This postprint was originally published by Elsevier as:
Feddermann, M., Baumert, J., & Möller, J. (2022). Just selection and preparation? CLIL effects on second language learning. *Learning and Instruction*, *80*, Article 101578. https://doi.org/10.1016/j.learninstruc.2021.101578

Just selection and preparation? CLIL effects on second language learning

Maja Feddermann [a,*], Jürgen Baumert [b], Jens Möller [a]

[a] Institute for Psychology of Learning and Instruction, Kiel University, Olshausenstraße 75, 24118, Kiel, Germany
[b] Max-Planck-Institute for Human Development, Lentzeallee 94, 14196, Berlin, Germany
[*] Corresponding author.
E-mail addresses: mfeddermann@ipl.uni-kiel.de (M. Feddermann), jmoeller@ipl.uni-kiel.de (J. Baumert), jmpobaumert@mpib-berlin.mpg.de (J. Möller).

ABSTRACT

Since most studies failed to account for selection and preparation effects, previous studies overestimated the positive effects of Content and Language Integrated Learning (CLIL) on the development of students' receptive foreign language skills. We examined the English listening and reading comprehension development of $N = 448$ German CLIL and $N = 4{,}191$ non-CLIL grammar school students from grade five to grade nine using full survey data from a 2014–2019 cohort. At the beginning of the study, students were on average $M = 10.38$ ($SD = 0.54$) years old. Prior achievement and sociodemographic variables showed significant selection effects. After propensity score matching, data indicated significant preparation effects of additional English lessons. However, when both selection and preparation effects were controlled, no significant additive CLIL effect showed up. We discuss the results in light of earlier contradictory findings and recommend considering selection and preparation effects when studying CLIL effects.

Keywords
Content and language integrated learning (CLIL), English as a foreign language, Second language learning, Skill development, Propensity score matching

Supplementary material to this article is available. For more information see https://hdl.handle.net/21.11116/0000-000A-CAEA-B.

1. Introduction

One way to learn a foreign language is content-based instruction in Content and Language Integrated Learning (CLIL) programs (Genesee, 2014). In CLIL, one or more content subjects are taught in a foreign language, while the common language is usually used as the language of instruction in the other subjects. CLIL was implemented due to high-level policymaking and grass-roots activities, which were often backed by teachers and parents (Dalton-Puffer, 2011). CLIL contributes to the language learning goals of the European Union, including access to lifelong language learning, improving language education, and creating a language-friendly environment. The European Union wants its citizens to learn other European languages to achieve these objectives (European Commission, 2004). CLIL programs expanded considerably in the context of educational initiatives since the 1990s (Rumlich, 2018). However, this advancement is accompanied by substantial gaps in empirical studies on the effects of CLIL (Goris, Denessen, & Verhoeven, 2019).

Several studies with convenience samples found positive effects of CLIL on language competencies (e.g., Dallinger, Jonkmann, Holm, & Fiege, 2016; Lasagabaster, 2008; Ruiz de Zarobe, 2008). However, often the empirical studies had some shortcomings. For example, they failed to control for selection and preparation effects. Selection effects arise when students choose a CLIL program themselves or are chosen by schools. For example, CLIL students had better prior knowledge, higher cognitive skills, and a more favorable family background in a study by Dallinger et al. (2016). Preparation effects were caused by an increase in English language lessons for future CLIL students to prepare for the upcoming CLIL instruction (Rumlich, 2018). Failure to recognize both effects can lead to inaccurate results (Goris et al., 2019). Our literature review revealed that the selection, preparation, and CLIL-effects have only been assessed in two previous studies (Feddermann, Baumert, & Möller. under review; Feddermann, Möller, & Baumert, 2021) using longitudinal data from large and comprehensive samples. The current research applies a similar approach to assess CLIL and non-CLIL students' English skills, but deals with more recent data from grade five to grade nine.

2. Content and Language Integrated Learning (CLIL)

The subject content of a CLIL class is taught in a foreign language (mostly English), while the curriculum is maintained. The integration of content and foreign language learning is reflected in the 4Cs Framework (Coyle, Hood, & Marsh, 2010). Here, content and foreign language learning are integrated into the respective context. This holistic model serves the didactic modelling of CLIL and comprises four central and contextualized elements: Content (subject matter), Communication (language learning and using), Cognition (learning and thinking

processes) and Culture (developing intercultural understanding and global citizenship). According to Coyle et al. (2010), there is a symbiotic relationship between content and foreign language learning. From a didactic perspective, it is advantageous that the foreign language (L2) can be used in an authentic context without creating a separate lesson (Möller et al., 2017). CLIL promises to improve L2 competencies for several reasons. At present, interactionist and sociocultural approaches to foreign language acquisition are current. These are based on the assumption that the foreign language input is crucial for the interactive negotiation and output in language production (Königs, 2010). The Input Hypothesis (Krashen, 1985) can be assigned to these approaches and indicates that language is acquired by the presence of comprehensible input and the possibility to use the foreign language and receive feedback for it. CLIL programs provide comprehensible input and the opportunity to voluntarily use the foreign language (Dallinger et al., 2016). According to Dalton-Puffer (2008), CLIL contributes to the reading comprehension skills, which are investigated in the present study, by offering additional reasons for reading. The Natural Approach (Krashen & Terrell, 1998) states that foreign language competence is best achieved by communicating in the foreign language under conditions that resemble native language acquisition. These conditions are met in CLIL programs: They focus on meaning and not on form (Dalton-Puffer, 2008). Further, students are encouraged but not forced to use the foreign language (Dallinger et al., 2016). At least, topics in the CLIL classes are determined by the curriculum of the content subject whereas the foreign language simply operates as a means of communication, leading to a high level of authenticity in language use (Surmont, van de Craen, Struys, & Somers, 2014).

2.1. CLIL in Europe

In the 1950s and 1960s, CLIL started with a few innovative projects (Eurydice, 2006). Until 1970, the need to design language- and content-integrated programs was a natural consequence of various geographical, demographic and economic aspects. The teaching of individual subjects in a foreign, minority, or regional language was mainly used in certain language regions (e.g., near national borders or in large cities). The aim was to provide children in these regions with the necessary language skills for communication with the locals (Hanesová, 2015). In Europe, CLIL programs in a foreign language became available in the 1980s/1990s (Eurydice, 2006). In 1978, the European Commission already issued a proposal to promote the teaching of more than one language in schools (European Commission, 1978). In 1983, the European Parliament asked the European Commission to develop a new program to improve foreign language skills (European Parliament, 1983). Consequently, the number of schools in Europe that taught some subjects in a foreign language increased steadily. Due to the development of different teaching methods and different historical, sociological and educational factors within each region, various types of integrated approaches to foreign language teaching emerged, including CLIL (Hanesová, 2015). The term CLIL was coined in 1994 by David Marsh (Marsh, 1994). During the 1990s, CLIL became the most widely used term for integrated content and foreign language teaching. Politics, and especially education policy, promoted the introduction of CLIL in Europe (Hanesová, 2015). According to the European Commission, everyone should have the opportunity to acquire and maintain communication skills in at least two languages in addition to their first language. However, there are major differences between countries: In Sweden, for example, almost 80%, in the Netherlands 69%, and in Germany 62% learn two or more foreign languages at school. In Spain, on the other hand, only 30% and in Italy only 25% do so (European Commission, 1995; Fäcke, 2021). One idea to realize this goal is to use the first foreign language learned at the secondary level as the language of instruction in some subjects (European Commission, 1995). From the beginning of the 21st century, CLIL is part of regular schooling in primary and secondary education in most European countries (Eurydice, 2006).

Goris et al. (2019) provided the most recent comprehensive overview on the assessment of CLIL's contribution to English skills. They included only those 21 studies that were conducted in the past 20 years, written in English, included a measure of at least one English as a foreign language (EFL) skill, used students in mainstream primary and secondary education in a European country, and used a longitudinal design. The results of Goris et al. (2019) do not unambiguously support the assumption that CLIL students will gain more EFL skills than non-CLIL students in the same time. While null effects revealed in Germany and other northern European countries, positive effects were more common in Spain. For example, Admiraal, Westhoff, and de Bot (2006) compared $N = 584$ CLIL students with $N = 721$ non-CLIL students from a convenience sample in the first four years of secondary school in the Netherlands in English, Dutch, geography and history. CLIL students performed significantly better in verbal English skills and reading comprehension, even when controlling for initial vocabulary test performance, gender, cognitive ability, language background, exposure to English, and motivation to learn English. As Mearns, de Graaff, and Coyle (2017) reported, CLIL students display more language learning motivation than non-CLIL students, which highlights the need to control language learning motivation. In the vocabulary test, which was the only repeated test, the initial proficiency of the CLIL group was significantly better than the initial proficiency of the non-CLIL group, when the above-mentioned variables were controlled. However, no significant difference in the development of both groups could be found in the study by Admiraal et al. (2006). In other words, CLIL students maintained their lead, but did not extend it. Alonso, Grisaleña, and Campo (2008) investigated the English competencies of a convenience sample of $N = 159$ CLIL students and $N = 70$ non-CLIL students in Spanish secondary schools. They tested students from grades seven to eight (twelve to 14 years old), nine to ten (14–16 years old), and eleven to twelve (16–18 years old) and investigated the effects of CLIL on English competencies in listening, reading, speaking, and writing over one and a half years. To be able to compare the English competencies of CLIL and non-CLIL students, it was ensured that the control groups did not differ from the CLIL groups in terms of previous school grades, gender, and motivation to learn. The CLIL groups showed better English competencies than the non-CLIL groups at both measurement points, in all grades, and in all tests. Furthermore, CLIL was associated with an increased learning rate in this study.

EFL skills in general are relatively low in Spain and other southern European countries compared to high EFL skills in countries like Germany and other more northern European countries, according to Goris et al. (2019). They revealed that the philosophy of CLIL in Spain is to provide everyone a better EFL learning opportunity, while CLIL is much more elitist and highly selective in Germany and the Netherlands. Those differences were seen as an explanation for the positive effects of CLIL in southern European countries such as Spain and Italy and the lower impacts in northern European countries such as Sweden, the Netherlands, and Germany. However, other studies found contradictory results. The efficacy of CLIL was studied in the Netherlands by Verspoor, de Bot, and Xu (2015). They reported that CLIL students improved significantly more in English proficiency than non-CLIL students in the first year when controlling important covariates. However, CLIL students received two more English lessons per week than non-CLIL students, but Verspoor et al. (2015) were not able to control for these effects. In year three and when controlling the covariates, CLIL students did not improve relatively more than non-CLIL students. However, they maintained their lead. Goris et al. (2019) concluded that selection and preparation effects must be considered – otherwise, studies may lead to positive biases towards CLIL effects.

2.2. CLIL in Germany

The development of CLIL teaching in Germany is described by the

Kultusministerkonferenz (2013b). German CLIL teaching has its origins in the Franco-German reconciliation after the Second World War. This reconciliation should have an effect on the following generation, which is why the Élysée Treaty particularly emphasised the mutual teaching of language and culture. The first CLIL lessons took place in French at a grammar school in Baden-Württemberg in 1969. From 1980 to 1995, there was an increased search for concepts that lead to better results in foreign language learning and thus enable students to achieve higher foreign language competencies. This was followed by the development and testing of numerous methods, which agreed that the focus should be less on the formal aspects of foreign language learning and more on the meaning aspects in the communicative context. An integrated and process-oriented approach should be found. The duration of contact with the foreign language was also seen as crucial for success. Therefore, subject teaching in the foreign language should also increase the quantity of foreign language teaching. Only gradually the teaching of subject content moved into the foreground. As a result, the foreign language was seen more as a langugage of instruction in CLIL lessons. From the 1980s onwards, the importance of English in Europe increased. Consequently, the number of English-language CLIL programs in particular increased significantly in the 1980s and 1990s. The target group also expanded: while CLIL instruction was initially offered to particularly high-achieving students at grammar schools, CLIL instruction later became established at primary schools, *Realschulen*, and comprehensive schools as well[1] (Kultusministerkonferenz, 2013b). From 25 schools across Germany offering CLIL in 1987, the number rose to 366 schools in 1999, to 776 schools by 2006, and to more than 1,500 schools in 2013 (Köller, Leucht, & Pant, 2012; Rumlich, 2018). This number is likely to have continued to rise sharply since then. Nevertheless, CLIL is labelled as a contribution to the promotion of the best-performing students in the linguistic field and schools typically state some prerequisites to participate in a CLIL program, such as the willingness to learn or good school performance, which still makes CLIL an elitist program in Germany (Kultusministerkonferenz, 2013b; Rumlich, 2018).

English instruction in Germany starts typically in grade three in primary school. In some federal states, students already receive English instruction in grade one (Kultusministerkonferenz, 2013a). Before the start of the new CLIL instruction, students in Germany usually receive preparatory language instruction by increasing the number of English lessons in grades five and six. CLIL classes are therefore already formed at the beginning of secondary school in year five. CLIL is a voluntary offer and the decision for a CLIL program is made by the parents on the basis of a discussion between parents, child, teacher, and/or school management (Rumlich, 2018). In Germany, CLIL instruction usually started in grade seven (Rumlich, 2018) when students have acquired literacy skills in their native language (L1) (Dalton-Puffer, 2011). CLIL instruction is implemented in one to three content subjects, while the common language is used as the language of instruction in the other subjects. In Germany, mainly history, geography, biology, music, or politics are taught in English (Möller et al., 2017; Rumlich, 2018). CLIL teachers at German schools are typically non-native speakers of the foreign language and content experts rather than foreign language experts (Dalton-Puffer, 2011). Usually, foreign language instruction in CLIL classes is accompanied by explicit foreign language instruction (Dalton-Puffer, 2011).

As mentioned above, the national background should be taken into account in empirical research on CLIL effects. We present the empirical results for Germany thoroughly since our study used a German sample to examine the CLIL effect on English performance. Six German studies on CLIL effects measured at least selection effects. Only two studies went beyond that and additionally measured preparation effects. Three studies revealed the effects of CLIL on listening and/or reading comprehension in English.

First, the DESI study (Deutsch Englisch Schülerleistungen International [Assessment of Student Achievements in German and English as a Foreign Language], Klieme, 2008) tested $N = 1{,}945$ students who represented students in Germany. German performance, the socioeconomic status measured by the socioeconomic index, cognitive ability, educational pathway, first language, and gender served as variables to control for selection effects. CLIL students improved significantly more in English listening comprehension than non-CLIL students when controlling selection effects (Nold, Hartig, Hinz, & Rossa, 2008). Since the students were only tested at the beginning and end of grade nine, preparation effects could not be estimated and might have promoted a bias (Nold et al., 2008).

Second, $N = 9{,}867$ students from a representative sample were examined in Köller et al. (2012). They controlled parents' school-leaving qualifications, socioeconomic status, and linguistic skills in German. The study revealed that CLIL student's English achievement in reading and listening tests was significantly higher than the English achievement of non-CLIL students. However, they could not control for preparatory effects as students were only tested in grade nine.

Third, Dallinger et al. (2016) studied $N = 1{,}806$ eighth graders from a convenience sample and controlled prior achievement, general abilities, motivation, demographics, class composition, and teaching quality. CLIL students showed significantly greater skill development in English listening comprehension than non-CLIL students. The development of general English skills measured by C-test was comparable between both groups. The authors made no comments on the effect size of preparatory lessons. However, the study revealed that CLIL students possessed significantly better prior achievement and motivation, higher cognitive abilities, and higher socioeconomic status indicating selection effects.

Fourth, Rumlich (2018) examined a convenience sample of approximately 1,000 sixth and eighth graders. He statistically controlled combined selection and preparation effects. The improvement of C-test performance in the CLIL group compared to the control groups was no longer significant when age, gender, first language, and initial English test performance were controlled at the start of grade six. The results did not correspond with the previously mentioned studies as they contradict the assumption of a positive net effect of CLIL. However, Rumlich could not entangle selection and preparation effects.

Fifth, Feddermann et al. (2021) re-analyzed the longitudinal KESS data (Kompetenzen und Einstellungen von Schülerinnen und Schülern [Competencies and attitudes of students], Bos, Bonsen, & Gröhlich, 2009; Bos & Gröhlich, 2010; Bos & Pietsch, 2006), which allowed disentangling selection, preparation, and CLIL effects. Complete survey data from a 2002–2007 cohort were used to assess English skill development from seventh to eighth grade in C-test. The sample consisted of $N = 5{,}963$ German students. A propensity score matching in grade four led to comparable groups concerning primary school performance, sociodemographic variables, and cognitive abilities. The authors reported significant selection effects ($\beta = 0.34$) and additional significant preparation effects of similar size ($\beta = 0.37$). CLIL students maintained the lead they built up through selection and preparation but did not improve further with a slightly positive non-significant effect of $\beta = 0.06$.

Sixth, Feddermann et al. (under review) re-analyzed longitudinal and full survey data from the LAU study (Aspekte der Lernausgangslage und der Lernentwicklung [Aspects of the initial learning situation and learning development], Behörde für Schule und Berufsbildung, 2011, 2012). $N = 6{,}733$ German students were tested in English using C-tests in grades seven, nine, and eleven from 1996 to 2003. Propensity score matching in grade four produced comparable groups in terms of prior

[1] Grammar school: The most academic type of secondary school in Germany (grades 5–12 or 13) with the aim of preparing students to take up studies at a university. *Realschule*: A medium-academic type of secondary school (Grades 5–10) with the aim of preparing students for an apprenticeship. Comprehensive school: A type of school where all general secondary qualifications can be obtained.

achievement, sociodemographic variables, and cognitive abilities. The authors reported selection effects ($\beta = 0.13$) and significant preparation effects ($\beta = 0.64$). Controlling both led to similar results as in KESS. CLIL compensated for the assumed fading out-effect but did not produce significant added value ($\beta = 0.03$ from grade seven to nine; $\beta = 0.11$ from grade nine to eleven; $\beta = 0.13$ from grade seven to eleven).

Consequently, representative studies estimating selection, preparation, and CLIL effects (in countries where preparation is common, e.g., in Germany and Italy) are rare. The present study investigated German students from fifth to ninth grade. Due to the more recent longitudinal data set and the use of propensity score matching, the current research allowed well-founded and up-to-date statements about the effects of selection, preparation, and CLIL on English listening comprehension and/or reading comprehension. Like LAU and KESS, the present study also outperformed previous ones in method and content using a large and representative sample to separately estimate selection, preparation, and CLIL-effects on English proficiency. Going beyond LAU and KESS, the present study used listening and reading comprehension as indicators of second language competence and more recent data.

3. The present study

The KERMIT study (Kompetenzen ermitteln [Identifying competencies], Lücken et al., 2017) has been conducted since the 2012/2013 school year. The present study analyzed the data of the 2014–2019 cohort. The complete survey of Hamburg school students took place at a total of six points in time. In this paper, only data from grades five, seven, and nine were analyzed as the other measurement points were not relevant to investigate selection, preparation, and CLIL-effects and no English test was available (grades two and three) or differed in terms of the English tests (grade eight), which led to a lack of comparability between the measurement points.

The data were obtained from 49 Hamburg grammar schools (5th to 9th grade). Among those grammar schools, eight offered an English CLIL program, and 41 provided only regular monolingual instruction. We had to identify which students received CLIL instruction by conducting a follow-up survey. In doing so, we asked for the CLIL class names in order to be able to assign the classes in the data set to CLIL or non-CLIL. Since several schools taught CLIL in their course system and some schools did not provide any information on the CLIL classes, the number of schools in this study was somewhat lower than in LAU (Behörde für Schule und Berufsbildung, 2011; 2012) and KESS (Bos et al., 2009; Bos & Gröhlich, 2010; Bos & Pietsch, 2006). Only grammar schools were considered in the current research.

Given the selection criteria for admittance to a CLIL class, the choice of a CLIL program, and previous studies' results, we assumed that selection and preparation effects would become apparent. More importantly, we expected only small positive longitudinal effects on the English skills of CLIL students when selection and preparation effects were controlled for.

3.1. Sample

The data set included $N = 4{,}639$ grammar school students from KERMIT 5 (at the beginning of the school year 2014/2015) to KERMIT 9 (at the beginning of the school year 2018/19) with information about CLIL participation. The sample consisted of 47.7% boys and 52.3% girls. $N = 3{,}318$ students participated at every measurement point. $N = 448$ students participated in CLIL programs and $N = 4{,}191$ students were taught in regular classrooms and received the usual English as a foreign language instruction. Students received English instruction since grade one. Usually, students in grades one and two receive one English lesson and students in grades three and four receive three English lessons per week (Behörde für Schule und Berufsbildung, 2014a). The training and examination regulations stipulated 836 English lessons to be taught in Hamburg grammar schools from grades five to ten (§ 42 HmbGVBl). As a result, an average of 4 h of English instruction were provided per year in the non-CLIL group. CLIL students received an average of two additional English lessons in grades five and six to prepare for CLIL, starting in grade seven (Behörde für Schule und Berufsbildung, 2014b). We estimated the selection effect using the KERMIT 5 data. The second survey of the same cohort was conducted right after the beginning of the school year 2016/2017 in the seventh grade (KERMIT 7). We used the second survey data to estimate the preparation effect. In addition to the usual English teaching, the CLIL subjects were taught at least 3 h per week (Behörde für Schule und Berufsbildung, 2014b). In grade nine, CLIL subjects were usually taught 3 h per week. Two or more CLIL subjects were offered during lower secondary level (grades five to ten). Each was taught in at least two consecutive years (Behörde für Schule und Berufsbildung, 2014b). The KERMIT 9 data enabled the estimation of the CLIL effect. The number of English lessons for CLIL and non-CLIL students from grade one to grade nine is summed up in Table 1. The data used in the present study are marked bold.

Table 1
Number of English lessons per week for CLIL and non-CLIL students from grade one to grade nine.

		Number of English lessons per week	
Grade	Part of the KERMIT study	CLIL ($N = 448$)	Non-CLIL ($N = 4191$)
1	No	1	1
2	Yes	1	1
3	Yes	3	3
4	No	3	3
5	**Yes**	**6**	**4**
6	No	6	4
7	**Yes**	**4 + 3**	**4**
8	Yes	4 + 3	4
9	**Yes**	**4 + 3**	**4**

Note. In grades five to nine, the number of English lessons per week varied between the schools as they could freely distribute the compulsory 836 h of English instruction among the class levels. The data used in the present study are

3.2. Variables

The test scores of the proficiency variables were scaled via a one-parameter IRT model. Weighted likelihood estimates (WLE scores) were calculated for each domain. The tasks used in the different tests were based on the VERA study (Vergleichsarbeiten [written comparison tests], Institut zur Qualitätsentwicklung im Bildungswesen, 2021) and the IQB trends in student achievement. For each domain, two pseudo-parallel test versions with a varying order of the tasks were available. An overview of the test design is given in Table 2. Afterwards, the tests are described in greater detail.

3.2.1. Grade five: English listening comprehension

Thirty-one or thirty-two tasks on six recordings were applied to assess English listening comprehension skills in KERMIT 5. Students had to work on six open questions, five true/false tasks, 16 single-choice tasks, and four or five matching tasks (depending on the test version). For example, a matching task required that six different short instructions for gymnastic exercises are assigned to corresponding pictures. The text is based on six independent descriptions or instructions describing the actions a gymnast performs. The test reliability was $\alpha = .82$ in both test versions. The WLE reliability was $Rel(WLE) = 0.86$.

3.2.2. Grade five: German reading comprehension

The reading comprehension test used in KERMIT 5 consisted of a factual text and a short story. Out of the 19 or 21 questions (depending on the test version), there were six or eleven open questions, four or five true/false tasks, five or seven single-choice tasks, and one matching task. The reliabilities of the two test versions were $\alpha = .80$ and $\alpha = .79$. The

WLE reliability was $Rel(WLE) = 0.80$.

3.2.3. Grade five: mathematics proficiency

Thirty-four or thirty-five tasks (depending on the test version) from the fields of numbers, measurement, space and shape, functional relationships, and data and chance were used to test mathematical performance in KERMIT 5. The questions were designed as 23 or 25 open questions, one or none true/false task, and nine or eleven single-choice tasks. The reliabilities of the mathematics test versions were $\alpha = .84$ and $\alpha = .83$. The WLE reliability was $Rel(WLE) = 0.90$.

Table 2
Overview of the test design.

	English		German Reading	Mathematics	Life science
Grade	Listening	Reading			
5	31 or 32 tasks	Not tested in KERMIT	19 or 21 tasks	34 or 35 tasks	19 or 24 tasks
7	35 tasks	35 tasks	Data not used in the present study		
9	Not tested in KERMIT	30 tasks	Data not used in the present study		

Note. The item number depends on the test version. German, mathematics, and life science test scores were used for propensity score matching.

3.2.4. Grade five: life science proficiency

The KERMIT 5 life science test consisted of 19 or 24 tasks (depending on the test version). The tasks derived from the fields of biology, physics, and chemistry. Students had to work on four or five open questions, one or two true/false tasks, twelve or 16 single-choice tasks, and one or two matching tasks. The reliabilities of the test versions were $\alpha = .73$ and $\alpha = .72$. The WLE reliability was $Rel(WLE) = 0.76$.

3.2.5. Grade seven: English listening and reading comprehension

A listening comprehension test and a reading comprehension test were used in KERMIT 7 to assess English language proficiency.

In the KERMIT 7 listening comprehension test, students listened to six audio files and answered 35 related questions (two to eleven questions per recording), which were designed as 23 open questions and twelve single-choice tasks. Two texts from KERMIT 5 with eight and four questions each were used again in KERMIT 7. The reliabilities of the test versions were $\alpha = .86$ and $\alpha = .85$. The WLE reliability of the English listening comprehension test was $Rel(WLE) = 0.89$.

In the KERMIT 7 reading comprehension test, students read six texts and answered five or six questions on each text. In total, students answered 35 questions, which consisted of 15 open questions, nine single-choice tasks, and eleven matching tasks. The reliability of this test was $\alpha = .86$ in both test versions. The WLE reliability was $Rel(WLE) = 0.91$.

3.2.6. Grade nine: English reading comprehension

In grade nine, a reading comprehension test was conducted to assess students' English proficiency. Both test versions cosisted of 30 questions each. Students read seven texts and had to work on 19 open questions and eleven matching tasks. Two texts from KERMIT 7 with five questions each were used again in KERMIT 9. The reliabilities of the reading comprehension test versions were $\alpha = .89$ and $\alpha = .88$. The WLE reliability was $Rel(WLE) = 0.93$.

4. Analyses

WLEs were estimated separately for KERMIT 5 to KERMIT 7 and KERMIT 7 to KERMIT 9. Like other (inter)national large-scale student assessment studies, the WLEs were linearly transformed to a mean of 500 and a standard deviation of 100. This transformation was done separately for KERMIT 5 to 7 and KERMIT 7 to 9. To eliminate significant differences between CLIL and non-CLIL students, propensity score matching was conducted. Since propensity score matching required a data set without missing values, we previously used multiple imputation. Subsequently, we estimated the probability of belonging to the treatment group (propensity score) for each student and matched CLIL and non-CLIL students with equal propensity scores. This procedure eliminated significant differences between both groups.

Missing data is a well-known problem in longitudinal studies. On average, about 11% of the data were missing per variable in the present research. In grades five, seven, and nine, the percentage of missing values varied between 9.3% and 21% in the English tests and between 9% and 20.4% in all other tests. The sociodemographic variables had between 0% and 2.4% missing values. In the literature, multiple imputation is recommended to deal with missing values (Lüdtke, Robitzsch, Trautwein, & Köller, 2007, see Supplement A for further information on multiple imputation). Multiple imputation was performed with the R-package mice (version 3.12.0, Van Buuren & Groothuis-Oudshoorn, 2020). We considered the multi-level data structure using the class ID as a grouping variable and the R-package pan (version 1.6, Schafer & Zhao, 2018). The quickpred function in mice led to suitable variables predicting the missing values by conducting correlation analysis (Van Buuren & Groothuis-Oudshoorn, 2011). The correlation coefficients are presented in Table 3 (non-CLIL group) and Table 4 (CLIL group). Overall, the non-CLIL group (Table 3) shows higher and more often significant correlation coefficients than the CLIL group (Table 4). For example, all variables correlate significantly with the RISE-status in the non-CLIL group, whereas only six variables (performance variables and migration background) show significant correlation coefficients with the RISE-status in the CLIL group. This may be due to a larger sample size in the non-CLIL group and due to selection leading to limited variabiliy in the CLIL group. It is interesting that the correlation between German reading grades and RISE-status decreases in higher classes in both groups. The highest correlation coefficients can be found between the different measurement points in the same subject and between mathematics and life science grades. The proportion of the total variance of a variable that can be attributed to the variance between imputations (*fraction of missing values, FMI*) was 0.882 (Madley-Dowd, Hughes, Tilling, & Heron, 2019). Therefore, in this study, $m = 100 \times 0.882 = 88.2 \approx 90$ imputations were performed to obtain reliable estimates (White, Royston, & Wood, 2011). Thus, we created 90 complete data sets with $N = 4{,}639$ cases each. Diagrams showed the mean values and the variance of all imputations and the fluctuation per iteration of each variable (Van Buuren & Groothuis-Oudshoorn, 2011) and revealed that convergence was achieved.

In the next step, we used the imputed data for propensity score matching with the aim to draw strong inferences on the effects of CLIL on English reading comprehension (see Supplement A for more detailed information on propensity score matching). Ten variables related to prior achievement or sociodemographic background were believed to be associated with the positive selection of CLIL students in the present research (Möller et al., 2017; Rumlich, 2018). The set of variables included English listening comprehension, German reading comprehension, mathematics, and life science test scores at the beginning of year five and sociodemographic variables such as gender, year of birth, migration background, language between parents and child, socioeconomic status, and recommended school type. The score was estimated individually for each student using the R-package MatchThem (version 0.9.3, Pishgar & Greifer, 2020). We used three routines to match CLIL and non-CLIL students with equal propensity scores: 1:1 without replacement, 1:1 with replacement, and 1:9.35 with replacement, each combined with nearest neighbor matching, a caliper of $c = .025$, and logit distance matching for sociodemographic and prior achievement variables. All three routines eliminated pre-treatment differences in relevant covariates between CLIL and non-CLIL students successfully. The subsequent analyses and results are based on the 1:9.35 matching with replacement as it revealed the smallest standardized bias (% bias = 0.80 for 1:1 matching without replacement, % bias = 0.69 for 1:1

Table 3
Correlation coefficients within the non-CLIL group for all variables ($N = 4191$).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1) Gender	1.00																			
(2) Year of birth	.08**	1.00																		
(3) Language between parents and child	.04**	-.15**	1.00																	
(4) RISE-status	-.03	.12**	-.36**	1.00																
(5) Migration background	.03*	-.10**	.50**	-.36**	1.00															
(6) Grammar school recommendation	-.02	.10**	-.16**	.10**	-.15**	1.00														
(7) English listening grade five	-.05**	.00	-.07**	.09**	-.01	.15**	1.00													
(8) German reading grade five	.01	-.01	-.25**	.19**	-.25**	.25**	.28**	1.00												
(9) Mathematics grade five	-.18**	.00	-.21**	.20**	-.24**	.27**	.34**	.47**	1.00											
(10) Life science grade five	-.14**	.00	-.28**	.23**	-.28**	.21**	.25**	.47**	.55**	1.00										
(11) English listening grade seven	.01	.03	-.07**	.15**	-.02	.16**	.51**	.35**	.33**	.31**	1.00									
(12) English reading grade seven	.01	.02	-.06**	.11**	-.03*	.17**	.45**	.35**	.35**	.28**	.66**	1.00								
(13) German reading grade seven	.00	.06**	-.21**	.17**	-.22**	.21**	.25**	.45**	.38**	.43**	.37**	.38**	1.00							
(14) German orthography grade seven	.13**	.06**	-.12**	.13**	-.13**	.25**	.27**	.37**	.36**	.27**	.38**	.39**	.37**	1.00						
(15) Mathematics grade seven	-.18**	.05**	-.18**	.16**	-.19**	.24**	.27**	.39**	.68**	.51**	.36**	.39**	.42**	.42**	1.00					
(16) Life science grade seven	-.15**	.04**	-.25**	.19**	-.25**	.19**	.23**	.39**	.53**	.59**	.34**	.31**	.44**	.30**	.56**	1.00				
(17) English reading grade nine	.05**	.09**	-.13**	.11**	-.11**	.17**	.36**	.30**	.37**	.35**	.53**	.52**	.40**	.40**	.38**	.35**	1.00			
(18) German reading grade nine	.00	.11**	-.22**	.17**	-.22**	.22**	.20**	.39**	.34**	.35**	.32**	.32**	.41**	.32**	.39**	.37**	.46**	1.00		
(19) Mathematics grade nine	-.20**	.11**	-.22**	.17**	-.23**	.23**	.22**	.35**	.64**	.50**	.31**	.31**	.35**	.34**	.68**	.54**	.42**	.44**	1.00	
(20) Life science grade nine	-.14**	.12**	-.26**	.17**	-.26**	.20**	.24**	.39**	.55**	.56**	.33**	.32**	.41**	.28**	.58**	.61**	.45**	.48**	.65**	1.00

Note. $^{*}p \leq .05$, $^{**}p \leq .01$. The RISE-status is calculated on the basis of seven indicators (e.g., the proportion of unemployed persons or lower school qualifications) and divides residential districts in Hamburg into more or less socially disadvantaged areas. The RISE-status refers to the residential districts of the students and ranges between 1 (very low) and 4 (high). A higher value describes a better status of the residential district.

Table 4
Correlation coefficients within the CLIL group for all variables ($N = 448$).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1) Gender	1.00																			
(2) Year of birth	.13**	1.00																		
(3) Language between parents and child	-.01	-.18**	1.00																	
(4) RISE-status	.00	.02	-.26**	1.00																
(5) Migration background	-.12*	-.03	.40**	-.28**	1.00															
(6) Grammar school recommendation	.01	.07	-.15**	.09	-.20**	1.00														
(7) English listening grade five	-.04	-.01	.06	.03	.13*	.04	1.00													
(8) German reading grade five	.10	.01	-.18**	.23**	-.29**	.28**	.21**	1.00												
(9) Mathematics grade five	-.18**	-.06	-.06	.07	-.05	.17**	.29**	.49**	1.00											
(10) Life science grade five	-.13*	-.05	-.20**	.19**	-.19**	.11*	.24**	.44**	.57**	1.00										
(11) English listening grade seven	.02	-.02	.03	.06	.10*	-.09	.54**	.33**	.37**	.34**	1.00									
(12) English reading grade seven	.04	.02	.05	-.01	.07	-.02	.50**	.37**	.37**	.33**	.70**	1.00								
(13) German reading grade seven	.06	-.03	-.13**	.11*	-.17**	.16**	.20**	.49**	.41**	.43**	.32**	.38**	1.00							
(14) German orthography grade seven	.15**	.04	-.02	-.03	.00	.08	.19**	.29**	.37**	.30**	.30**	.38**	.38**	1.00						
(15) Mathematics grade seven	-.11*	.07	-.07	.04	.01	.11*	.24**	.40**	.69**	.50**	.35**	.42**	.42**	.43**	1.00					
(16) Life science grade seven	-.14**	.05	-.16**	.12*	-.21**	.14**	.21**	.45**	.56**	.58**	.27**	.31**	.45**	.26**	.54**	1.00				
(17) English reading grade nine	.16**	.05	-.08	.07	.01	.10	.32**	.42**	.44**	.39**	.48**	.51**	.42**	.38**	.45**	.35**	1.00			
(18) German reading grade nine	.05	.06	-.18**	.10	-.21**	.21**	.22**	.30**	.35**	.39**	.26**	.23**	.39**	.28**	.37**	.41**	.40**	1.00		
(19) Mathematics grade nine	-.11*	.04	-.11*	.08	-.08	.07	.24**	.36**	.67**	.55**	.29**	.33**	.37**	.35**	.71**	.57**	.42**	.43**	1.00	
(20) Life science grade nine	-.01	.03	-.16**	.11*	-.12*	.11*	.17**	.40**	.53**	.54**	.27**	.31**	.49**	.27**	.58**	.61**	.44**	.47**	.62**	1.00

Note. $^{*}p \leq .05$, $^{**}p \leq .01$. The RISE-status is calculated on the basis of seven indicators (e.g., the proportion of unemployed persons or lower school qualifications) and divides residential districts in Hamburg into more or less socially disadvantaged areas. The RISE-status refers to the residential districts of the students and ranges between 1 (very low) and 4 (high). A higher value describes a better status of the residential district.

matching with replacement, and % bias = 0.47 for 1:9.35 matching with replacement). Only individuals within the area of common support were compared. In other words, the analyses applied only to CLIL and non-CLIL students with equal propensity scores. The matching procedure eliminated relevant pre-treatment differences and led to comparability between CLIL and non-CLIL students.

Before the start of CLIL, differences in prior achievement and sociodemographic variables between CLIL and non-CLIL students were tested for statistical significance. We included the class ID as a cluster variable when estimating selection effects on English performance in the unmatched data set. Propensity score matching led to controlled selection effects. We then analyzed the English listening and reading comprehension performance of KERMIT 7 in the matched data set to estimate preparation effects. Therefore, we also used the class ID as a cluster variable.

The estimation of the CLIL effect was more complex. There were no students with preparatory lessons but without CLIL instruction. However, this group would be necessary to estimate the exact CLIL effect. For this reason, we had to make additional assumptions. The ANCOVA and the Difference-in-Differences analysis enable statements on the separate CLIL effect. Both address the same research question as they intend to estimate causal effects. However, they are based on different assumptions, which were hardly met in the present design (Lüdtke & Robitzsch, 2020). No possible fading out-effect of the preparatory lessons was assumed when conducting Difference-in-Differences analyses to estimate the additive CLIL effect between KERMIT 7 and 9 (see Supplement A for further information on Difference-in-Differences analyses). We used the class ID as a cluster variable. However, the common trend assumption did not seem plausible. Therefore, we conducted an ANCOVA based on the empirically observed stability controlling selection and preparation effects. In case of no subsequent CLIL instruction, a fading out-effect of the preparatory lessons was assumed. This procedure revealed the combined fading out and CLIL-effect over two years. We used the English performance of KERMIT 7 as a predictor and the class ID as a cluster variable. However, the exact size of a possible fading out-effect remained unclear. Both analyses were conducted in *Mplus* (Muthén & Muthén, 1998-2017) before and after propensity score matching. The true CLIL effect seems to be between the outcomes of both analyses: The DiD score provides the lower bound, and the ANCOVA estimator provides the upper bound for the true CLIL effect (Lüdtke & Robitzsch, 2020).

5. Results

We firstly present the results on the selection effect, followed by the outcomes of preparation and CLIL effects before and after matching.

5.1. Selection effect: differences before propensity score matching

We tested predictor variables for significant differences between CLIL and non-CLIL students before and after propensity score matching. The results and the descriptive data before matching are presented in Table 5. The average standardized bias, which is the mean difference divided by the pooled standard deviation (Rosenbaum & Rubin, 1985), before matching is 25.2%. The groups differed significantly in all predictors. CLIL students showed significantly better prior achievement and significant advantages in sociodemographic variables. The CLIL sample consisted of more girls, and we found significant differences in English, German, mathematics, and life science test scores in year five, year of birth, language between parents and child, RISE-status, migration background, and grammar school recommendation (see Table 5), indicating selection effects. The selection effect was $\beta = 0.59$ ($p < .001$) in English listening comprehension.

Fig. 1 shows the area of common support. In this area, treatment and comparison groups overlap (Retelsdorf, Becker, Köller, & Möller, 2012). Conclusions on treatment effects are only valid for CLIL students who were matched with non-CLIL students with similar propensity scores, that is, for students within the area of common support (Garrido et al.,

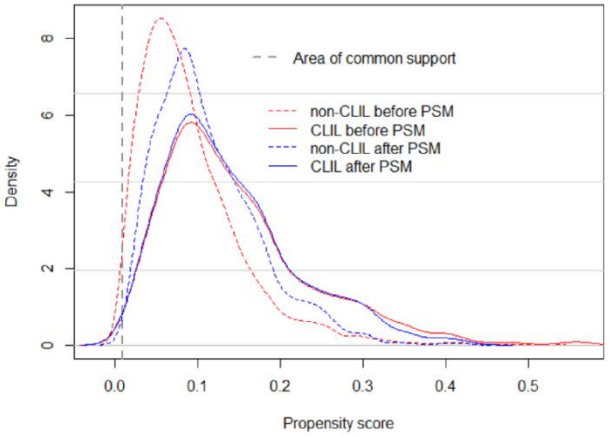

Fig. 1. Area of common support before and after matching (exemplary for one imputation).

Table 5
Before matching: Imbalance in predictor variables between CLIL and non-CLIL students ($N = 4{,}639$).

	CLIL ($N = 448$)		Non-CLIL ($N = 4191$)				
Predictor variable	*M*	*SD*	*M*	*SD*	*t*	*p*	% bias
KERMIT 5 performance							
English listening WLE	412.74	59.61	378.98	56.18	5.91	<.001	58.29
German reading WLE	505.74	88.20	488.45	87.32	2.26	<.05	19.70
Mathematics WLE	526.53	73.20	507.92	75.50	3.24	<.01	25.03
Life science WLE	501.08	85.16	479.49	83.04	3.09	<.01	25.67
Sociodemographic variables							
Gender (0 = male)	0.58	0.49	0.52	0.50	2.70	<.01	13.49
Year of birth	2003.70	0.49	2003.61	0.54	3.54	<.001	16.72
Language between parents and child (0 = German, 1 = other)	0.10	0.30	0.16	0.37	−3.84	<.001	−17.47
RISE-status	3.37	0.71	3.05	0.77	8.86	<.001	43.17
Migration background (0 = no, 1 = yes)	0.37	0.48	0.41	0.49	−2.01	<.05	−9.85
Grammar school recommendation (0 = no, 1 = yes)	0.85	0.36	0.76	0.43	4.84	<.001	22.32
Total bias (Mean [|% bias|])							25.17

Note. *p*-value for a two-tailed test. Performance estimates are clustered for class ID. The RISE-status is calculated on the basis of seven indicators (e.g., the proportion of unemployed persons or lower school qualifications) and divides residential districts in Hamburg into more or less socially disadvantaged areas. The RISE-status refers to the residential districts of the students and ranges between 1 (very low) and 4 (high). A higher value describes a better status of the residential district.

2014). Therefore, we compared the lowest and highest propensity scores of both groups. The overlapping region of the two intervals calls the area of common support. Almost every student was located in this area, indicating that nearly every CLIL student matched with a non-CLIL student in the present research. After propensity score matching, the distributions of the propensity scores were very similar, demonstrating the comparability of CLIL and non-CLIL students. Therefore, significant pre-treatment differences in sociodemographic variables and prior achievement between CLIL and non-CLIL students diminished.

Table 6 shows the descriptive statistics after propensity score matching. The average standardized bias before matching was 25.2% (see Table 5) and decreased to 0.5% after matching (see Table 6). A standardized bias of less than 5% indicates a balanced distribution among the covariates (Caliendo & Kopeinig, 2008). Differences between the predictor variables were no longer significant. Thus, selection effects were controlled by matching.

Table 6
After matching: predictor variables between CLIL and non-CLIL students.

	CLIL		Non-CLIL				
Predictor variable	*M*	*SD*	*M*	*SD*	*t*	*p*	% bias
KERMIT 5 performance							
English listening WLE	409.72	56.94	410.07	58.57	−0.07	.95	−0.60
German reading WLE	505.05	88.41	504.00	83.75	0.13	.89	1.22
Mathematics WLE	525.38	72.52	525.98	76.62	−0.10	.92	−0.80
Life science WLE	499.82	84.91	499.79	82.78	0.01	1.00	0.04
Sociodemographic variables							
Gender (0 = male)	0.58	0.49	0.58	0.49	−0.01	1.00	0.00
Year of birth	2003.69	0.49	2003.69	0.51	0.04	.97	0.20
Language between parents and child (0 = German, 1 = other)	0.10	0.30	0.11	0.31	−0.04	.97	−0.33
RISE-status	3.36	0.72	3.35	0.61	0.18	.86	1.05
Migration background (0 = no, 1 = yes)	0.36	0.48	0.36	0.48	−0.07	.95	−0.42
Grammar school recommendation (0 = no, 1 = yes)	0.85	0.36	0.85	0.36	−0.01	.99	0.00
Total bias (Mean [|% bias|])							0.47

Note. *p*-value for a two-tailed test. Performance estimates are clustered for class ID. The RISE-status is calculated on the basis of seven indicators (e.g., the proportion of unemployed persons or lower school qualifications) and divides residential districts in Hamburg into more or less socially disadvantaged areas. The RISE-status refers to the residential districts of the students and ranges between 1 (very low) and 4 (high). A higher value describes a better status of the residential district.

5.2. Preparation effect: differences after propensity score matching

Table 7 shows the descriptive statistics of the English listening and reading comprehension performance for grades five to seven before and after matching. Table 8 presents the preparation effects. Differentiated statements about the effect of preparatory instruction in grades five and six were possible as the KERMIT 7 survey was carried out at the beginning of grade seven. In grade seven, significant effects were observed in listening comprehension ($\beta = 0.92$) and reading comprehension ($\beta = 0.68$) between the two groups before propensity score matching (PSM). These effects decreased to still significant $\beta = 0.56$ in listening comprehension and $\beta = 0.35$ in reading comprehension after propensity score matching (see Table 8). As CLIL did not start until grade seven and differences in relevant covariates were eliminated in KERMIT 5, the significant effects of $\beta = 0.56$ and $\beta = 0.35$ determined in this study were attributable to the preparatory classes, including an average of two additional English lessons in grades five and six.

Table 7
Before and after matching: English listening and reading comprehension of CLIL and non-CLIL students for grades five to seven.

			CLIL		Non-CLIL	
Grade	Test	PSM	*M*	*SD*	*M*	*SD*
5	Listening	No	412.74	59.61	378.98	56.18
	Listening	Yes	409.72	56.94	410.07	58.57
7	Listening	No	629.12	71.48	560.02	72.10
	Listening	Yes	627.24	70.86	585.04	75.22
	Reading	No	606.41	67.88	561.35	64.66
	Reading	Yes	604.83	67.46	581.37	67.41

Note. PSM = Propensity score matching. WLEs were estimated separately for grades five to seven.

5.3. CLIL effect: effect sizes before and after propensity score matching

Propensity score matching led to reduced differences in reading comprehension performance between CLIL and non-CLIL students in grades seven and nine (see Table 9). Table 10 shows that ignoring selection and preparation effects could lead to an overestimation of the CLIL effect. Before propensity score matching and without controlling English performance in grade seven, the significant difference between the two groups in English reading comprehension was $\beta = 0.64$ in grade nine. Considering selection effects, reduced the difference in grade nine to a still significant $\beta = 0.36$. Additionally considering preparation effects by controlling for English reading performance in grade seven further reduced the effect of CLIL to a significant $\beta = 0.18$. We also estimated the effect of CLIL on English reading performance in grade nine mediated by English reading performance in grade seven and found a significant indirect effect of $\beta = 0.18$. However, these effects include a possible fading out-effect. We controlled selection effects, considered preparation effects, and excluded a possible fading out-effect in the subsequent Difference-in-Differences analysis. As a result, we could no longer find any (additive) CLIL effect ($\beta = 0.04$) between grades seven and nine.

6. Discussion

Selection and preparation effects were rarely considered in previous empirical research on CLIL effects. These studies often revealed positive CLIL effects, especially in countries with low levels of English proficiency. Therefore, this study aimed to estimate selection and preparation effects and the effect of CLIL on second language learning in Germany. We used a large sample from a full longitudinal survey to assess selection and preparation effects and considered both to compare the English skills of CLIL and non-CLIL grammar school students from grade five to grade nine and to estimate the CLIL effect.

6.1. Selection effect

Significant differences in prior achievement and sociodemographic variables indicated an advantage of future CLIL students in relevant covariates for English proficiency at the beginning of grade five. Due to the requirements stated by the schools for participation in a CLIL program (Rumlich, 2018) and the label as a contribution to the promotion

Table 8
Before and after matching: Preparation effects on English listening and reading comprehension for grade seven (ANOVA with *z-standardized*, latent variables, robust standard errors).

	Before PSM				After PSM			
Effects	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*
Listening performance 5 on listening performance 7 *(stability)*	0.633	0.019	33.297	<.001	0.613	0.038	16.304	<.001
Listening performance 5 on reading performance 7 *(stability)*	0.566	0.020	28.558	<.001	0.579	0.036	16.119	<.001
CLIL on listening performance 7 *(preparation effect)*	0.923	0.098	9.425	<.001	0.555	0.103	5.404	<.001
CLIL on reading performance 7 *(preparation effect)*	0.679	0.089	7.614	<.001	0.345	0.096	3.600	<.001

Note. PSM = Propensity score matching.

of the best-performing students in the linguistic field (Kultusministerkonferenz, 2013b), CLIL attracted higher-performing students with a more favorable sociodemographic background. Before the start of CLIL and preparatory lessons, a significant selection effect occurred favoring future CLIL students in English listening comprehension. In light of the earlier research by Verspoor et al. (2015) revealing a higher language learning motivation of the CLIL students compared to the non-CLIL students at the beginning of the CLIL program in the Netherlands, selection effects also become apparent in terms of motivation. No significant differences in initial English performance in writing and vocabulary were found between CLIL students and non-CLIL students from schools without CLIL programs. But, when comparing CLIL students and non-CLIL students from the same schools, CLIL students showed significantly better initial English skills than non-CLIL students, confirming selection effects. Admiraal et al. (2006) also reported significantly better English vocabulary knowledge of CLIL students compared to non-CLIL students at the beginning of the CLIL program for the Netherlands. The present study confirmed that controlling selection effects is indispensable when estimating the CLIL effect (see also Dallinger et al., 2016; Feddermann et al., under review; Feddermann et al., 2021; Rumlich, 2018). Furthermore, the selection effect in the present study (β = 0.59) exceeds the selection effects in LAU (β = 0.13) and KESS (β = 0.34). Thus, the selection effect seems to have increased over the last decades. It is beyond the scope of our research to analyze the reasons for this increase, however, it is possible that with growing exposure and competence to the English language it seems only attractive to a strongly elitist sample to take part in such programs. Among other advantages, CLIL students also showed a favorable socioeconomic status (RISE). At the same time, the importance of English has risen considerably. It is conceivable that in times of increasing internationalisation parents with a higher RISE status attach more importance to their children's good English skills than before and than parents with a lower RISE status.

Table 9
Before and after matching: English reading comprehension of CLIL and non-CLIL students for grades seven to nine.

			CLIL		Non-CLIL	
Grade	Test	PSM	*M*	*SD*	*M*	*SD*
7	Reading	No	486.75	68.21	444.49	64.18
	Reading	Yes	485.26	67.87	463.46	66.75
9	Reading	No	606.33	74.25	558.24	74.03
	Reading	Yes	604.95	73.87	577.76	74.36

Note. PSM = Propensity score matching. WLEs were estimated separately for grades seven to nine.

6.2. Preparation effect

In Germany and some other countries, students receive additional English lessons to be prepared for the upcoming CLIL instruction. We assumed that these preparatory lessons would improve the English skills. However, previous research rarely considered this preparation effect when estimating the CLIL effect. Only KESS and LAU considered the effect of preparatory lessons when estimating the CLIL effect (Feddermann et al., under review; Feddermann et al., 2021).

Large significant effects in English listening and reading comprehension test scores occurred between CLIL and non-CLIL students at the beginning of grade seven when selection and preparation effects were not controlled. Propensity score matching reduced the (still significant) effects in grade seven, indicating the preparation effects from two years of additional English instruction. The preparation effect on listening comprehension exceeded the preparation effect on reading comprehension. KESS and LAU used C-tests and also revealed large significant effects at the beginning of grade seven without matching (Feddermann et al., under review; Feddermann et al., 2021). After propensity score matching, the preparation effect decreased to a still significant but much smaller effect (Feddermann et al., under review; Feddermann et al., 2021). In comparison, the preparation effects in KERMIT (β = 0.56 in listening comprehension, β = 0.35 in reading comprehension) were smaller than the one in LAU (β = 0.64), whereas the preparation effect in reading comprehension was comparable to the one in KESS and the preparation effect in listening comprehension exceeded the one in KESS (β = 0.37).

Consequently, not only the selection effect but also the large preparation effect has to be considered when estimating the CLIIL effect. Preparatory classes are not common in all countries. However, we could show the potential of preparatory lessons for improving English reading comprehension and especially English listening comprehension performance. Further, Dallinger et al. (2016) showed that prior achievement was the most important confounder when estimating the CLIL effect on English proficiency. Thus, preparatory lessons can help to follow CLIL instruction more easily. The present study showed that ignoring preparation effects could lead to an overestimated CLIL effect. Therefore, future research has to consider preparation effects if additional English

Table 10
Before and after matching: CLIL effect on English reading comprehension for grade nine (ANCOVA with z-standardized, latent variables, robust standard errors).

	Before PSM				After PSM			
Effects	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*
Performance 7 on performance 9 *(stability)*	0.564	0.028	20.069	<.001	0.566	0.038	14.995	<.001
CLIL on performance 9 *(overall effect)*	0.638	0.094	6.799	<.001	0.363	0.100	3.626	<.001
Specific effect CLIL on performance 9 when controlling for performance 7 *(CLIL effect)*	0.274	0.081	3.386	<.01	0.179	0.086	2.083	<.05
Indirect effect CLIL on performance 9 *(preparation via performance 7 on performance 9)*	0.364	0.053	6.829	<.001	0.184	0.052	3.552	<.001
Additive effect of CLIL year 7–9	−0.006	0.085	−0.071	.94	0.039	0.092	0.422	.67

Note. PSM = Propensity score matching.

lessons are conducted before students enter a CLIL program. Our results indicate that it might be more worthwhile to offer two additional English lessons in grades five and six for all grammar school students to improve their English proficiency instead of implementing a CLIL curriculum for the privileged few.

6.3. CLIL effect

The selection and preparation effects exceeded the CLIL effect itself. The CLIL effect on English reading comprehension was much smaller when selection and preparation effects were statistically controlled for. These findings showed the indispensability of controlling selection and preparation effects. Otherwise, differences in test scores between CLIL and non-CLIL students could not be clearly attributed to CLIL as these differences may have existed before the start of the CLIL program and be favored by differences in performance-related covariates and preparatory lessons. Consequently, ignoring relevant covariates and additional English lessons resulted in an overestimation of the CLIL effect. Additionally, the indirect effect of CLIL ($\beta = 0.184$) was slightly greater than the direct effect ($\beta = 0.179$) and also became significant. In other words, 51% of the total CLIL effect ($\beta = 0.36$) were mediated by prior achievement in grade seven, which showed the importance of prior achievement for English proficiency in accordance with Dallinger et al. (2016). However, the estimation of the pure CLIL effect is not possible. As CLIL instruction and preparatory lessons are confounded, we could rather estimate the effect of the CLIL-component of a complex treatment of preparatory lessons and CLIL. We additionally conducted a Difference-in-Differences analysis to assess the additive CLIL effect under the assumption of perfect stability between grades seven and nine. However, after finishing the preparatory lessons, fading out-effects may have occurred, as the indirect effect of preparation indicates. Consequently, CLIL maintains the advantage in English reading comprehension built up through selection and preparation by compensating for the fading out-effect. However, in our study CLIL did not contribute any significant added value to English reading comprehension skills.

The non-significant additive CLIL effect on English reading comprehension in the present research confirms the results of previous studies. Rumlich (2018) analyzed the English performance increase in C-test from grade six to grade eight and reported non-significant effects for CLIL and non-CLIL students. KESS and LAU also analyzed C-tests and revealed that CLIL compensated for the assumed fading out-effect but did not result in significant added value (Feddermann et al., under review; Feddermann et al., 2021). Admiraal et al. (2006) studied the effect of CLIL on the development of voabulary knowledge in the Netherlands and also reported no significant differences between CLIL and non-CLIL students. In contrast, the study of Dallinger et al. (2016) revealed significantly greater skill development of CLIL students compared to non-CLIL students in English listening comprehension considering general abilities, demographics, motivation, and quality of instruction. These findings indicate that the CLIL effect varies according to the investigated English skill. Students are confronted with different speakers (Dalton-Puffer, 2008) and encouraged to use the foreign language in CLIL classes (Dallinger et al., 2016). Therefore, listening comprehension could benefit more from CLIL than reading comprehension.

Furthermore, we compared the effects in LAU, KESS, and KERMIT. The extramural exposure to English increased over time (LAU started in 1996, KESS began in 2003, KERMIT started in 2014), and English instruction started two years earlier than in KESS and four years earlier than in LAU. Even though the selection, preparation, and CLIL effects may not be directly comparable, they were able to show the trend of the CLIL effect's development. Further, in LAU and KESS, C-tests were used to assess students' English skills, whereas a reading comprehension test was conducted in KERMIT. In their research, Sylvén (2013) has already shown that extramural exposure is a key factor in explaining the CLIL effect differences across Europe. The CLIL effect variation between LAU, KESS, and KERMIT may be explained by extramural exposure to English. Over the last years, students encounter English increasingly outside school, and the start of English instruction in primary school strongly changed. In LAU, English instruction started in grade five (Behörde für Schule und Berufsbildung, 2012). Students learned English mainly at school. In the following years, social media, streaming services for films, series, and music emerged, and students encounter English increasingly outside of school (Sundqvist & Sylvén, 2016). Besides, English instruction started in grade three in KESS (May, 2006). Since the school year 2011/2012, English instruction started in grade one (Behörde für Schule und Berufsbildung, 2020). Further, social media and English films, series, and music became part of KERMIT students' everyday life. Peters (2018) showed for Belgian students aged 15 or 16 and 19 that the effect of extramural exposure to English on vocabulary knowledge is greater than the effect of in-school English instruction. Sylvén (2006) reported for Swedish pupils that English vocabulary knowledge can be attributed more to extramural exposure to the foreign language, especially to reading English texts, than to CLIL. This is presumably also related to the pronounced use of English outside of school in Sweden (Sylvén, 2013). Complementarily, a study by De Wilde, Brysbaert, and Eyckmans (2020) in Belgium revealed that social media use and out-of-school conversations in English were significant predictors of vocabulary knowledge, reading and listening comprehension, and verbal English skills among students aged ten to twelve. In addition, playing English-language computer games was a significant predictor of vocabulary knowledge and verbal English skills. Students often use English outside of school unconsciously and in authentic situations (Sylvén & Olsson, 2015). This could be associated with improved English competencies according to the Natural Approach (Krashen & Terrell, 1998). In addition, opportunities for input, output, and interaction have increased due to expanded out-of-school opportunities for English acquisition and earlier English language teaching, which is also seen as beneficial for the development of English language competencies (Krashen, 1985; Long, 1996; Sundqvist & Olin-Scheller, 2013; Swain, 1985). We assumed that smaller CLIL, or, more precisely, smaller fading out-effects would show up in KESS than in LAU due to the higher level of extramural exposure to English and the earlier introduction of English instruction in KESS. These changes could also explain the even smaller CLIL effect in KERMIT. Sylvén (2013) explained the small CLIL effects with the high level of students' initial English proficiency, which led to limited room for improvement. Therefore, this explanation could also be valid for the different CLIL effects across European countries and the varying CLIL effects over time in Germany.

6.4. Limitations

We could show that selection and preparatory lessons have a greater effect on English performance than CLIL itself. In light of findings by Dallinger et al. (2016), additional influences have to be considered in future studies. Dallinger et al. (2016) reported that motivation, among others, is a significant predictor of differences between CLIL and non-CLIL students in C-test scores and listening comprehension test scores. CLIL students showed significantly higher motivation than non-CLIL students. Following Mearns et al. (2017), the higher motivation of CLIL students is not a result of CLIL instruction, but CLIL students are inherently more motivated. Additionally, Dallinger et al. (2016) revealed that CLIL-teachers' higher enthusiasm for teaching results from the greater need to prepare CLIL lessons and seems to positively affect their instructional quality in history. However, no significant differences could be found in instructional quality in English. No such data were available in KERMIT. Therefore, we could not consider these aspects in the present research, but they should be included in future studies. Besides, we considered prior achievement by matching students in terms of their performance in listening comprehension. However, we also analyzed a reading comprehension test in KERMIT 7. Thus, it is possible that part of the prior achievement was not considered in propensity

score matching due to the different English tests. Part of the preparation effect on reading comprehension could be a selection effect. Other influences besides preparatory lessons and CLIL could also cause the effects in grades seven and nine. However, due to the imitation of a randomized design by propensity score matching and the controlled covariates, such effects produced by uncontrolled variables seem relatively unlikely. An experimental design is not possible but would be useful to create truly randomized CLIL and non-CLIL groups as it would lead to even more meaningful results on the CLIL effect.

As CLIL programs in other countries might differ in terms of selection, preparation, and implementation, the results of this study do not generally apply to all CLIL programs. For example, some countries did not offer preparatory lessons (e.g., the Netherlands), used a less selective approach than Germany (e.g., Spain), and/or the design of CLIL programs varied (e.g., in terms of the amount of CLIL lessons) (Goris, Denessen, & Verhoeven, 2013; Goris et al., 2019). Thus, biases in selection and preparation are not equally likely sources of influence on the differences in English proficiency of CLIL and non-CLIL students across countries. To study the effects of selection, preparation, and CLIL in other school systems, similar research is required in other countries. Further, differences in CLIL programs occur not only between countries but also between schools or classes within Germany. Although the present study does not provide any information on the exact CLIL conditions in the respective schools or classes, it is likely that there are differences in the number of CLIL lessons, teacher qualifications, or CLIL subjects between schools and classes in Hamburg. To clarify how uniformly English language proficiency develops in CLIL classes in our sample, we estimated the ICC between CLIL classes in English reading comprehension in grade nine after matching. The small ICC between CLIL classes ($ICC = 0.12$ [0.03; 0.21]) was below the ICC between non-CLIL classes ($ICC = 0.17$ [0.12; 0.23]) and contradicted systematically different effects of implementation variations, which justified an analysis of the CLIL effect across all classes. This may be due to the Hamburg school law specifying requirements for CLIL teaching, for example regarding the minimum number of CLIL lessons and subjects. Nevertheless, it would be interesting to investigate the effects of different CLIL conditions on English competencies in future studies.

6.5. Conclusions

The present study exceeded previous ones when disentangling selection, preparation, and CLIL-effects. Furthermore, large and representative samples were rarely used in the past. As a result, the benefits of CLIL, given its small and non-significant additive effect, seem to be missing. Instead, CLIL students are a positive selected group and the results indicate that simply raising the number of English lessons for all could have a greater effect than CLIL. We measured English performance by a reading comprehension test in KERMIT and by C-tests in LAU and KESS. These three studies revealed that CLIL compensated for the fading out-effect but had no significant additive effect beyond that. However, LAU, KESS, and KERMIT did not assess speaking or listening comprehension skills after the implementation of CLIL. Dallinger et al. (2016) found significant differences in the development of listening comprehension skills in favor of CLIL students. Following Dalton-Puffer (2008), CLIL positively affects listening comprehension skills by enlarging the number of different speakers which students are confronted with face-to-face. In addition, CLIL students often display greater fluency, quantity, and creativity when speaking in the foreign language and show the kind of higher risk-taking inclinlation, which is often associated with good language learners. Speaking skills may also be positively affected because students seem to lose their inhibitions to use the foreign language spontaneously for face-to-face interaction after a certain amount of time spent in CLIL lessons (Dalton-Puffer, 2008). As CLIL might have a greater effect on these domains and also on knowledge in the content subjects, future studies should focus on these effects. Dallinger et al. (2016) reported only non-significant effects of CLIL on content subject learning in history. Further studies regarding these skills and subjects, the effects of teacher characteristics and instructional quality, and historical developments are necessary to justify future decisions regarding the expansion of CLIL programs. If substantial advantages in favor of CLIL students were missing in future research, the benefit of the implementation of CLIL would have to be questioned from an empirical point of view. Without an empirical basis, over-optimistic reports of the CLIL effect on English proficiency should be critically reviewed.

Author statement

Maja Feddermann: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft, Writing - Review & Editing. **Jürgen Baumert:** Formal Analysis, Writing - Review & Editing. **Jens Möller:** Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This paper uses data from the longitudinal study KERMIT. This data set was generated by the Free and Hanseatic City of Hamburg through the Ministry of Schools and Vocational Training between 2014 and 2018 and has been provided to us for further examinations of CLIL effects.

This research was supported by the German Science Foundation (DFG) [grant number MO 648/26-1].

Appendix A. Supplementary data

Supplementary material to this article is available. For more information see https://hdl.handle.net/21.11116/0000-000A-CAEA-B.

References

Admiraal, W., Westhoff, G., & de Bot, K. (2006). Evaluation of bilingual secondary education in The Netherlands: Students' language proficiency in English. *Educational Research and Evaluation, 1*(12), 75–93.

Alonso, E., Grisaleña, J., & Campo, A. (2008). Plurilingual education in secondary schools: Analysis of results. *International CLIL Research Journal, 1*(1), 36–49.

Behörde für Schule und Berufsbildung. (2011). [Authority for school and vocational training]. In *LAU – Aspekte der Lernausgangslage und der Lernentwicklung. Klassenstufen 5, 7 und 9 [LAU – aspects of the initial learning situation and learning development. Grades 5, 7 and 9]. HANSE – Hamburger Schriften zur Qualität im Bildungswesen: Bd. 8. Münster: Waxmann*.

Behörde für Schule und Berufsbildung. (2012). [Authority for school and vocational training]. In *LAU – Aspekte der Lernausgangslage und der Lernentwicklung. Klassenstufen 11 und 13 [LAU – aspects of the initial learning situation and learning development. Grades 11 and 13]. HANSE – Hamburger Schriften zur Qualität im Bildungswesen: Bd. 9. Münster: Waxmann*.

Behörde für Schule und Berufsbildung. (2014a). [Authority for school and vocational training]. In *Fremdsprachenunterricht im Schuljahr 2014/15 [Foreign language instruction during the 2014/15 school term]*. Retrieved from https://epub.sub.uni-hamburg.de/epub/volltexte/2014/33595/pdf/bbs_br_fremdsprachenunterricht_2014_15_.pdf. (Accessed 29 January 2021).

Behörde für Schule und Berufsbildung. (2014b). [Authority for school and vocational training]. In *Schulaufsichtliche Weisung für bilinguale Zweige an weiterführenden allgemeinbildenden Schulen in Hamburg [School supervision guidelines for bilingual branches at secondary general schools in Hamburg]*. Retrieved from http://www.schulrechthamburg.de/jportal/portal/bs/18/page/sammlung.psml?doc.hl=1&doc.id=VVHA-VVHA000000199&documentnumber=1&numberofresults=1&doctyp=vvhhschulr&showdoccase=1&doc.part=F&paramfromHL=true. (Accessed 29 January 2021).

Behörde für Schule und Berufsbildung. (2020). [Authority for school and vocational training]. In *Fremdsprachenunterricht im Schuljahr 2020/21 [Foreign language instruction during the 2020/21 school term]*. Retrieved from https://welcome.hamburg.de/contentblob/64460/552fb658fb4d53f88b70b65a230b7935/data/bbs-br-fremdsprachenunterricht.pdf. (Accessed 29 January 2021).

Bos, W., Bonsen, M., & Gröhlich, C. (Eds.). (2009). *KESS 7 - Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7 [KESS 7 - competencies and attitudes of students at schools in Hamburg at the beginning of*

grade 7]. HANSE - Hamburger Schriften zur Qualität im Bildungswesen. Münster: Waxmann.
Bos, W., & Gröhlich, C. (Eds.). (2010). *KESS 8 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8 [KESS 8 – competencies and attitudes of students at the end of grade 8]. HANSE - Hamburger Schriften zur Qualität im Bildungswesen*. Münster: Waxmann.
Bos, W., & Pietsch, M. (Eds.). (2006). *KESS 4 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen [KESS 4 - competencies and attitudes of students at the end of grade 4 in Hamburg elementary schools]. HANSE - Hamburger Schriften zur Qualität im Bildungswesen*. Münster: Waxmann.
Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x
Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge: Cambridge Univ. Press.
Dallinger, S., Jonkmann, K., Holm, J., & Fiege, C. (2016). The effect of content and language integrated learning on students' English and history competences – killing two birds with one stone? *Learning and Instruction, 41*, 23–31.
Dalton-Puffer, C. (2008). Outcomes and processes in content and language integrated learning (CLIL): Current research from Europe. In W. Delanoy, & L. Volkmann (Eds.), *Future perspectives for English language teaching*. Heidelberg: Carl Winter.
Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics, 31*, 182–204. https://doi.org/10.1017/S0267190511000092
De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of- school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition, 23*(1), 171–185. https://doi.org/10.1017/S1366728918001062
European Commission. (1978). *Education action programme at community level. The teaching of languages in the community*. Retrieved from http://aei.pitt.edu/3791/1/3791.pdf. (Accessed 29 January 2021).
European Commission. (1995). *White paper on education and training. Teaching and learning – towards the learning society*. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/d0a8aa7a-5311-4eee-904c-98fa541108d8/language-en. (Accessed 29 January 2021).
European Commission. (2004). *Promoting language learning and linguistic diversity: An action plan 2004-06*. Luxembourg: Office for Official Publications of the European Communities.
European Parliament. (1983). Resolution concerning language teaching in the community. *Official Journal of the European Communities, 26*(C 68), 105.
Eurydice. (2006). *Content and language integrated learning (CLIL) at school in Europe. EURYDICE Survey*. Brüssel: Eurydice.
Fäcke, C. (2021). foreign language didactics and foreign language education since 1945. *European Journal of Applied Linguistics, 9*(1), 1–19. https://doi.org/10.1515/eujal-2020-029
Feddermann, M., Baumert, J., & Möller, J. (under review). A replication study to assess CLIL effects on second language learning in Germany: More than selection and preparation effects? International Journal of Bilingual Education and Bilingualism. Submitted for publication.
Feddermann, M., Möller, J., & Baumert, J. (2021). Effects of CLIL on second language learning: Disentangling selection, preparation, and CLIL-effects. *Learning and Instruction, 74*, 101459. https://doi.org/10.1016/j.learninstruc.2021.101459
Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., et al. (2014). Methods for constructing and assessing propensity scores. *Health Services Research, 49*(5), 1701–1720. https://doi.org/10.1111/1475-6773.12182
Genesee, F. (2014). Is early second language learning really better? Evidence from research on students in CLIL programs. *Babylonia, 1*(14), 26–30.
Goris, J., Denessen, E., & Verhoeven, L. (2013). Effects of the content and Language integrated learning approach to EFL teaching: A comparative study. *Written Language & Literacy, 16*(2), 186–207. https://doi.org/10.1075/wll.16.2.03gor
Goris, J., Denessen, E., & Verhoeven, L. (2019). Effects of content and language integrated learning in Europe. A systematic review of longitudinal experimental studies. *European Educational Research Journal, 18*(6), 675–698. https://doi.org/10.1177/1474904119872426
Hanesová, D. (2015). History of CLIL (2015). In S. Pokrivčáková, et al. (Eds.), *CLIL in Foreign Language education: E-Textbook for Foreign Language teachers*. Nitra: Constantine the Philosopher University in Nitra. https://doi.org/10.17846/CLIL.2015.7-16.
Institut zur Qualitätsentwicklung im Bildungswesen. (2021). *VERA – Ein Überblick. [VERA – an overview]*. Retrieved from https://www.iqb.hu-berlin.de/vera. (Accessed 29 January 2021).
Klieme, E. (Ed.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie [Instruction and Competence Acquisition in German and English. Findings of the DESI Study]*. Weinheim, Basel: Beltz.
Köller, O., Leucht, M., & Pant, H. A. (2012). Effekte bilingualen Unterrichts auf die Englischleistungen in der Sekundarstufe 1 [The effects of CLIL on the English performance in lower secondary schools]. *Unterrichtswissenschaft, 40*(4), 334–350.
Königs, F. G. (2010). Zweitsprachenerwerb und Fremdsprachenlernen: Begriffe und Konzepte [Second language acquisition and foreign language learning: Terms and concepts]. In H.-J. Krumm, C. Fandrych, B. Hufeisen, & C. Riemer (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft: Vol. 35.1. Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch*. Berlin/New York: De Gruyter.
Krashen, S. D. (1985). *The input hypothesis. Issues and implications*. London: Longman.
Krashen, S. D., & Terrell, T. D. (1998). *The natural approach. Language acquisition in the classroom*. Hemel Hempstead: Prentice Hall.
Kultusministerkonferenz. (2013a). [Standing conference of the ministers of education and cultural affairs]. In *Bericht "Fremdsprachen in der Grundschule – Sachstand und Konzeptionen 2013" (Beschluss der Kultusministerkonferenz vom 17.10.2013). [Report "Foreign languages in primary school – state of affairs and concepts 2013" (resolution of the Standing Conference of the Ministers of Education and Cultural Affairs from 17.10.2013)]*. Retrieved from https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2013/2013_10_17-Fremdsprachen-in-der-Grundschule.pdf. (Accessed 29 January 2021).
Kultusministerkonferenz. (2013b). [Standing conference of the ministers of education and cultural affairs]. In *Bericht „Konzepte für den bilingualen Unterricht – Erfahrungsbericht und Vorschläge zur Weiterentwicklung" (Beschluss der Kultusministerkonferenz vom 17.10.2013) [Report "concepts for bilingual teaching – experience report and further development proposals" (resolution of the Standing Conference of the Ministers of Education and Cultural Affairs from 17.10.2013)]*. Retrieved from https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2013/201_10_17-Konzepte-bilingualer-Unterricht.pdf. (Accessed 29 January 2021).
Lasagabaster, D. (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal, 1*(1), 31–42.
Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of second language acquisition*. San Diego, CA: Academic Press.
Lücken, M., Thonke, F., Pohlmann, B., Hoffmann, H., Golecki, R., Rosendahl, J., & Poerschke, J. (2017). *KERMIT – kompetenzen ermitteln [KERMIT – identifying competencies]*. Retrieved from https://www.kermit-hamburg.de/index.php?action=download&id=1300. (Accessed 29 January 2021).
Lüdtke, O., & Robitzsch, A. (2020). Commentary regarding the section "modelling the effectiveness of teaching quality". Methodological challenges in assessing the causal effects of teaching. In A.-K. Praetorius, J. Grünkorn, & E. Klieme (Eds.), *Empirische Forschung zu Unterrichtsqualität: Theoretische Grundfragen und quantitative Modellierungen (Zeitschrift für Pädagogik. 66. Beiheft*. Weinheim, Basel: Beltz Juventa.
Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung [How to deal with missing values in psychological research]. *Psychologische Rundschau, 58*(2), 103–117. https://doi.org/10.1026/0033-3042.58.2.103
Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology, 110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016
Marsh, D. (1994). Bilingual education & content and language integrated learning. In *Paris: International association for Cross-cultural communication, language teaching in the Member states of the European union (Lingua)*, university of Sorbonne.
May, P. (2006). Englisch-Hörverstehen am Ende der Grundschulzeit [English listening comprehension at the end of primary school]. In W. Bos, & M. Pietsch (Eds.), *HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 1. KESS 4 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen (pp. 203–224)*. Münster: Waxmann.
Mearns, T., de Graaff, R., & Coyle, D. (2017). Motivation for or from bilingual education? A comparative study of learner views in The Netherlands. In *International Journal of bilingual education and Bilingualism. Advance online publication*. https://doi.org/10.1080/13670050.2017.1405906
Möller, J., Fleckenstein, J., Hohenstein, F., Preusler, S., Paulick, I., & Baumert, J. (2017). Varianten und Effekte bilingualen Lernens in der Schule [Variations and impacts of CLIL in school]. *Zeitschrift für Erziehungswissenschaft, 21*(1), 4–28. https://doi.org/10.1007/s11618-017-0791-x
Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus user's guide (8th ed.). Los Angeles, CA: Muthén & Muthén.
Nold, G., Hartig, J., Hinz, S., & Rossa, H. (2008). Klassen mit bilingualem Sachfachunterricht: Englisch als arbeitssprache [CLIL classes: English as a language of instruction]. In E. Klieme (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie*. Weinheim, Basel: Beltz.
Peters, E. (2018). Approaches to learning, testing, and researching L2 vocabulary. *ITL - International Journal of Applied Linguistics, 169*(1), 142–168. https://doi.org/10.1075/itl.00010.pet
Pishgar, F., & Greifer, N. (2020). *Package 'MatchThem'. [Computer software manual] (version 0.9.3)*.
Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. The *British Journal of Educational Psychology, 82*(4), 647–671. https://doi.org/10.1111/j.2044-8279.2011.02051.x
Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.
Ruiz de Zarobe, Y. (2008). CLIL and Foreign Language learning: A longitudinal study in the Basque country. *International CLIL Research Journal, 1*(1), 60–73.
Rumlich, D. (2018). Englischnoten und globale englische Sprachkompetenz in bilingualen Zweigen [English grades and global English language proficiency in CLIL programs]. *Zeitschrift für Erziehungswissenschaft, 21*(1), 29–48. https://doi.org/10.1007/s11618-017-0801-z
Schafer, J. L., & Zhao, J. H. (2018). Package pan: Multiple imputation for multivariate panel or clustered data. *[Computer software manual] (version 1.6)*.
Sundqvist, P., & Olin-Scheller, C. (2013). Classroom vs. extramural English: Teachers dealing with demotivation. *Language and Linguistics Compass, 7*(6), 329–338. https://doi.org/10.1111/lnc3.12031
Sundqvist, P., & Sylv´en, L. K. (2016). *Extramural English in teaching and learning: From theory and research to practice*. London: Palgrave Macmillan UK.

Surmont, J., van de Craen, P., Struys, E., & Somers, T. (2014). Evaluating a CLIL student: Where to find the CLIL advantage. In R. Breeze, C. Llamas Saíz, C. Martínez Pasamar, & C. Tabernero Sala (Eds.), *Integration of theory and practice in CLIL*. Amsterdam: Brill/Rodopi.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass, & C. G. Madden (Eds.), *Input in second language acquisition*. Rowley, MA: Newbury House.

Sylvén, L. K. (2013). CLIL in Sweden – why does it not work? A metaperspective on CLIL across contexts in Europe. *International Journal of Bilingual Education and Bilingualism, 16*(3), 301–320. https://doi.org/10.1080/13670050.2013.777387

Sylvén, L. K., & Olsson, E. (2015). Extramural English and academic vocabulary. A longitudinal study of CLIL and non-CLIL students in Sweden. *Apple - Journal of applied language studies, 9*(2), 77–103.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van Buuren, S., & Groothuis-Oudshoorn, K. (2020). *Package 'mice'. [Computer software manual] (version 3.12.0)*.

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

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399. https://doi.org/10.1002/sim.4067