	-.299
	p	.550	.659	.446	.763	.625	.213
L1 vocabulary (self-evaluation)	r_s	-.134	.414	.271	-.155	.294	-.545
	p	.679	.180	.393	.526	.221	.016

Additionally, evidence emerged of a relationship between SES and the PNT totals for the combined non-CLIL cohort in the form of a strong (.657) correlation and a significant (.002) *p* value. The correlations involving the self-evaluations did not produce any patterns of interest, except for the fact that the three self-evaluations for the combined non-CLIL cohort all have negative correlations of varying strengths with the PNT RT figures, contrary to the corresponding results for the CLIL group, which did not show such a pattern.

7 Discussion and conclusions

This study has some limitations that should be considered before discussing the results in detail. Firstly, the small sample size invariably restricts the analysability of the data and the generalisability of the findings. For instance, regarding the PNT, there were not enough participants and pictures to allow for the total exclusion of all potentially problematic responses, such as those where a sound (e.g., a filler) was made before a response and those where the respondent self-corrected. Naturally, sample size also plays a role when it comes to comparing group-level results, such as via an independent samples *t* test or Mann-Whitney U test. Although bootstrapping has been used to mitigate this limitation in this study, it does not resolve all issues related to small sample sizes, as it may, for example, multiply the occurrences of outliers (Nikitina et al., 2019). Regarding the VFT, ideally, the researcher would have had access to a corpus containing all the items produced as this would have improved the validity of the frequency values. However, given this limitation, the researcher's creation of a pool using the participants' actual responses provides some indication of the relative frequency of items produced. With respect to individual variation among participants, although the use of the L1 grades and SES provides some useful insight into individual differences, these would ideally be supplemented by, for example, scores from a robust test of either general or verbal intelligence.

Firstly, the differences in L1 self-evaluations among groups (Table 3) reveal that the CLIL students do not appear to be at a disadvantage compared to the non-CLIL students on the basis of their own perceptions. Additionally, the CLIL students' L1 grades indicate that their academic achievement in L1 as a subject was either equally strong or stronger than that of non-CLIL students (Table 4). The significant advantage in favour of the CLIL group with respect to their English grades reflects the fact that the CLIL students are likely to have stronger English skills than the non-CLIL students, given that they participate in bilingual education.

In the VFT, the CLIL students' performance was equal to or better than the non-CLIL students' performance in this task (Table 6), which required participants to produce exemplars under the category of animals. As discussed by Shao et al. (2014), retrieving the first word in a given category seems to activate additional, semantically related words. Given that the CLIL students' ability to produce valid items was equal to or better than that of the non-CLIL students, the VFT finding in this study is suggestive of a potential bilingual advantage in the CLIL group with respect to their ability to retrieve lexis. Specifically, this could be an indication of greater executive ability in the CLIL students than in the non-CLIL students, given that this ability has previously been found to be reflected in VFT results (Shao et al., 2014). According to Blumenfeld and Marian (2011), bilinguals may develop cognitive control mechanisms that bolster their executive function. Assuming that the CLIL students' results in the VFT can be attributed to such cognitive abilities, this could also explain why the results in the study by Baus et al. (2013) did not reflect a bilingual cost in the overall VFT totals, despite such a finding emerging in the PNT RTs in that study.

Regarding the frequency of words produced, the significant differences (Table 8) observed between the CLIL group and both non-CLIL groups for the remaining responses suggest that the CLIL group's possible advantage in lexical access applied not only to the total number of words produced, but also to items with low frequency values relative to the other items produced during the task. In other words, the CLIL students were not subject to a bilingual cost in terms of word frequency, contrary to other bilingual participants in previous research in bilingualism (e.g., Baus et al., 2013; Faroqi-Shah et al., 2021). Instead, in the present study, it seems CLIL students were able to access such items more readily than non-CLIL students. Although this is indicative of an advantage to the CLIL students, the finding must be considered in light of the limitations inherent to the frequency calculation. One possible reason that the CLIL results did not reflect such a cost is that these students are not

participating in such an extensive CLIL implementation as was the case in, for example, studies conducted by Lim Falk (2019) and Holmberg (2019), where one CLIL group had comparatively very little instruction in their L1. Although the CLIL students in the present study are undoubtedly more proficient in English than the non-CLIL students – judging by their English grades, for example – if they had participated in an extensive implementation of CLIL or immersion where the use of L1 Finnish would have been minimal or non-existent, then the emergence of a bilingual cost in terms of L1 ability would, arguably, have been more likely.

The PNT totals are reflective of two possible phenomena: participants not knowing a word and participants knowing a word but not retrieving a word within the time limit. In either case, lower totals in one group would have been reflective of a linguistic cost relative to the other groups. This was not the case in the present study, which suggests that, on the surface, neither a bilingual cost nor advantage is reflected in the PNT results. This contrasts with research in bilingualism where slower RTs have been observed amongst bilinguals (Faroqi-Shah et al., 2021; Gollan & Goldrick, 2016). Moreover, given that the CLIL group had a lower standard deviation than both non-CLIL groups for the PNT totals and the RTs (Table 9), along with the fact that the CLIL group did not have any individual students who correctly named all ten pictures, it seems that the ability to retrieve lexis among the CLIL students' may have been less subject to variation than among the non-CLIL students. Additionally, the non-CLIL students' PNT totals (Table 11) correlated strongly with their RTs (r_s = .593, p = .008). This means that, among the non-CLIL students only, producing more valid responses was strongly associated with the mean speed with which the responses were provided. This could mean that the non-CLIL students are less comfortable with linguistic ambiguity than the CLIL students, who may take a more flexible approach to such ambiguity.

Regarding the picture-based results, the CLIL students' PNT totals showed some level of association with word frequency, which is suggestive of the CLIL students' ability to name pictures being more affected by word frequency than that of the non-CLIL students (Table 10). However, all groups' PNT RTs displayed some level of negative correlation with frequency, which means that students in all groups often needed more time to retrieve low-frequency words than high-frequency words. Additionally, the non-CLIL groups' PNT totals were more reflective of a positive correlation with word length than the PNT totals in the CLIL group, which could indicate that the non-CLIL students were less hindered by word length than the CLIL students. All groups' PNT totals showed some indications of a

relationship with AoA. Overall, the picture-based results did not reflect any conclusive group-level differences; however, they do provide some indications of the CLIL students performing better with short high-frequency words than with long low-frequency words.

The correlation analysis indicated that CLIL students' results in the VFT appear to be less strongly associated with SES than the results of the non-CLIL students. In addition, patterns of interest emerged in the results of the combined non-CLIL cohort with respect to L1 grades and SES. The correlations observed in Table 11 indicate that, despite there being no significant differences at a group level, the CLIL students' PNT totals were less strongly associated with individual differences in L1 grades (p = .290) and SES (p = .972) than the corresponding results in the non-CLIL cohort, whose p values were .002 in both cases. These findings are reflective of the results found in a large-scale study (Lorenzo et al., 2021) on CLIL and SES in Spain, in which CLIL was found to mitigate the effect of SES on L1 ability. In that study, the groups of students with the highest and lowest mean L1 ability were the non-CLIL students with the highest and lowest SES, respectively, whereas the CLIL students' L1 ability was less responsive to changes in SES (in either direction) than the non-CLIL students' L1 ability. This suggests that CLIL may, in mitigating the impact of SES, have a homogenising effect on L1 ability across groups. Furthermore, the fact that the CLIL students' results are not as reflective of differences in L1 grades as the non-CLIL students' results adds further weight to the notion that CLIL participation may help to homogenise L1 ability in spite of individual differences. Although L1 self-evaluations were also included in the correlation analyses, they did not lead to additional insights regarding individual differences in task performance between CLIL and non-CLIL students.

The results of the present study seem to indicate that the L1 ability of these CLIL students is not subject to a bilingual cost, contrasting with the results found in previous research in bilingualism (Faroqi-Shah et al., 2021; Haman et al., 2017; Sadat et al., 2016). This suggests that CLIL, as a form of bilingual education, can be considered a viable alternative for parents considering traditional methods of bilingualism, such as the one parent, one language approach. In addition, if further evidence emerges of CLIL's possible role in mitigating the effects of SES, this may have considerable ramifications for education policymakers who aim to promote equality in education.

In conclusion, this study aimed to elucidate group-level differences between CLIL and non-CLIL students in a Finnish context. The methods utilised include a verbal fluency task (VFT),

a picture-naming task (PNT) and a survey which captured background information, self-evaluations of L1 ability, socioeconomic status (SES) and grades in L1 as a subject (L1 grades). In this study, the CLIL students' performance was equal to or better than that of the non-CLIL students in the VFT, but the PNT totals were not reflective of any group-level differences. However, when examining additional variables, the CLIL students' results seemed less responsive to changes in individual differences in SES and L1 grades than those of the non-CLIL students. Therefore, the PNT results lend weight to the notion that CLIL could play a role in promoting equality by mitigating the impact of, for example, SES. This suggests that the pedagogical challenges teachers face in managing a socioeconomically diverse classroom may differ between CLIL and non-CLIL groups. Moreover, given that CLIL teaching naturally implies that students receive less content-specific instruction in their L1 than students not in CLIL, the fact that CLIL students would outperform non-CLIL students in any L1-related task seems to be reflective of a bilingual advantage rather than a cost. However, explaining its overall effect on L1 ability also requires an exploration of individual differences related to ability and background, such as SES. Future research in this area holds the potential to inform educational policy and practices with a view to maximising the effectiveness of bilingual education in achieving its outcomes in the Finnish context.

Author

Peter Launonen

ptlaun@utu.fi

Rehtorinpellonkatu 3. 20500. Turku, Finland.

References

Antoniou, M., Liang, E., Ettlinger, M., & Wong, P. C. M. (2015). The bilingual advantage in phonetic learning. *Bilingualism (Cambridge, England), 18*(4), 683–695. https://doi.org/10.1017/S1366728914000777

Arens, A. K., & Jansen, M. (2016). Self-concepts in reading, writing, listening, and speaking: A multidimensional and hierarchical structure and its generalizability across native and foreign languages. *Journal of Educational Psychology, 108*(5), 646–664. https://doi.org/10.1037/edu0000081

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic Status, School Quality, and National Economic Development: A Cross-National Analysis of the "Heyneman-Loxley Effect" on Mathematics and Science Achievement. *Comparative Education Review*, *46*(3), 291–312. https://doi.org/10.1086/341159

Baus, C., Costa, A., & Carreiras, M. (2013). On the effects of second language immersion on first language production. *Acta Psychologica*, *142*(3), 402–409. https://doi.org/10.1016/j.actpsy.2013.01.010

Białecka, M., Wodniecka, Z., Muszyńska, K., Szpak, M., & Haman, E. (2023). Both L1 and L2 proficiency impact ToM reasoning in children aged 4 to 6. Painting a more nuanced picture of the relation between bilingualism and ToM. *Bilingualism (Cambridge, England)*, 1–19. https://doi.org/10.1017/S1366728923000652

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240–250. https://doi.org/10.1016/j.tics.2012.03.001

Blumenfeld, H. K., & Marian, V. (2011). Bilingualism influences inhibitory control in auditory comprehension. *Cognition*, *118*(2), 245–257. https://doi.org/10.1016/j.cognition.2010.10.012

Bruton, A. (2011). 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

Cenoz, J., Genesee, F., & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. *Applied Linguistics*, *35*(3), 243–262, https://doi.org/10.1093/applin/amt011

City Digital Group (2021). *Suomi24 virkkeet -korpus 2001-2020, Korp-versio [korpus]*. Kielipankki. Retrieved from: http://urn.fi/urn:nbn:fi:lb-2021101525

Council of the European Union (2018). *Council recommendation of 22 May 2018 on key competences for lifelong learning* (Text with EEA relevance) (2018/C 189/01). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:C:2018:189:TOC

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

Cucina, J. M., Peyton, S. T., Su, C., & Byle, K. A. (2016). Role of mental abilities and mental tests in explaining high-school grades. *Intelligence (Norwood)*, *54*, 90–104. https://doi.org/10.1016/j.intell.2015.11.007

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence (Norwood)*, *35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

European Commission (1995). *White paper on education and training: Teaching and learning - towards the learning society*. Retrieved from https://europa.eu/documents/comm/white_papers/pdf/com95_590_en.pdf

European Parliament & Council of the European Union (1998). *Council resolution of 20 November 1997 on a European year of languages (1998)*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31998Y0103(01)

Faroqi-Shah, Y., Kevas, Y., & Li, R. (2021). Lexical category differences in bilingual picture naming: Implications for models of lexical representation. *Bilingualism (Cambridge, England), 24*(5), 849–863. https://doi.org/10.1017/S1366728921000213

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*(2), 234–248. https://doi.org/10.1111/desc.12019

Fernández-Sanjurjo, J., Fernández-Costales, A., & Arias Blanco, J. M. (2019). Analysing students' content-learning in science in CLIL vs. non-CLIL programmes: Empirical evidence from Spain. *International Journal of Bilingual Education and Bilingualism, 22*(6), 661–674. https://doi.org/10.1080/13670050.2017.1294142

Finnish National Agency for Education. (n.d.). *Evaluation of students' learning and competence and the criteria for final assessment.* https://www.oph.fi/sites/default/files/documents/Perusopetuksen%20päättöarvioinnin%20kriteerit%2C%20AI.pdf

Friedrich, T. S., & Schütz, A. (2023). Predicting school grades: Can conscientiousness compensate for intelligence? *Journal of Intelligence, 11*(7), 146-. https://doi.org/10.3390/jintelligence11070146

Genesee, F., & Fortune, T. (2014). Bilingual education and at-risk students. *Journal of Immersion and Content-Based Language Education, 2*(2), 196-209. https://doi.org/10.1075/jicb.2.2.03gen

Gollan, T. H., & Goldrick, M. (2016). Grammatical constraints on language switching: Language control is not just executive control. *Journal of Memory and Language*, *90*, 177–199. https://doi.org/10.1016/j.jml.2016.04.002

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787–814. https://doi.org/10.1016/j.jml.2007.07.001

Haman, E., Wodniecka, Z., Marecka, M., Szewczyk, J., Białecka-Pikul, M., Otwinowska, A., … Foryś-Nogala, M. (2017). How does L1 and L2 exposure impact L1 performance in bilingual children? Evidence from Polish-English migrants to the United Kingdom. *Frontiers in Psychology*, *8*, 1444–1444. https://doi.org/10.3389/fpsyg.2017.01444

Holmberg, P. (2019). The development of academic vocabulary in Swedish. In L. Sylvén (Ed.), *Investigating Content and Language Integrated Learning: Insights from Swedish High Schools* (pp. 173-186). Blue Ridge Summit, PA: Multilingual Matters.

Itoi, K. (2024). Fostering Inclusive Learning and 21st-Century Skills: Creating Translanguaging Spaces in University Content and Language Integrated Learning Courses. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.12643

Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica*, *127*(2), 277–288. https://doi.org/10.1016/j.actpsy.2007.06.003

Kovas, Y., Haworth, C. M. A., Dale, P. S., & Plomin, R. (2007). The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monographs of the Society for Research in Child Development*, *72*(3), 1–144.

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, *19*(6), 339–345. https://doi.org/10.1177/0963721410389459

Launonen, P., Roiha, A., & Maijala, M. (2024). Exploring the Relationship between CLIL and L1 Ability in Finland: Analyzing Written and Oral Production. *Latin American Journal of Content & Language Integrated Learning*, *15*(2), 1–30. https://doi.org/10.5294/laclil.2022.15.2.8

Lauermann, F., Meißner, A., & Steinmayr, R. (2020). Relative importance of intelligence and ability self-concept in predicting test performance and school grades in the math and language arts domains. *Journal of Educational Psychology*, *112*(2), 364–383. https://doi.org/10.1037/edu0000377

Lecheile, B. M., Spinrad, T. L., Xu, X., Lopez, J., & Eisenberg, N. (2020). Longitudinal relations among household chaos, SES, and effortful control in the prediction of language skills in early childhood. *Developmental Psychology*, *56*(4), 727–738. https://doi.org/10.1037/dev0000896

Levie, R., Ben-Zvi, G., & Ravid, D. (2017). Morpho-lexical development in language impaired and typically developing Hebrew-speaking children from two SES backgrounds. *Reading & Writing*, *30*(5), 1035–1064. https://doi.org/10.1007/s11145-016-9711-3

Lim Falk, M. (2019). The development of linguistic correctness in CLIL and non-CLIL students' writing in the L1 at upper secondary school. In L. Sylvén (Ed.), *Investigating Content and Language Integrated Learning : Insights from Swedish High Schools* (pp. 187-215). Blue Ridge Summit, PA: Multilingual Matters.

Lindstromberg, S. (2016). Inferential statistics in language teaching research: A review and ways forward. *Language Teaching Research: LTR*, *20*(6), 741–768. https://doi.org/10.1177/1362168816649979

Lorenzo, F., Granados, A., & Rico, N. (2021). Equity in bilingual education: Socioeconomic status and content and language integrated learning in monolingual southern Europe. *Applied Linguistics*, *42*(3), 393–413. https://doi.org/10.1093/applin/amaa037

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Anđelković, D., … Ünal-Logacev, Ö. (2016). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, *48*(3), 1154–1177. https://doi.org/10.3758/s13428-015-0636-6

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and

multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Merisuo-Storm, T. (2007). Pupils' attitudes towards foreign-language learning and the development of literacy skills in bilingual education. *Teaching and Teacher Education, 23*(2), 226–235. https://doi.org/10.1016/j.tate.2006.04.024

Nikitina, L., Paidi, R., & Furuoka, F. (2019). Using bootstrapped quantile regression analysis for small sample research in applied linguistics: Some methodological considerations. *PloS One, 14*(1), e0210668–e0210668. https://doi.org/10.1371/journal.pone.0210668

Nikula, T., & Mård-Miettinen, K. (2014). Language learning in immersion and CLIL classrooms. In J.-O. Östman and J. Verschueren (Eds.) *Handbook of pragmatics, 18*. 2014 Installment. Amsterdam: John Benjamins. https://doi.org/10.1075/hop.18.lan10

Nishimoto, T., Miyawaki, K., Ueda, T., Une, Y., & Takahashi, M. (2005). Japanese normative set of 359 pictures. *Behavior Research Methods, 37*(3), 398–416. https://doi.org/10.3758/BF03192709

Ohlsson, E. (2021). Perspectives on written L1 in Swedish CLIL education. *Apples (Jyväskylä, Finland), 15*(2). https://doi.org/10.47862/apples.98178

Paap, K. R., Myuz, H. A., Anders, R. T., Bockelman, M. F., Mikulinsky, R., & Sawi, O. M. (2017). No compelling evidence for a bilingual advantage in switching or that frequent language switching reduces switch cost. *Journal of Cognitive Psychology (Hove, England), 29*(2), 89–112. https://doi.org/10.1080/20445911.2016.1248436

Pavón Vázquez, V. (2018). Learning outcomes in CLIL programmes: A comparison of results between urban and rural environments. *Porta Linguarum: Revista Internacional de Didáctica de Las Lenguas Extranjeras, 29*, 9–28. https://doi.org/10.30827/Digibug.54020

Pérez Cañado, M. (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. (2018). The effects of CLIL on L1 and content learning: Updated empirical evidence from monolingual contexts. *Learning and Instruction, 57*, 18–33. https://doi.org/10.1016/j.learninstruc.2017.12.002

Petrill, S. A., Deater-Deckard, K., Thompson, L. A., De Thorne, L. S., & Schatschneider, C. (2006). Reading skills in early readers: Genetic and shared environmental Influences. *Journal of Learning Disabilities, 39*(1), 48–55. https://doi.org/10.1177/00222194060390010501

Rascón Moreno, D. (2018). Socioeconomic Status and its Impact on Language and Content Attainment in CLIL Contexts. *Porta Linguarum Revista Interuniversitaria de Didáctica de Las Lenguas Extranjeras*, 29, 115–135. https://doi.org/10.30827/Digibug.54025

Rojczyk, A. (2018). Time-limited verbal fluency task with Polish−English unbalanced bilinguals. In *Individual Learner Differences in SLA* (pp. 210–225). Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781847694355-015

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception (London)*, *33*(2), 217–236. https://doi.org/10.1068/p5117

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence (Norwood)*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Sadat, J., Martin, C. D., Magnuson, J. S., Alario, F., & Costa, A. (2016). Breaking down the bilingual cost in speech production. *Cognitive Science*, *40*(8), 1911–1940. https://doi.org/10.1111/cogs.12315

San Isidro, X., & Lasagabaster, D. (2019). The impact of CLIL on pluriliteracy development and content learning in a rural multilingual setting: A longitudinal study. *Language Teaching Research*, *23*(5), 584–602. https://doi.org/10.1177/1362168817754103

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*, 772–772. https://doi.org/10.3389/fpsyg.2014.00772

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174. https://doi.org/10.3758/s13428-013-0376-4

Tabari, F. (2021). Lexicon on board: A MEG study based on expressive picture-naming. *East European Journal of Psycholinguistics*, *8*(2), 233–254. https://doi.org/10.29038/EEJPL.2021.8.2.TAB

Torrance, M., Nottbusch, G., Alves, R. A., Arfé, B., Chanquoy, L., Chukharev-Hudilainen, E., … Wengelin, Å. (2018). Timed written picture naming in 14 European languages. *Behavior Research Methods*, *50*(2), 744–758. https://doi.org/10.3758/s13428-017-0902-x

Turner, M. (2013). Content-based Japanese Language Teaching in Australian Schools: Is CLIL a Good Fit? *Japanese Studies*, *33*(3), 315–330. https://doi.org/10.1080/10371397.2013.846211

Villabona, N., & Cenoz, J. (2022). The integration of content and language in CLIL: a challenge for content-driven and language-driven teachers. *Language, Culture, and Curriculum*, *35*(1), 36–50. https://doi.org/10.1080/07908318.2021.1910703

Wauters, L., & Marquardt, T. P. (2018). Category, letter, and emotional verbal fluency in Spanish–English bilingual speakers: A preliminary report. *Archives of Clinical Neuropsychology*, *33*(4), 444–457. https://doi.org/10.1093/arclin/acx063

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd ed.). St. Louis: Elsevier Science & Technology. https://doi.org/10.1016/C2010-0-67044-1