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

# Effects of CLIL on second language learning: Disentangling selection, preparation, and CLIL-effects

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

[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. Möller), jmpobaumert@mpib-berlin.mpg.de (J. Baumert).

**Abstract**

The positive effects of Content and Language Integrated Learning (CLIL) on the development of students' foreign language skills may have been overestimated by previous studies, since most studies failed to consider selection and preparation effects appropriately. Therefore, the present study used complete survey data from a 2002–2007 cohort to investigate English skill development of $N = 385$ German CLIL and $N = 5{,}578$ non-CLIL grammar school students from grade seven ($M = 12.46$, $SD = 0.53$) to grade eight ($M = 14.46$, $SD = 0.53$). Firstly, we found significant selection effects for performance in primary school, sociodemographic variables, and cognitive abilities. Secondly, after propensity score matching, data revealed significant preparation effects of additional second language instruction for the CLIL students. When controlling selection and preparation effects, only a small and non-significant CLIL effect occurred measured by a C-test. We discuss the results with regard to previous inconsistent findings.

**Keywords**

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

## 1. Introduction

In Content and Language Integrated Learning (CLIL) programs, students learn a foreign language through content-based instruction (Genesee, 2014). This entails that one or more content subjects are taught in a foreign language, while in the other subjects the common language is used as the language of instruction. High-level policymaking and grass-roots actions led to the implementation of CLIL, which was also supported by teachers and parents (Dalton-Puffer, 2011). CLIL has a major contribution to make to the European Union's language learning goals, which consist of the access to lifelong language learning, the improvement of language teaching, and the creation of a language-friendly environment. To reach these goals, the European Union encourages its citizens to learn other European languages (European Commission, 2005). Within the context of European educational initiatives in the 1990s and 2000s, CLIL programs increased enormously (Rumlich, 2018). But, this development is surprising, given the large gaps in empirical research on the effects of CLIL (Goris, Denessen, & Verhoeven, 2019).

Several studies with convenience samples regarding the effects of CLIL were presented in the last years. However, there is a lack of longitudinal studies with a representative and sufficiently large sample taking into account selection and preparation effects. Selection effects occur because students choose a CLIL program themselves or are selected by schools. CLIL students, therefore, are likely to possess, for example, higher prior knowledge in English, higher cognitive abilities, and a more favorable family background (Dallinger, Jonkmann, Holm, & Fiege, 2016). Preparation effects may result from the increased number of English lessons for future CLIL students to prepare for CLIL instruction, starting in later grades (Rumlich, 2018). Ignoring these aspects may lead to biased results (Goris et al., 2019). Therefore, the present study analyzed CLIL- and non-CLIL students' English competencies from grade four to grade eight in a comprehensive large German sample controlling for preparation and selection effects.

## 2. Content and Language Integrated Learning (CLIL)

In Germany, the first CLIL programs were introduced in 1969. These were German/French branches. The number of CLIL programs grew and in the 1970s and 1980s also German/English branches were established. In the 1990s, CLIL was expanded to other languages or types of school (Eurydice, 2006). The data of this study originate from Hamburg. At the beginning of the KESS study in 2002/2003, seven grammar schools offered an English CLIL program in Hamburg. 2009/2010 the number

increased to 15 grammar schools. Ten years later, 22 of 73 grammar schools in Hamburg offered an English CLIL program (Behörde für Schule und Berufsbildung, 2009, 2020).

In CLIL programs, the subject content of a class is taught in a foreign language, mostly English. Dalton-Puffer (2011) explicitly used the term „CEIL, or content-and-English integrated learning“ (p. 183). However, there are also French, German, Spanish, or Italian CLIL programs (see Eurydice, 2006). In this paper, unless explicitly stated otherwise, the foreign language used in CLIL is English. The curriculum is maintained in the subject, which is why CLIL is internationally referred to. In this case, instruction in the foreign language takes place in one to three content subjects while the language of instruction of the other subjects is the common language. In Germany, the content subjects which are taught in a foreign language are mostly history, geography, biology, music, or politics (Möller et al., 2017; Rumlich, 2018). In Germany, the actual CLIL instruction is usually implemented once students have acquired literacy skills in their common language, which is usually at the secondary level. From a didactic point of view, the advantage of this teaching method is that the foreign language can be used in an authentic context without having to make it a separate lesson. Moreover, in most cases, foreign language learning is additionally supported by explicit foreign language teaching (Dalton-Puffer, 2011).

From 2002 to 2007, Hamburg students received English lessons starting in grade three in primary school. In general, English was taught two hours per week. CLIL students received a minimum of six hours of English teaching in grades five and six. In grade six, they could also receive five hours of regular English teaching and one hour of foreign language instruction in a content subject. This is, CLIL students received one to two more English lessons per week than non-CLIL students. From grade seven on, a minimum of three hours divided into one to three content subjects were taught in the foreign language in addition to the regular English lesson. The goal of CLIL programs in Hamburg is to achieve an approximate bilingualism. CLIL instruction aims to promote a deeper cultural and linguistic understanding. It teaches the skills and content of the subject as well as foreign language and intercultural communication skills. CLIL instruction was generally given by teachers who were qualified to teach both the subject and the foreign language. The teaching qualification for the foreign language could be waived for native speakers and teachers with a proven linguistic qualification at the level of at least C1 of the Common European Framework of Reference for Languages. The structure and design of the CLIL programs in Hamburg did not change since 2007, but since 2011/2012, all primary schools have offered English lessons starting in grade one (Behörde für Schule und Berufsbildung, 2014, 2020, 2009; Kultusministerkonferenz, 2006).

In general, CLIL is characterized by a dual focus on language acquisition and content knowledge. However, depending on the country, other objectives are also pursued. According to Eurydice (2006), these objectives vary between socioeconomic, sociocultural, linguistic, and educational objectives. In Germany, CLIL aims to prepare students for life in an internationalized society and to offer them better job prospects on the labor market (socioeconomic objectives). In addition, CLIL aims to teach German students values of tolerance and respect for other cultures (sociocultural objectives). The objectives of developing linguistic communication skills and motivating students to learn a language by using it for real practical purposes are pursued in almost every country, including Germany (linguistic objectives). Educational objectives refer to the promotion of subject-related knowledge and learning ability by using a different and innovative approach. This is not the main focus in Germany. Due to these varying goals, CLIL subject pedagogies also differ depending on the country (Van Kampen, Mearns, Meirink, Admiraal, & Berry, 2018). A framework was designed by Coyle, Hood, and Marsh (2010) to guide the fundamental elements of CLIL pedagogies from a holistic perspective. The 4 Cs Framework contains the interlinked elements of content, cognition, and communication that are embedded within a cultural context. Content is about the subject or theme being learned, cognition means the cognitive processing required for learning activities, communication refers to the language learning and using, and culture is about how teaching promotes the intercultural and interpersonal understanding as well as how to apprentice students into the discourses, genres, and approaches specific to each subject (Van Kampen et al., 2018). Thus, CLIL is not just about language acquisition and content knowledge.

Theories on the acquisition of a second language suggest positive effects of CLIL on English skills. The input hypothesis (Krashen, 1985) and the interaction approach (Gass & Mackey, 2015) both assume that language is acquired through the presence of comprehensible input and opportunities to interact with the foreign language and receive feedback for it. CLIL classes in particular therefore offer sufficient opportunities as the foreign language is used to a much greater extent than in regular classes, thus, encouraging students to use the foreign language as often as possible (Eurydice, 2006). The natural approach (Krashen & Terrell, 1998) states that competence in a foreign language can most effectively be developed by communicating in the foreign language in real situations as happens in CLIL programs with its focus on meaning and not on form (Dalton-Puffer, 2008). The foreign language can be used on a voluntary basis, and the authenticity of the language use in CLIL classrooms is high. The topics are determined by the curriculum, and the foreign language serves as a means of communication. As the focus in CLIL is not primarily on the correct language use, the pressure for avoiding mistakes decreases (Surmont, van de Craen, Struys, & Somers, 2014). From these theories, it can be concluded that CLIL students benefit from CLIL instruction, especially in listening and reading comprehension, since the input they receive is often in written or oral form. Speaking and writing can also profit from CLIL instruction but to a lesser degree as students are encouraged, not forced, to communicate and write in the foreign language (Dallinger et al., 2016).

In Germany, the actual CLIL instruction, which only begins in grade seven of the secondary level, is often preceded by preparatory instruction for CLIL students with an increased number of English lessons in grades five and six. This is to ensure that the students have acquired sufficient English competencies at the beginning of the CLIL program to be able to follow the CLIL lessons (Köller, Leucht, & Pant, 2012). This, nevertheless, makes it more difficult to make statements about the effect of CLIL instruction. Studies analyzing CLIL effects without controlling such preparation effects may lead to biased results (Rumlich, 2018). A further bias may be due to the input selectivity of these programs. CLIL programs are attractive for students with positive linguistic, cognitive, and socioeconomic backgrounds (Möller et al., 2017). Secondary schools in general use selection criteria for admittance to a CLIL class involving enhanced motivation and above-average English skills (Goris et al., 2019). Mearns, de Graaff, and Coyle (2017) reported that Dutch CLIL students showed a significantly higher level of motivation at the beginning of CLIL compared to non-CLIL students and that CLIL had little influence on motivation. Thus, motivation predates entry into a CLIL program and may therefore affect selection. In addition, most secondary schools with CLIL programs also offer a monolingual branch. In these parallel classes, the use of the foreign language is limited to explicit foreign language instruction. The resulting choice leads to the initial selectivity as Dallinger et al. (2016) revealed.

The most recent comprehensive overview on the evaluation of the

contribution of CLIL to competencies in English during the past 20 years was done by Goris et al. (2019). They included only those 21 studies that contained a measure of one or more English as a foreign language (EFL) skills, were conducted in the last 20 years and written in English, used students in mainstream primary or secondary education in a European country, and, most importantly, used a longitudinal design. The findings of Goris et al. (2019) do not unequivocally support the assumption that CLIL students will develop more EFL proficiency than non-CLIL students in the same time. Whereas in Germany and other northern European countries mostly null effects were found, in Spain significant effects were more frequent. Following Goris et al. (2019), the Spanish ideology of CLIL was to provide better EFL learning opportunities for all, whereas in Germany or the Netherlands CLIL is more elitist and highly selective. Verspoor et al. (2015) investigated the effectiveness of CLIL in the Netherlands. They analyzed the development of English proficiency of two cohorts ($N = 399$) at Dutch high schools during one year. English proficiency was tested three times in each cohort and assessed by means of the combined scores of a vocabulary test and a productive informal writing task. Regarding the selective approach in the Netherlands, they reported that CLIL students scored significantly higher on the scholastic aptitude test, showed significantly higher scores in initial English proficiency, and were more motivated than non-CLIL students. These covariates as well as the English performance at measure two and the out-of-school contact explained 60.9% of the variance in the students' proficiency at cohort one and 62.8% at cohort three. Finally, they revealed that CLIL students made a significantly higher gain than non-CLIL students in year one when controlling these covariates. However, in year three they did not improve relatively more than non-CLIL students in English proficiency when controlling the covariates. They kept the gains they had made but they did not extend their lead. According to Goris et al. (2019), EFL proficiency in Spain and other southern European countries was rather low in contrast to high EFL-proficiency countries like Germany and other northern European countries. They considered these differences as an explanation for the contradictory findings, i.e., the positive effects of CLIL in southern European countries like Spain or Italy, and the smaller effects in northern European countries like Sweden, the Netherlands, and Germany. Goris et al. (2019) claimed that selection and preparation effects must be considered, but criticized that the initial EFL level as well as the preparation effect were not consistently taken into account throughout the studies leading to biases in CLIL.

As this study used a German sample to analyze the CLIL effect on English performance, the empirical findings for Germany are presented in more detail. In this study, the KESS data were re-analyzed. Prior studies and results were documented in mainly descriptive reports. One of these reports was written under the direction of Bos (Bos, Bonsen, & Gröhlich, 2009; Bos & Gröhlich, 2010; Bos & Pietsch, 2006). With the focus on CLIL effects, this report estimated selection effects in grade four and the difference between CLIL and non-CLIL classes in grade seven without considering selection effects. However, the effect in grade seven does not reflect the effect of CLIL lessons, as they did not start until grade seven. Therefore, $N = 218$ CLIL and $N = 1931$ non-CLIL grammar school students and the results of the listening comprehension test in grade four as well as the results of the C-test in grade seven were used. The analysis compared the English proficiency of CLIL and non-CLIL students and already revealed advantages in English performance in favor of later CLIL-students of $\delta = 0.5$ in grade four. After two years of increased English instruction, CLIL students could once again significantly improve their performance compared to non-CLIL students. At the beginning of grade seven, CLIL students showed an advantage in English performance of $\delta = 0.98$ compared to non-CLIL students (May 2009). In contrast to this analysis, the present study also used the data of grade eight and could thus estimate the CLIL effect. Additionally, the present study could separately estimate preparation effects by matching students at the end of grade four in terms of prior achievement, sociodemographic variables, and cognitive abilities. Therefore, the present study goes far beyond the former analyses.

Four German studies controlled selection effects. The DESI-study (Deutsch Englisch Schülerleistungen International) examined a sample of $N = 1945$ students, who were representative of students in Germany, and formed a monolingual comparison group based on German performance, socioeconomic status measured by the highest international socioeconomic index of occupational status (HISEI), basic cognitive ability, educational pathway, first language, and gender. The study showed advantageous English skills in C-test, listening and reading comprehension, grammar, and creative writing. However, the increase of English performance in C-test was comparable for CLIL and non-CLIL students (Nold, Hartig, Hinz, & Rossa, 2008). A C-test consists of texts in which halves of words are deleted according to certain principles. Participants are required to complete the words. The C-test measures the ability to apply and integrate vocabulary, grammar, and textual knowledge (Harsch & Schröder, 2007 ). Estimations of selection effects were not included, and, statements on preparation effects were not possible because the students were tested at the beginning and at the end of grade nine (Nold et al., 2008). Köller et al. (2012) examined $N = 9867$ students from a representative sample controlling for parents' school-leaving qualifications, socioeconomic status, and linguistic competence in German. As a result, the English performance of CLIL students in reading comprehension and listening comprehension was significantly higher than the English performance of non-CLIL students. Köller et al. (2012) found advantages in the order of half a standard deviation in favor of CLIL students. However, selection effects were not explicitly estimated. Statements on preparation effects were not possible, as the students were tested only in grade nine. Dallinger et al. (2016) examined a convenience sample of $N = 1806$ eighth graders and controlled prior achievement, general abilities, motivation, demographics, classroom composition, and instructional quality. They found significant differences in the results of the listening comprehension test in favor of CLIL students. The C-test scores, however, showed no significant differences. They did not make any statements on effect sizes of selection or preparation. However, they could show that CLIL students possessed significantly better prior achievement and motivation, higher cognitive abilities, as well as higher socioeconomic status. In addition, CLIL teachers reported more enthusiasm for teaching than non-CLIL teachers. Rumlich (2018) carried out the only German study with statistical control of selection and preparation effects in his longitudinal data. He found significant differences between the two groups of approximately $N = 1000$ sixth and eighth graders from a convenience sample in scores on C-tests. When age, gender, first language, and initial English test performance at the beginning of grade six were controlled, the increase in performance of the CLIL group at the end of grade eight compared to the control groups was no longer significant. In contrast to the previously mentioned studies, he could contradict the assumption of a positive net effect of CLIL, but did not make any precise statements on the effect sizes of selection and preparation.

To go beyond selection effects, there remains a lack of studies estimating selection and preparation effects, using a representative sample, and considering preparation effects in those CLIL programs preparing CLIL students by offering more EFL lessons to them before starting the CLIL instruction. References to preparatory lessons are rarely mentioned in the studies. In Germany and Italy, the use of additional EFL lessons is common. In the Netherlands, however, no preparatory instruction is offered (Goris et al., 2019). The lack of estimation and consideration of selection and preparation effects in representative samples illustrates the need for the present study, which used a propensity score matching, a large and comprehensive survey sample, and the longitudinal data structure from fourth to eighth grade, to make well-founded statements about the effect of selection, preparation, and the effect of CLIL on English performance. Thus, the present study exceeds previous ones in terms of content and method.

## 3. The present study

In this study, CLIL and non-CLIL students' English competencies were examined and compared. The previous empirical findings showed the importance of longitudinal studies, along with the consideration of background variables. In a longitudinal study, the same subjects are questioned and tested over a period of time. This study design enables statements on performance development and the consideration of prior knowledge from earlier measurement points, which is an important predictor for future English performance (Goris et al., 2019). We estimated the effect of CLIL using data from a full survey of Hamburg students, while considering and estimating selection and preparation effects by longitudinally controlling the English performance at the end of primary school and at the beginning of grade seven, prior achievement, sociodemographic variables, and cognitive abilities.

We assumed that selection effects and preparation effects would show up. More importantly, when we controlled selection and preparation effects, we expected only small positive longitudinal effects on English performance for CLIL students.

### *3.1. Sample*

The data used in this analysis came from the longitudinal school performance study "Competencies and Attitudes of Students" (KESS, see Bos et al., 2006). The data were obtained from 66 Hamburg grammar schools (seventh and eighth grade) and 256 Hamburg primary schools (fourth grade). Among those grammar schools, six offered an English CLIL program in addition to the regular monolingual instruction, two schools provided only English CLIL instruction, and 55 provided only regular monolingual instruction. Since CLIL was only available at grammar schools (highest secondary school track with Abitur as exit exam in 12th or 13th grade when students are on average between 18 and 19 years old), only this type of school was considered. The aim of this Hamburg panel study is not only to present the learning status and performance development of an entire school cohort but also to make statements about the factors that can influence the development of scholastic achievement. The complete survey of Hamburg school students took place at a total of five points in time. The data used in this paper originate from the first three measurement points. The KESS study began at the end of the school year 2002/2003 with the first survey in June 2003 at the end of grade four (KESS 4). 263 primary schools in Hamburg took part at this measurement point (Bos et al., 2006). The longitudinal study was continued with the second survey of the same cohort right after the beginning of the school year 2005/2006 in September 2005 in the seventh grade (KESS 7). 174 schools with a lower secondary level were tested at the beginning of grade seven (Bonsen, Bos, Gröhlich, & Rau, 2009). In grades five and six, CLIL students received one to two additional English lessons to prepare for CLIL, starting in grade seven. The CLIL subjects were taught at least three hours per week. The third survey wave, KESS 8, took place at the end of the school year 2006/2007 in spring 2007, testing eighth graders from 170 schools (Bos, Gröhlich, Guill, Scharenberg, & Wendt, 2010).

The combined data set contained $N = 6020$ grammar school students from KESS 7 and 8 with information about CLIL participation (see Table 1). In a first step, students who attended CLIL at only one measurement point were excluded ($N = 9$). After the data from KESS 4 were added, students who participated in a non-English CLIL program ($N = 45$) and all cases without valid values ($N = 3$) were excluded from the analysis. As a result, $N = 5963$ students remained in the data set. It consisted of 45.4% boys and 50.3% girls, the missing rate for gender information was 4.5%. $N = 5524$ students participated in KESS 4, $N = 5530$ in KESS 7, and $N = 5273$ in KESS 8. $N = 5520$ students took part in KESS 4 and KESS 7, $N = 4837$ in KESS 4 and KESS 8, and $N = 4840$ in KESS 7 and KESS 8. $N = 4833$ students participated at every measurement point. $N = 385$ students were taught in CLIL programs, and $N = 5578$ students were taught monolingual and received the usual English as a foreign language instruction.

**Table 1**
Merging and selecting data.

| Steps | $N$ |
|---|---|
| Grammar school students from KESS 7 and 8 with information about CLIL | 6020 |
| Students who continuously attended a CLIL or non-CLIL program | 6011 |
| Match KESS 4 data | 6011 |
| Students who attended an English CLIL or non-CLIL program | 5966 |
| Cases with at least one valid value | 5963 |

### *3.2. Instruments*

The Item-Response-Theory (IRT) is used for performance test scaling in all current national and international performance comparison studies. It was assumed that the abilities of individuals could be estimated using the applied tests. Based on the IRT, response probabilities were calculated for all individuals based on the answers to the test items. These probabilities were regarded as a function of student ability, which as a latent continuous person variable is the basis of the response behavior and could be deduced from the actual answers. Thus, the ability parameters estimated for each person on the basis of an explicitly formulated measurement model were used as measurements of student performance (Bos et al., 2010). The solution probability of an item depended on the person's ability and the item difficulty. Personal ability and item difficulty must be estimated using certain models (Bortz & Döring, 2002; Rentzsch & Schütz, 2009).

The IRT method made it possible to map student abilities and item difficulties on a common metric and, thus, to relate them directly to each other. If a student's ability exceeded the difficulty of an item, it was likely that the student would also solve items of lesser difficulty correctly. Ultimately, this could be used to determine whether students had achieved the desired learning goals (Bos et al., 2006).

In addition, IRT-based estimates of personal abilities allowed test scores to be determined on a common metric for all students, despite differences in the item quantities processed (Rauch & Hartig, 2012). This was possible by using anchor items, i.e., tasks that were used at several measurement points (Bos et al., 2010). In the KESS study, weighted likelihood estimates (WLEs) were estimated separately for each domain (mathematics, natural science, German, English) from the overall scores. The WLEs were used to estimate the personal ability, since they were less biased and were regarded as superior estimation method (Warm, 1989). This was carried out in R, using the package sirt (Robitzsch, 2019).

#### *3.2.1. Grade four*

For the required estimated item parameters, the items of all measurement points were scaled together and one-dimensionally without a background model. The one-parameter model (the solution probability of an item is dependent on personal ability and item difficulty) was chosen as the scaling model (Bos et al., 2010; Moosbrugger, 2012). The identification of the model was carried out by standardizing the personal abilities. Thereafter, these item parameters were used for the estimation of personal abilities.

*3.2.1.1. Cognitive ability.* To test the cognitive abilities of the students, the figural and verbal analogy subtests of the cognitive ability test (KFT; Heller & Perleth, 2000) were carried out in KESS 4. The figural analogy subtest consisted of 25 items to test the logical and spatial thinking. The students had eight minutes to finish the test. Each task started with a pair of figures or drawings that fitted together in a certain way. For each task, the students had to decide how the two figures were related to each other. There was also a third figure, which was the first figure of a second pair. Of the five figures on the right, the one that fitted the third

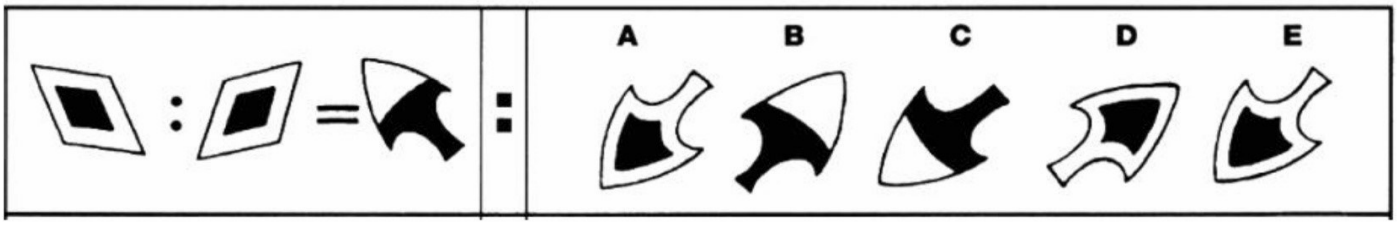


**Fig. 1.** Example of a task of the figural analogy subtest


**Fig. 2.** Example of a task of the verbal analogy subtest.

figure as well as the second figure to the first one had to be found. An example is shown in Fig. 1. The internal consistency is $\alpha = 0.93$ for the figural analogy subtest.

The verbal analogy subtest consisted of 20 items to test the verbal abilities. The students had seven minutes to finish the test. The verbal analogy subtest was similar to the figural analogy subtest. Each task of this test started with a pair of words that were related in a certain way. For each task, the students had to find out how the two words belonged together. There was also a third word, which was the first word of a new pair. From the five words in the line below, the word that fitted the third word as well as the second word to the first had to be chosen. An example is shown in Fig. 2. The internal consistency is $\alpha = 0.80$ for the verbal analogy subtest.

*3.2.1.2. Sociodemographic variables.* Sociodemographic variables (highest international socioeconomic index of occupational status (HISEI), gender, year of birth, language spoken at home, time spent in kindergarten/preschool, age at enrolment in primary school, total gross income per year, migration status (no parent/one parent/both parents with migration background), and recommended type of secondary school) as well as school grades in German language, German reading, life science, and mathematics were assessed through parent and student questionnaires (Bonsen et al., 2009). The German grading system comprises grades from 1 (excellent) to 6 (failed). Lower grades therefore indicate better performance.

*3.2.1.3. English listening comprehension.* A listening comprehension test (May 2006a) was used to assess English competencies in KESS 4. The test was divided into two parts, in which a native speaker read individual sentences and an entire story to the students in English. After a short break, both the individual sentences and the story were repeated. The first part consisted of twelve individual sentences in the form of questions. The students had to choose the correct answer out of four alternatives. For example, the native speaker asked “What can you eat?” and the answer alternatives were “water”, “bread”, “soap”, “spoon”. The answer alternatives were in German and therefore, answering the questions did not require reading skills in English. In the second part, a native speaker read a story to the students twice. Afterwards, they had to answer ten questions by choosing the right answer out of four alternatives. Additionally, for two questions the students had to write down a reason for their choice. The results of both parts combined resulted in an overall scale for assessing listening comprehension in English (May 2006a). The reliability of this scale was $\alpha = 0.62$. The English test’s WLE reliability ($Rel(WLE) > 0.70$) was satisfactory.

*3.2.1.4. German.* German performance in KESS 4 was assessed by using a reading comprehension test including tasks on texts from the LAU (Aspects of Learning Prerequisites and Learning Development) and PIRLS (Progress in International Reading Literacy Study) studies and an orthography test in the form of a dictation (Bos et al., 2006). For the orthography test, the HSP (Hamburg Writing Test) and the DoSE (Dortmund Writing Test) were used (May 2006b). The reading comprehension test consisted of three texts, with one text varying according to the test booklet. The texts were stories or factual texts. The students had 40 min to answer the 13 to 15 questions (depending on the test booklet) on the first text. For some questions, they had to choose the correct answer from four alternatives. For other questions, there were no given answer alternatives. In the second and third text, they had to choose the correct answer out of four alternatives. Here, students had 25 min to answer eleven questions, seven on the second text and four on the third text. Depending on the test version, the reliability of the reading comprehension test ranged from $\alpha = 0.75$ to $\alpha = 0.81$.

The orthography test consisted of two different spelling tests. On the one hand, sentences with 45 word gaps used from the DoSE were dictated from the test leader, which had to be filled in by all children at the same pace. On the other hand, a total of 42 words from the HSP were read aloud in the form of eight individual words and five sentences by the test leader and presented as picture impulses. The test leader read out all the individual words on a page before the students began. Finally, line drawings were shown as a reminder. The students could ask if they have forgotten a word. This allowed each student to work at their own pace and in the order they chose. Afterwards, the five sentences were read aloud individually. As a measure of orthographic competence, only whether the words were spelled correctly or incorrectly was taken into account (May 2006b). For both test versions, the reliability of the orthography test was $\alpha = 0.92$.

*3.2.1.5. Mathematics.* In KESS 4, 24 tasks from the fields of arithmetic, geometry, and word problems were used to estimate mathematical performances. The tasks derived from the PIRLS and LAU studies and were defined as multiple-choice tasks, except for one (Pietsch & Krauthausen, 2006). Depending on the test version, the reliability of this test ranged between $\alpha = 0.76$ and $\alpha = 0.78$.

*3.2.2. Grades seven and eight: English C-test*

A word completion test (C-test) (Hastings, 2002; Raatz & Klein-Braley, 1983) was used in KESS 7 and KESS 8 to measure general English language proficiency. The C-test is the standard test in large-scale student assessment studies to measure foreign language competence and is based on the reduced redundancy principle. In natural language use, different aspects contribute to the constitution of the meaning of a message. These redundancies ensure communication. The more competent a language user is, the less redundancies are needed for language processing. In addition to global comprehension, it also requires textual knowledge, knowledge of words, spelling, syntax, and grammar (Harsch & Schröder, 2007). “A C-test measures the ability to apply and integrate contextual, semantic, syntactic, morphological, lexical, and orthographic information and knowledge pertaining to a

particular written language" (Hastings, 2002, p. 66). High correlations of the C-test scores with school grades, teacher assessments of the students' language level, or the results of other language tests (e.g. the test of English as a foreign language (TOEFL)) as well as studies on construct validity makes it a valid instrument for assessing general language competence (Eckes & Grotjahn, 2016; Grotjahn, 2002). In the KESS 7 C-test, there were two booklet versions in which the different texts were edited without interruption. Both versions consisted of three texts, two with 24 items, and one with 26 items. The text with 26 items differed between the booklet versions. The students had seven minutes to complete each text. In KESS 8, three of the four texts from KESS 7 were used again. Also two booklet versions were available, but they were processed with breaks between the texts. Both versions consisted of two texts, one with 24 and one with 26 items. The text with 26 items differed between the booklet versions. As in KESS 7, the students had seven minutes to complete each text. Apart from the first and last sentence, only half of the letters of every fourth word were given and had to be completed by the students. In case of an odd number of letters, one letter less was given than was missing. An example of a sentence used in KESS 7 and KESS 8: "It was fee___-time for the monkeys a___ keeper went in___ the cage and f___ them". The internal consistency in KESS 7 was between $\alpha = 0.97$ and $\alpha = 0.98$, depending on the test version and between $\alpha = 0.92$ and $\alpha = 0.94$ in KESS 8. The anchor-item design of the C-test allows the English performance at both measurement points to be compared directly (Bos, Bonsen, Gröhlich et al., 2009 ; Nikolova & Ivanov, 2010). For this purpose, the data of all measurement points had to be scaled to a common metric. Therefore, the two-parameter model was used, which assumes that the probability of solving a task depends on the personal ability, the item difficulty, and the item discrimination (Moosbrugger, 2012). At first, the items of the C-test of all measurement points were scaled together and one-dimensionally with the school form and the measurement point information in the background model to estimate the item parameters. The identification of the model was carried out by standardizing the personal abilities. Thereafter, the item parameters were used for the estimation of personal abilities. The WLE reliability of the C-test was sufficient ($Rel(WLE) = 0.94$ in KESS 7; $Rel(WLE) = 0.91$ in KESS 8).

### *3.3 Analyses*

Missing data is a common problem in longitudinal studies. One recommended solution to handle this problem is multiple imputation (Lüdtke, Robitzsch, Trautwein, & Köller, 2007). In our present study, on average about 26% of the data were missing per variable. In grades seven and eight, the percentage of missing values varied between 14.57% and 55.11% in the English tests and between 8.77% and 14.81% in all other tests. In the tests of grade four, the percentage of missing values ranged between 25.78% and 42.34%, and in the biographical and socioeconomic variables, it ranged between 4.31% and 54.72%.

We used a multiple imputation to obtain a data set without missing values. This was necessary for the following propensity score matching. Due to significant differences in relevant covariates between CLIL and non-CLIL students (see Table 2), a propensity score matching was used. By estimating a propensity score, i.e., the probability of belonging to the treatment group, for every student and by matching CLIL and non-CLIL students with equal propensity scores, matching eliminated significant differences between the groups. Thus, the following ANCOVA and Difference-in-Differences analysis only used students with a matching partner.

A multiple imputation can be divided into three steps. First, each missing value was replaced by several plausible values with different values based on the information in the data set. This generated several complete data sets, which could then be analyzed in a second step using standard procedures. Finally, the results of the individual analyses were pooled using Rubin's (1987) rules. These were summarized by taking into account the uncertainty of the imputation. The uncertainty in replacing the missing values was therefore not taken into account within the respective data set, but by the differences in the analyses for the different data sets (Lüdtke et al., 2007). The multiple imputation was performed with the *R*-package mice (version 3.6.0, Van Buuren & Groothuis-Oudshorn, 2019). In order to consider the multi-level structure of the data, the class ID was used as a grouping variable, as well as

**Table 2**
Imbalance in predictor variables between CLIL- and non-CLIL students before matching (N = 5963).

| Predictor variable | CLIL ($N = 385$) | | Non-CLIL ($N = 5578$) | | | | |
|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *t* | *p* | % bias |
| *KESS 4 performance* | | | | | | | |
| English listening WLE | 122.3 | 29.42 | 112.62 | 28.35 | 3.76 | <.001 | 33.51 |
| German reading WLE | 0.59 | 1.12 | 0.15 | 1.17 | 6.24 | <.001 | 38.42 |
| German orthography WLE | 123.81 | 23.27 | 118.17 | 22.53 | 4.31 | <.001 | 24.63 |
| Mathematics WLE | 0.31 | 1.10 | 0.00 | 1.10 | 4.62 | <.001 | 28.18 |
| Grades in German Language | 1.88 | 0.51 | 2.13 | 0.59 | −7.98 | <.001 | −45.33 |
| Grades in German Reading | 1.68 | 0.59 | 1.94 | 0.59 | −6.87 | <.001 | −44.07 |
| Grades in Mathematics | 1.99 | 0.54 | 2.14 | 0.60 | −4.19 | <.001 | −26.28 |
| Grades in Science | 1.86 | 0.52 | 2.08 | 0.57 | −6.32 | <.001 | −40.32 |
| Cognitive Ability Test (CAT) - Verbal analogies | 15.51 | 2.79 | 14.59 | 2.94 | 5.69 | <.001 | 32.10 |
| Cognitive Ability Test (CAT) - Figural analogies | 20.61 | 4.69 | 19.66 | 5.05 | 3.81 | <.001 | 19.49 |
| *Sociodemographic variables* | | | | | | | |
| Gender (0 = male) | 0.52 | 0.50 | 0.53 | 0.50 | −0.27 | 0.79 | −2.00 |
| Year of birth | 1992.66 | 0.51 | 1992.53 | 0.53 | 4.23 | <.001 | 25.00 |
| Language spoken at home (1 = German, 2 = mostly German) | 1.33 | 0.65 | 1.40 | 0.75 | −1.77 | 0.08 | −9.97 |
| Time spent in kindergarten/preschool (0 = less than one year, 1 = one year, 2 = 1–2 years, 3 = 2 years, 4 = more than 2 years) | 3.65 | 0.76 | 3.47 | 1.00 | 3.67 | <.001 | 20.27 |
| Age at enrolment in primary school (1 = 5 years old, 2 = 6 years old, 3 = 7 years old, 4 = at least 8 years old) | 2.06 | 0.50 | 2.15 | 0.53 | −2.47 | 0.01 | −17.47 |
| Total gross income per year (1 = less than 20,000, 2 = 20,000–29,999, 3 = 30,000–39,999 etc.) | 4.49 | 1.65 | 3.99 | 1.79 | 4.58 | <.001 | 29.05 |
| Highest ISEI in household | 61.67 | 15.61 | 56.29 | 16.92 | 5.79 | <.001 | 33.05 |
| Migration status (0 = no parent born abroad 1 = at least one parent born abroad, 2 = both parents born abroad) | 0.41 | 0.68 | 0.56 | 0.82 | −3.63 | <.001 | −19.91 |
| Recommended type of secondary school (1 = lower secondary school, 2 = upper secondary school) | 1.98 | 0.15 | 1.81 | 0.39 | 17.29 | <.001 | 57.54 |
| Total bias (Mean [\|%bias\|] | | | | | | | 28.77 |

*Note.* *p*-value for a two-tailed test. Performance estimates are clustered for class ID.

the *R*-package pan (version 1.6, Schafer & Zhao, 2018). With the quickpred function in mice, correlation analysis was performed to select suitable variables for predicting the missing values (Van Buuren & Groothuis-Oudshoorn, 2011). In the present study, the *fraction of missing values (FMI),* which describes the proportion of the total variance (variance within and between imputations) of a variable that can be attributed to the variance between imputations, was 0.97 (Madley-Dowd, Hughes, Tilling, & Heron, 2019). Due to the uncertainty associated with such a high *FMI* in estimating missing values, a large number of imputations was required to obtain reliable estimates. Therefore, in this study, $m = 100 \times 0.97 = 97 \approx 100$ imputations were performed (White, Royston, & Wood, 2011). At the same time, this fulfilled the requirements of Little, Lang, Wu, and Rhemtulla (2016), who generally recommended $m = 100$ imputations. Thus, 100 complete data sets with $N = 5963$ cases each were created. The parameters of all imputed data sets were evaluated according to Rubin's (1987) rules. Diagrams showing the mean values as well as the variance of all imputations and the fluctuation per iteration of each variable were used to assess convergence (Van Buuren & Groothuis-Oudshoorn, 2011). The diagrams showed that convergence was achieved with the mice algorithm for estimating the mean values and the variance of the imputed data.

The imputed data could then be used for the propensity score matching. In this design, causal inferences should be made. Following the counterfactual model, causal inferences require the outcomes of an individual when it receives or does not receive the treatment. The problem of making causal inferences is how to reconstruct the outcomes that are not observed, called counterfactuals. The propensity score matching allows researchers to reconstruct counterfactuals using observational data (Li, 2013). "The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum & Rubin, 1983, p. 41). I.e., students of the treated and untreated group with the same propensity score have identical distributions for the observed covariates. In other words, the propensity score is a function of the observed covariates such as the conditional distributions of the covariates of CLIL and non-CLIL students given a propensity score are the same (Rosenbaum & Rubin, 1983). The goal of the propensity score matching is to obtain two groups (with or without treatment) whose members are comparable concerning the observed covariates and therefore have a similar propensity score in order to imitate a randomized design (Beal & Kupzyk, 2014). Thus, the benefit of the propensity score matching in this design was particularly clear, given the lack of randomization due to the existing allocation of students to a CLIL or non-CLIL program. Using the *R*-package MatchThem (version 0.9.0, Pishgar & Greifer, 2019), the score was estimated individually for each person. It describes the probability of belonging to a certain treatment in view of the covariates, which can range from 0 to 1. For the following propensity score matching, i.e., the matching of CLIL and non-CLIL students with equal propensity scores, we used three routines: 1:1 without replacement, 1:1 with replacement, and 1:15 with replacement, each in combination with nearest neighbor matching, a caliper of $c = 0.025$, and logit distance matching for cognitive, socioeconomic, and prior achievement variables. All three routines removed pre-treatment differences between the groups successfully. As the 1:1 matching without replacement revealed the smallest standardized bias ($\%\mathrm{bias} = 0.42$ for 1:1 matching without replacement, $\%\mathrm{bias} = 0.53$ for 1:1 matching with replacement, and $\%\mathrm{bias} = 1.22$ for 1:15 matching with replacement), the subsequent analyses and results are based on this matching procedure. In the analyses, only individuals within the area of common support were compared, which means that the analyses applied to the subgroup of students consisting of CLIL and non-CLIL students with equal propensity scores. After eliminating relevant pre-treatment differences and achieving comparability between the two groups, no further adjustments were necessary when comparing the outcome differences.

To analyze the difference in English performance between CLIL and non-CLIL students before and after the propensity score matching, the test performance of both groups was compared in KESS 7 and KESS 8, estimating an ANCOVA as a structural equation model in Mplus (Muthén & Muthén, 1998–2017). The selection effect was estimated in grade four, using a regression analysis in Mplus (Muthén & Muthén, 1998–2017). In the ANCOVA, the English performance of KESS 7 was used as a predictor and the class ID as a cluster variable. This enabled statements on selection, preparation, and CLIL effects. Between KESS 7 and KESS 8, a subsequent Difference-in-Differences analysis was conducted to obtain the additive CLIL-effect. A Difference-in-Differences analysis is a way to estimate causal relationships by estimating the difference in outcomes before and after the implementation of the treatment (CLIL) for the treatment and control groups (CLIL and non-CLIL students) (Bertrand, Duflo, & Mullainathan, 2004).

## 4. Results

### *4.1. Selection effect: differences before propensity score matching*

Predictors were tested for significant differences between CLIL and non-CLIL students before and after propensity score matching. Table 2 shows an average standardized bias in percent, i.e., the mean difference in percent of the average standard deviation (Rosenbaum & Rubin, 1985), before matching of 28.77% as well as significant differences between the groups in almost all predictors that confirm the assumption of selection effects.

We found a significant difference in English proficiency between future CLIL and non-CLIL students at the end of grade four (KESS 4), which strongly indicates selection effects. In the listening comprehension test, the primary school students who became CLIL students in the future scored significantly better than their non-CLIL classmates (see Table 2). Based on this, the assumption of positive selection in CLIL programs, which Rumlich (2018) and Dallinger et al. (2016) have already reported for Germany, is supported.

To prevent initial selectivity and to analyze CLIL effects, it was necessary to match the students of both groups in terms of their English proficiency and other relevant background variables at the end of grade four. The study examined 20 variables related to prior achievement, cognitive abilities, or sociodemographic background, as they were believed to be associated with the positive selection of CLIL students (Möller et al., 2017; Rumlich, 2018). The set of variables was comprised of performance variables of the fourth grade (English, German, and mathematics), the school grades in German, math's, and life science, as well as sociodemographic variables such as gender, year of birth, language spoken among parents and child, time spent in kindergarten and preschool, age at enrolment, socioeconomic status of the family, migration background, recommended school type, as well as cognitive abilities of the child. In the end, a propensity score matching was

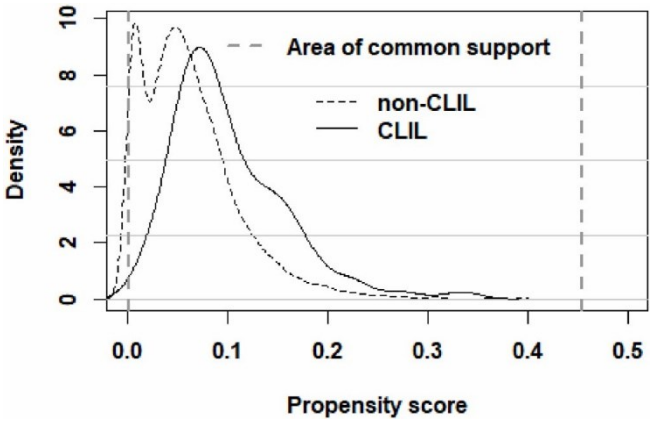


**Fig. 3.** Area of common support before matching (exemplary for one imputation).

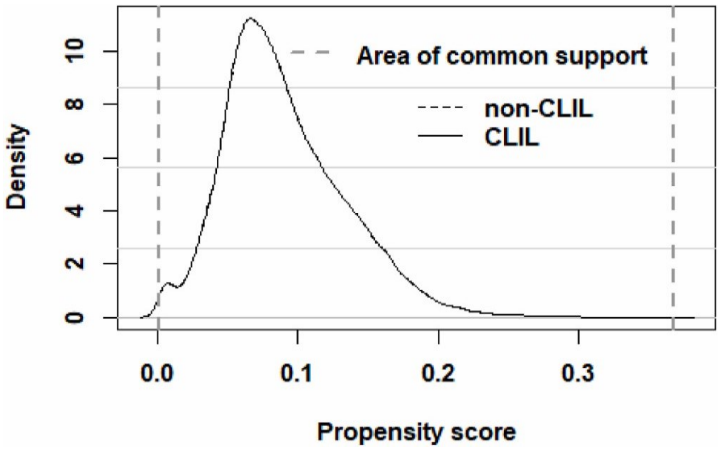


**Fig. 4.** Area of common support after matching.

performed in all 100 complete data sets.

A regression analysis revealed a small effect of $\beta = 0.340$ ($p < .001$) in favor of the later CLIL students in listening comprehension. A selection effect, which shows that the higher-performing students with socioeconomically more advantageous backgrounds later attended a CLIL program, was already shown by the descriptive statistics in Table 2 and was estimated and confirmed by this.

### *4.2. Preparation effect: differences after propensity score matching*

Fig. 3 illustrates the area of common support, which is the area where treatment and comparison groups overlap (Retelsdorf, Becker, Köller, & Möller, 2012). By using a propensity score matching, conclusions on treatment effects could only be drawn for a treated individual for whom there was a comparative individual with a similar propensity score, that is, for students within the area of common support (Garrido et al., 2014). Therefore, the lowest and highest propensity scores of both groups were compared. The area of common support is the overlapping region of the two intervals. In this study, almost every student was located in this area and, thus, almost every CLIL student could be matched with a non-CLIL student (see Fig. 3). Fig. 4 demonstrates the comparability of CLIL and non-CLIL students as the distributions of the propensity scores after matching were the same, that is, there were no relevant pre-treatment differences concerning cognitive, socioeconomic, and prior achievement variables between CLIL and non-CLIL students.

The descriptive statistics after the propensity score matching can be found in Table 3. In comparison with the average standardized bias before matching of 28.77% (see Table 2), the bias after matching was 0.42% (see Table 3). According to Caliendo and Kopeinig (2008), a standardized bias of less than 5% is an indicator that matching has led to a balanced distribution among the covariates. Moreover, significant differences between the background variables could no longer be found. Selection effects were thus controlled by matching. Since the students of both groups showed comparable English skills at the end of the fourth grade and the KESS 7 survey was carried out at the beginning of grade seven, the analysis allowed

**Table 3**
Imbalance in predictor variables between CLIL- and non-CLIL students after matching.

| Predictor variable | CLIL | | Non-CLIL | | | | |
|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *t* | *p* | % bias |
| *KESS 4 performance* | | | | | | | |
| English listening WLE | 121.50 | 28.82 | 121.60 | 28.20 | −0.04 | 0.97 | −0.37 |
| German reading WLE | 0.56 | 1.11 | 0.55 | 1.16 | 0.10 | 0.92 | 0.88 |
| German orthography WLE | 123.50 | 23.21 | 123.30 | 22.55 | 0.11 | 0.92 | 0.89 |
| Mathematics WLE | 0.29 | 1.09 | 0.29 | 1.11 | −0.02 | 0.99 | −0.18 |
| Grades in German Language | 1.89 | 0.50 | 1.89 | 0.51 | −0.03 | 0.98 | −0.20 |
| Grades in German Reading | 1.70 | 0.59 | 1.69 | 0.54 | 0.01 | 0.99 | 0.18 |
| Grades in Mathematics | 2.00 | 0.53 | 2.00 | 0.55 | −0.02 | 0.98 | −0.19 |
| Grades in Science | 1.88 | 0.51 | 1.88 | 0.48 | 0.00 | 1.00 | 0.00 |
| Cognitive Ability Test (CAT) - Verbal analogies | 15.48 | 2.79 | 15.48 | 2.65 | −0.03 | 0.98 | −0.22 |
| Cognitive Ability Test (CAT) - Figural analogies | 20.55 | 4.71 | 20.51 | 4.56 | 0.10 | 0.92 | 0.82 |
| *Sociodemographic variables* | | | | | | | |
| Gender (0 = male) | 0.52 | 0.50 | 0.52 | 0.50 | −0.03 | 0.98 | −0.20 |
| Year of birth | 1992.65 | 0.50 | 1992.65 | 0.51 | 0.05 | 0.96 | 0.39 |
| Language spoken at home (1 = German, 2 = mostly German) | 1.32 | 0.63 | 1.32 | 0.68 | −0.14 | 0.89 | −1.07 |
| Time spent in kindergarten/preschool (0 = less than one year, 1 = one year, 2 = 1–2 years, 3 = 2 years, 4 = more than 2 years) | 3.64 | 0.76 | 3.64 | 0.78 | 0.07 | 0.95 | 0.65 |
| Age at enrolment in primary school (1 = 5 years old, 2 = 6 years old, 3 = 7 years old, 4 = at least 8 years old) | 2.07 | 0.50 | 2.06 | 0.48 | 0.01 | 1.00 | 0.21 |
| Total gross income per year (1 = less than 20,000, 2 = 20,000–29,999, 3 = 30,000–39,999 etc.) | 4.47 | 1.65 | 4.47 | 1.65 | 0.02 | 0.99 | 0.12 |
| Highest ISEI in household | 61.39 | 15.55 | 61.43 | 16.02 | −0.03 | 0.98 | −0.24 |
| Migration status (0 = no parent born abroad 1 = at least one parent born abroad, 2 = both parents born abroad) | 0.40 | 0.67 | 0.41 | 0.71 | −0.07 | 0.94 | −0.58 |
| Recommended type of secondary school (1 = lower secondary school, 2 = upper secondary school) | 1.98 | 0.15 | 1.98 | 0.15 | −0.07 | 0.94 | −0.67 |
| Total bias (Mean [\|%bias\|] | | | | | | | 0.42 |

*Note.* *p*-value for a two-tailed test. Performance estimates are clustered for class ID.

**Table 4**
Descriptive statistics of English proficiency of CLIL and non-CLIL students for grades four to eight before and after matching.

| Grade | Test | PSM | CLIL | | non-CLIL | |
|---|---|---|---|---|---|---|
| | | | *M* | *SD* | *M* | *SD* |
| 4 | Listening | No | 122.3 | 29.42 | 112.62 | 28.35 |
| | Listening | Yes | 121.50 | 28.82 | 121.60 | 28.20 |
| 7 | C-test | No | 0.62 | 0.97 | −0.03 | 0.93 |
| | C-test | Yes | 0.60 | 0.97 | 0.25 | 0.91 |
| 8 | C-test | No | 1.58 | 1.05 | 0.89 | 0.88 |
| | C-test | Yes | 1.56 | 1.03 | 1.14 | 0.87 |

differentiated statements to be made about the effect of preparatory instruction in grades five and six.

Table 4 reveals the descriptive statistics of English proficiency. Table 5 presents the results of the structural equation model comparing C-test scores for CLIL and non-CLIL students. In grade seven, without propensity score matching (PSM) in grade four, a significant effect of $\beta = 0.686$ between the two groups was observed in the C-test scores. After propensity score matching, this effect decreased to a significant $\beta = 0.369$ (see Table 5). It should be noted that CLIL started at the beginning of grade seven and therefore cannot have had any influence on English

**Table 5**
Estimates of preparation and CLIL effects on English proficiency (C-test) of CLIL and non-CLIL students for grades seven and eight before and after matching (ANCOVA as SEM with z-standardized, latent variables, robust standard errors).

| Effects | Before PSM | | | | After PSM | | | |
|---|---|---|---|---|---|---|---|---|
| | β | *SE* | *t* | *p* | β | *SE* | *t* | *p* |
| Performance 7 on performance 8 (stability) | 0.639 | 0.020 | 32.06 | <0.001 | 0.549 | 0.079 | 6.95 | <0.001 |
| CLIL on performance 7 (preparation effect) | 0.686 | 0.139 | 4.95 | <0.001 | 0.369 | 0.153 | 2.41 | 0.02 |
| CLIL on performance 8 (overall effect) | 0.767 | 0.124 | 6.21 | <0.001 | 0.425 | 0.130 | 3.27 | <0.01 |
| Specific effect CLIL on performance 8 when controlling for performance 7 (CLIL-effect) | 0.329 | 0.120 | 2.74 | <0.01 | 0.222 | 0.131 | 1.70 | 0.09 |
| Indirect effect CLIL on performance 8 (preparation via performance 7 on performance 8) | 0.438 | 0.092 | 4.78 | <0.001 | 0.203 | 0.095 | 2.13 | 0.03 |
| Additive effect of CLIL | 0.082 | 0.143 | 0.57 | 0.57 | 0.056 | 0.156 | 0.36 | 0.72 |

performance measured at the beginning of grade seven. Differences in relevant background variables and resulting selection effects were eliminated by the propensity score matching in KESS 4. It follows that the significant effect of $\beta = 0.369$ determined in this study could be explained by the preparatory classes, which included one or two additional English lessons in grades five and six.

### *4.3. CLIL effect: effect sizes before and after propensity score matching*

When looking at the results of the C-test at the end of grade eight, it became clear that ignoring both selection and preparation effects could lead to an overestimation of the CLIL effect (see Table 5). Table 4 shows that differences in prior English achievement were eliminated by propensity score matching and, thus, differences in means between CLIL and non CLIL students in grades seven and eight were reduced. Unless propensity score matching was performed and English performance was not controlled in grade seven, a significant difference of $\beta = 0.767$ was observed between the two groups. When selection effects were taken into account, the difference remained significant, but was reduced to $\beta = 0.425$. When additionally preparation effects were taken into account by controlling for English performance in grade seven, the effect of CLIL was reduced to a non-significant $\beta = 0.222$. Estimating the effect of CLIL on English performance in grade eight mediated by English performance in grade seven revealed a significant indirect effect of $\beta = 0.203$. The subsequent Difference-in-Differences analysis revealed a non-significant effect of $\beta = 0.056$, controlling selection effects *and* taking preparation effects into account by estimating β for the differences in English performance from the beginning of grade seven to the end of grade eight between CLIL and non-CLIL students. This effect represented the additive effect of CLIL over nearly two years, thereby contradicting our main assumption of a positive CLIL effect.

## 5. Discussion

Empirical research often revealed positive effects of CLIL programs. It was noticeable that the effects differed according to the country: Studies from southern European countries were more likely to show positive effects of CLIL instruction than studies from northern European countries. Common to all studies was the insufficient consideration and estimation of both selection and preparation effects. Thus, the aim of this paper was to investigate selection and preparation effects as well as the effect of CLIL on second language learning in Germany. A large sample from a longitudinal full survey was examined, estimating selection and preparation effects and considering both to compare the English skills of CLIL and non-CLIL grammar school students from grade four to grade eight.

### *5.1. Selection effect*

Students choose a CLIL program themselves or are selected by schools. This leads to a CLIL group, which is likely to possess, for example, higher prior knowledge in English, higher cognitive abilities, and a more favorable family background (Dallinger et al., 2016). These prerequisites are assumed to affect English skills.

At the end of grade four, we estimated differences between future CLIL and non-CLIL groups in achievement, cognitive abilities, and sociodemographic variables. Future CLIL students showed significantly better results in prior achievement in English and all other subjects, higher cognitive abilities, and a more favorable sociodemographic background. Before the start of CLIL (and before the start of the preparatory courses), there was an effect of $\beta = 0.340$ in favor of future CLIL students in English listening comprehension, indicating selection effects and demonstrating the advantage of future CLIL students in English at the end of grade four.

As previously reported by Rumlich (2018), selection effects also occurred in this study, even on a large comprehensive sample. Rumlich (2018) estimated the CLIL effect considering selection and preparation effects by controlling age, gender, first language, and prior achievement. The students were tested at the end of grade six and at the end of grade eight. He was also able to show that CLIL students showed significantly higher English C-test scores even before CLIL lessons started, but he could not estimate the selection or preparation effect separately. In the German context, also Dallinger et al. (2016) considered selection effects when estimating the CLIL effect and controlled prior achievement, general abilities, motivation, demographics, classroom composition, and instructional quality at the beginning of grade eight. She reported substantially better scores of CLIL students in prior achievement, motivation, cognitive abilities, and socioeconomic status compared to non-CLIL students as well as higher instructional quality in CLIL classes and, thus, confirmed the importance of considering selection effects when estimating the CLIL effect. However, she could not make any statements on the size of selection or preparation effects as the difference in English achievement at the beginning of grade eight was already affected by both selection and preparation as well as CLIL lessons in grade seven. May (2009) also analyzed the KESS data and revealed that later CLIL students already exceeded later non-CLIL students in English performance measured by a C-test at the end of grade four. We found a medium effect of $\beta = 0.340$. May (2009) revealed a comparable medium effect of $\delta = 0.5$. In addition to showing the effect, we were also able to determine its size. The selection effect demonstrated that CLIL attracted higher-performing students, with higher cognitive skills, and a more favorable sociodemographic background. Thus, a propensity score matching was required. Otherwise, possible later effects between CLIL and non-CLIL students could not be clearly attributed to CLIL or preparatory lessons but may have existed prior to its start and be favored by differences in performance-related background variables.

### *5.2. Preparation effect*

In some countries, such as Germany, students are prepared for the upcoming CLIL classes through additional English lessons. It is assumed that an increased number of English lessons would improve the English performance. However, this preparation effect had rarely been taken into account in previous studies when estimating the effect of CLIL and had not been estimated yet.

Not taking selection and preparation effects into account resulted in

an effect of $\beta = 0.686$ in English C-test scores between the two groups at the beginning of grade seven. This large effect agreed with the results from May (2009) and Rumlich (2018). May (2009) also analyzed the KESS data and reported a large effect of $\delta = 0.98$ in the C-test at the beginning of grade seven. Rumlich (2018) revealed a large effect of $\delta = 0.84$ between CLIL and non-CLIL students from schools without CLIL programs and $\delta = 1.18$ between CLIL and non-CLIL students from parallel classes of schools with CLIL programs in the C-test at the end of grade six. However, both did not use a propensity score matching to estimate the preparation effect separately. In contrast, the use of a propensity score matching in the present study reduced the effect to $\beta = 0.369$, indicating the preparation effect resulting from two years of additional English lessons.

The results showed that in addition to the selection effect, the even greater preparation effect also has to be considered. This was confirmed by Dallinger et al. (2016), who revealed that prior achievement was the most important confounder when estimating the CLIL effect on English performance. Preparatory teaching is not used in all countries, but we were able to show its potential for improving English performance. Consequently, preparatory instruction can help to follow CLIL lessons more easily. If additional English lessons are conducted, it is necessary for research to consider their effect, as the present study could demonstrate that the CLIL effect could otherwise be overestimated.

### *5.3. CLIL effect*

Ignoring the adjustment for selection and preparation effects led to a significant difference between CLIL and non-CLIL students in English C- test scores of $\beta = 0.767$. Considering selection effects, the difference was reduced to $\beta = 0.425$, but was still significant. Additionally considering preparation effects further reduced the CLIL effect to a non-significant $\beta = 0.222$. Estimating the effect of CLIL on English performance in grade eight mediated by English performance in grade seven revealed a significant indirect effect of $\beta = 0.203$. The estimation of the indirect effect avoided an overestimation of the direct CLIL effect and showed the importance of prior achievement and, thus, of preparatory lessons. Conducting a Difference-in-Differences (DiD) analysis to estimate the additive CLIL effect from the beginning of grade seven to the end of grade eight revealed a slightly positive non-significant effect of $\beta = 0.056$. These results showed the importance of taking selection and preparation effects into account since ignoring them resulted in an overestimation of the effect of CLIL. It corresponds to the results of previous studies in particular by Rumlich (2018). He reported a non-significant effect of $\beta = 0.01$ for CLIL and non-CLIL students from schools without a CLIL program and $\beta = -0.05$ for CLIL and non-CLIL students from schools offering a CLIL program for the performance increase from grade six to grade eight in C-test. Dallinger et al. (2016) also revealed a non-significant effect in C-test between CLIL and non-CLIL classes considering prior achievement. These consistent findings lead to the conclusion that CLIL maintains the advantage in English skills built up through selection and preparation, but does not have a major effect beyond that for the development of global language skills in English. However, it must be noted that it is not possible to estimate the pure CLIL effect, but rather the effect of the CLIL-component of a complex treatment of preparatory lessons and CLIL. We estimated the additive CLIL effect under the assumption of perfect stability between grades seven and eight (DiD estimate). But, fading out-effects are to be taken into account after finishing the preparatory lessons as the indirect effect of preparation indicates. CLIL compensates for the fading out-effect but does not contribute any significant added value.

### *5.4. Strengths and limitations*

Our data originated from a full survey and went beyond previous analyses when controlling and estimating selection and preparation effects. Our findings should be replicated in future studies that also rigorously control selection and preparation effects as it has been shown that they have a greater effect on English performance than CLIL itself. To analyze the CLIL effect further, future studies should also include students' motivation, teacher characteristics, and instructional quality. Dallinger et al. (2016) included the previously mentioned variables. They revealed that CLIL students were significantly more motivated than non-CLIL students. Differences between CLIL and non-CLIL students in C-test scores and listening comprehension test scores were, besides others, also significantly predicted by motivation. Additionally, they revealed that CLIL-teachers showed more enthusiasm for teaching than non-CLIL-teachers due to the greater need of CLIL lesson preparation. They also found significant differences in instructional quality in history in favor of CLIL classes, but not in English. As no such data from the KESS study were available, these aspects could not be considered in the present study. In addition, future studies with exact data on the number of English lessons could estimate the effect of additional English lessons even more precisely. It must be kept in mind that the effects in grades seven and eight could also have been caused by other influences besides preparatory lessons or CLIL. However, such effects produced by uncontrolled variables seem relatively unlikely. We already used a propensity score matching to imitate a randomized design and control relevant covariates. But, even if it is not feasible, an experimental design would be helpful to create truly randomized CLIL and non-CLIL groups. This would lead to even more meaningful results on the CLIL effect.

This study focused on the CLIL programs in Hamburg and could just make statements on the effects of CLIL for German students. CLIL programs in other countries might differ in terms of selection, preparation, and implementation. For example, some countries did not offer preparatory lessons (e.g. the Netherlands), did not use such a selective approach as Germany (e.g. Spain), and/or offered different CLIL programs (e.g. in terms of the amount of CLIL lessons) (Goris, Denessen, & Verhoeven, 2013; Goris et al., 2019). This is, selection and preparation biases are not equally likely sources of influence on the differences in English proficiency of CLIL and non-CLIL students across countries. Thus, similar research is required in other countries to investigate the effects of selection, preparation, and CLIL in other school systems. In addition, a listening comprehension test was used in grade four to assess English proficiency, but in grades seven and eight, a C-test was used. Prior achievement in English was taken into account by matching students in terms of their performance in listening comprehension. Even if the C-test covered several areas of English proficiency and correlations between C-test scores and listening comprehension test scores were assumed and confirmed by Dallinger et al. (2016), it could be that part of the prior achievement was not taken into account in the propensity score matching due to the different English tests. Thus, part of the preparation effect could be a selection effect.

This study thus contributes greatly to the gain in knowledge regarding the importance of selection and in particular preparation effects in relation to the effect of CLIL. However, the question arose as to why one should introduce CLIL at all given its small and non-significant effect. Instead, one could argue that simply raising the number of English lessons could have a greater effect than CLIL. However, we measured English performance only by C-test and found small differences between CLIL and non-CLIL students. The C-test is an objective, reliable, and valid measuring instrument. It achieves values between $\alpha = 0.92$ and $\alpha = 0.98$ for the measurement consistency of the individual texts. In terms of the validity, the C-test scores correlate highly with school grades, teacher evaluations of student performance, or the results of other language tests (e.g. the test of English as a foreign language (TOEFL)). Furthermore, studies on construct validity reveal the C-test to be a valid instrument for measuring general language competence. Recent studies showed that participants considered the wider context when reconstructing the C-test gaps if necessary. Therefore, the C-test measures not only at the micro level but also at the macro level (Grotjahn, 2002). But, as the C-test measures global comprehension, knowledge of words, spelling, syntax, and grammar, it could be assumed that

the C-test is more appropriate for additional English lessons. Skills such as speaking and listening comprehension were not recorded. The impact of CLIL on these domains and also on knowledge in the content subjects might have been even stronger. Such effects will have to be investigated in future studies. If there were no great advantages in favor of CLIL students, as Dallinger et al. (2016) already showed for history, the benefit of the implementation of CLIL would have to be questioned from an empirical point of view. Consequently, there is a need for further studies regarding these skills and subjects as well as the effects of teacher characteristics and instructional quality to build an empirical basis for future decisions regarding the expansion of CLIL programs. The number of grammar schools offering an English CLIL program in Hamburg increased steadily over the last 15 years. In 2002/2003, i.e., at the beginning of the KESS study, seven grammar schools in Hamburg provided an English CLIL program. This number rose to 15 in 2009/2010 and 22 in 2020/2021. This is, 22 of 73 grammar schools in Hamburg, i.e., about 30%, provide English CLIL lessons (Behörde für Schule und Berufsbildung, 2009, 2020). Further studies are needed to justify this development. In the meanwhile, over optimistic summaries of the CLIL effects on English performance should be treated with caution.

## Declarations of competing interest

None.

## Acknowledgements

This research was supported by a grant from the German Research Foundation (DFG) to Jens Möller [grant number MO 648/26-1].

This paper uses data from the longitudinal study KESS. This data set was generated by the Free and Hanseatic City of Hamburg through the Ministry of Schools and Vocational Training between 2002 and 2012 and has been provided to the MILES scientific consortium (Methodological Issues in Longitudinal Educational Studies) for a limited period with the aim of conducting in-depth examinations of scientific questions. MILES is coordinated by the Leibniz Institute for Science and Mathematics Education (IPN).

## References

Beal, S. J., & Kupzyk, K. A. (2014). An introduction to propensity scores: What, when, and how. *The Journal of Early Adolescence, 34*(1), 66–92. https://doi.org/10.1177/0272431613503215

Behörde für Schule und Berufsbildung [authority for school and vocational training]. (2009). *Fremdsprachenunterricht in Hamburg: Schuljahr 2009/10 [Foreign Language instruction in Hamburg]*. Retrieved from https://silo.tips/download/fremdsprachenunterricht-schuljahr-2009-10-hamburg. (Accessed 29 January 2021).

Behörde für Schule und Berufsbildung [authority for school and vocational training]. (2014). *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-VHA000000199&documentnumber=1&numberofresults=1&doctyp=vvhhschulr&showdoccase=1&doc.part=F&paramfromHL=true. (Accessed 29 January 2021).

Behörde für Schule und Berufsbildung [authority for school and vocational training]. (2020). *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).

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics, 119*(1), 249–275. https://doi.org/10.1162/003355304772839588

Bonsen, M., Bos, W., Gröhlich, C., & Rau, A. (2009). Ziele der Untersuchung KESS 7 [educational objectives of the study of competencies and Attitudes of students at the beginning of year 7 (KESS 7)]. In W. Bos, M. Bonsen, & C. Gröhlich (Eds.), *HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 5. KESS 7 - Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7* (pp. 13–21). Münster: Waxmann.

Bortz, J., & Döring, N. (2002). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler [Methodologies and Evaluations in Research. Suitable for Researchers in the Humanities and Social Sciences* (3rd ed.). Berlin, Heidelberg: Springer.

Bos, W., Bonsen, M., & Gröhlich, C. (Eds.). (2009a). *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 in the Beginning of Grade 7]. HANSE - Hamburger Schriften zur Qualität im Bildungswesen* (Vol. 5). Münster: Waxmann.

Bos, W., Bonsen, M., Gröhlich, C., & Rau, A. (2009b). Kompetenzen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 6 [Competencies of Students at the End of Grade 6]. In W. Bos, M. Bonsen, & C. Gröhlich (Eds.), *HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 5. KESS 7 - Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7* (pp. 39–54). Münster: Waxmann.

Bos, W., Brose, U., Bundt, S., Gröhlich, C., Hugk, N., Janke, N., … Voss, A. (2006). Anlage und Durchführung der Studie ‚Kompetenzen und Einstellungen von Schülerinnen und Schülern – Jahrgangsstufe 4 (KESS 4)' [Development and Implementation of the Study ‚Competencies and Attitudes of Students - year 4 (KESS 4)']. 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. 9–31). Münster: Waxmann.

Bos, W., & Gröhlich, C. (Eds.). (2010). *KESS 8: Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8 [Competencies and Attitudes of Students at the End of Grade 8]. HANSE - Hamburger Schriften zur Qualität im Bildungswesen* (Vol. 6). Münster: Waxmann.

Bos, W., Gröhlich, C., Guill, K., Scharenberg, K., & Wendt, H. (2010). Ziele und Anlage der Studie KESS 8 [Educational Objectives and Approaches of the Study of Competencies and Attitudes of Students at the End of Year 8 (KESS 8)]. In W. Bos, & C. Gröhlich (Eds.), *HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 6 KESS 8: Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe* (Vol. 8, pp. 9–20). 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* (Vol. 1). 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 (Reprinted 2010)*. 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* (pp. 139–157). 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

Eckes, T., & Grotjahn, R. (2016). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa

European Commission. (2005). *Promoting language learning and linguistic diversity: An action plan 2004-06*. Luxembourg: Office for Official Publications of the European Communities.

Eurydice. (2006). *Content and Language integrated learning (CLIL) at school in Europe. EURYDICE Survey*. Brüssel: Eurydice.

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

Gass, S. M., & Mackey, A. (2015). Input, interaction, and output in second language acquisition. In B. VanPatten, & J. Williams (Eds.), *Theories in second language acquisition: An Introduction* (pp. 180–206). New York: Routledge.

Genesee, F. (2014). Is early Second Language learning really better? Evidence from research on students in CLIL programs. *Babylonia*, (14), 26–30, 01.

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

Grotjahn, R. (2002). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis [Construction and Usage of C-Tests: A Practical Guide]. In R. Grotjahn (Ed.), *Der C- Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 4, pp. 211–225). Bochum: AKS-Verlag.

Harsch, C., & Schröder, K. (2007). Textkonstruktion: C-test [text construction: The C-test]. In B. Beck, & E. Klieme (Eds.), *Beltz-Pädagogik. Sprachliche Kompetenzen Konzepte und Messung: DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 212–225). Weinheim, Basel: Beltz Verlag.

Hastings, A. J. (2002). Error analysis of an English C-Test: Evidence for integrated processing. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 4, pp. 53–66). Bochum: AKS-Verlag.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4–12+R) [A Cognitive Abilities Test for the 4th to the 12th Grade, Revised (KFT 4-12+R)]*. Gottingen: Beltz Test.

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.
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 [Standing Conference of the Ministers of Education and Cultural Affairs]. (2006). Bericht „Konzepte für den bilingualen Unterricht – Erfahrungsbericht und Vorschläge zur Weiterentwicklung“ (Bericht des Schulausschusses vom 10.04.2006) [Report ”Concepts for Bilingual Teaching – experience Reports and Suggestions for Further Development" (Report of the School Committee from 10.04.2006)]. Retrieved from https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2006/2006_04_10-Konzepte-bilingualer-Unterricht.pdf (Accessed 29 January 2021).
Li, M. (2013). Using the propensity score method to estimate causal effects. *Organizational Research Methods, 16*(2), 188–226. https://doi.org/10.1177/1094428112447816
Little, T. D., Lang, K. M., Wu, W., & Rhemtulla, M. (2016). Missing data. In D. Cicchetti (Ed.), *Developmental psychopathology. Theory and method* (Vol. 3, pp. 760–797). Hoboken, New Jersey: Wiley.
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
May, P. (2006a). 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.
May, P. (2006b). Orthographische Kompetenz und ihre Bedingungen am Ende der vierten Jahrgangsstufe [Orthographical Competence and its Prerequisites at the End of Year Four]. 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. 111–141). Münster: Waxmann.
May, P. (2009). Kompetenz in Englisch: Vertiefende analysen [English language proficiency: Extensive analyses]. In W. Bos, M. Bonsen, & C. Gröhlich (Eds.), *HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 5. KESS 7 - Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7* (pp. 54–68). 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. *International Journal of Bilingual Education and Bilingualism*. https://doi.org/10.1080/13670050.2017.1405906. Advance online publication.
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
Moosbrugger, H. (2012). Item-response-theorie (IRT) [Item-Response-Theory (IRT)]. In H. Moosbrugger, & A. Kelava (Eds.), *Springer-Lehrbuch. Testtheorie und Fragebogenkonstruktion* (2nd ed., pp. 228–274). Berlin, Heidelberg: Springer.
Muthén, L. K., & Muthén, B. (1998-2017) *Mplus user's guide*, 0 (8th ed.). Los Angeles, CA: Muthén & Muthén.
Nikolova, R., & Ivanov, S. (2010). Englischleistungen [achievements in English language acquisition]. In W. Bos, & C. Gröhlich (Eds.), *HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 6. KESS 8: Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8* (pp. 49–65). Münster: Waxmann.
Nold, G., Hartig, J., Hinz, S., & Rossa, H. (2008). Klassen mit bilingualem Sachfachunterricht: Englisch als Arbeitssprache [CLIL Classes: English as a Working Language]. In E. Klieme (Ed.), *Beltz Pädagogik. Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (pp. 451–457). Weinheim: Beltz.
Pietsch, M., & Krauthausen, G. (2006). Mathematisches Grundverständnis von Kindern am Ende der vierten Jahrgangsstufe [Basic Mathematical Knowledge of Children at the End of Grade Four]. 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. 143–163). Münster: Waxmann.
Pishgar, F., & Greifer, N. (2019). *Package MatchThem: Matching and weighting multiply imputed datasets. [Computer software manual] (version 0.9.0)*.
Raatz, U., & Klein-Braley, C. (1983). Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis [A New Approach to Measuring Language Performance. The C-Test: Theory and Practice]. In R. Horn, K. Igenkamp, & R. S. Jäger (Eds.), *Tests und Trends 3.1983. Tests und Trends 3: Jahrbuch der Pädagogischen Diagnostik* (pp. 107–138). Weinheim: Beltz.
Rauch, D., & Hartig, J. (2012). Interpretation von Testwerten in der IRT [Interpretation of Test Results in the IRT]. In H. Moosbrugger, & A. Kelava (Eds.), *Springer-Lehrbuch. Testtheorie und Fragebogenkonstruktion* (2nd ed., pp. 253–264). Berlin, Heidelberg: Springer.
Rentzsch, K., & Schütz, A. (2009). *Psychologische Diagnostik: Grundlagen und Anwendungsperspektiven [Psychological Diagnostics. Basics and application perspectives] (1. Aufl.). Kohlhammer-Urban-Taschenbücher* (Vol. 565). Stuttgart: Kohlhammer.
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. *British Journal of Educational Psychology, 82*(4), 647–671. https://doi.org/10.1111/j.2044-8279.2011.02051.x
Robitzsch, A. (2019). *Package sirt: Supplementary item response theory models. [Computer software manual] (Version 3.4-64)*.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.
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.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
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)*.
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* (pp. 55–72). Amsterdam: Brill/Rodopi.
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-Oudshorn, K. (2019). *Package mice: Multivariate imputation by chained equations. [Computer software manual] (version 3.6.0)*.
Van Kampen, E., Mearns, T., Meirink, J., Admiraal, W., & Berry, A. (2018). How do we measure up? A review of Dutch CLIL subject pedagogies against an international backdrop. *Dutch Journal of Applied Linguistics, 7*(2), 129–155. https://doi.org/10.1075/dujal.18004.kam
Verspoor, M., Bot, K. de, & 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
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450.
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