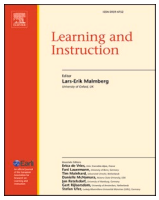
Learning and Instruction 92 (2024) 101923

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



CLIL effects on academic self-concepts: Positive effects in English but detrimental effects in math?

Marlene Wunberg [a,*], Jürgen Baumert [b], Maja Feddermann [c], Julian F. Lohmann [a], Jens Möller [a]

[a] Institute for Psychology of Learning and Instruction, Kiel University, Olshausenstraße 75, 24118, Kiel, Germany
[b] Max-Planck-Institute for Human Development, Lentzeallee 94, 14196, Berlin, Germany
[c] Institute for Educational Monitoring and Quality Development, Beltgens Garten 25, 20537, Hamburg, Germany

ARTICLE INFO

Keywords:
Content and language integrated learning (CLIL)
Academic self-concepts
Propensity score matching
Structural equation modeling

STRUCTURED ABSTRACT

Background: Content and Language Integrated Learning (CLIL) is considered a promising approach to enhancing foreign language skills and motivation. However, its impact on students' academic self-concepts remains largely unclear.
Aims: This study aimed to investigate whether CLIL positively affects students' English self-concepts but harms their math self-concepts in Grade 8 after two years of CLIL participation. Furthermore, the study intended to control for and disentangle selection and preparation effects caused by selective access and increased English instruction before the start of CLIL, as neglecting a priori differences between CLIL and non-CLIL students has led to overestimating CLIL effects in the past.
Sample: Participants were 5963 academic-track school students.
Methods: Propensity score matching was applied to control for selection effects. Structural equation modeling was used to estimate CLIL effects on English and math self-concepts. The inclusion of control variables allowed for accounting for preparation effects.
Results: CLIL students had significantly higher English self-concepts than non-CLIL students, which could be explained by selection and preparation effects. However, attending CLIL helped to maintain the advantage over non-CLIL students over the first two years of CLIL participation. CLIL had no detrimental effects on students' math self-concepts but left them unaffected.
Conclusions: The study contributes to a deeper understanding of the effects of CLIL on students' self-concepts in different subjects. Furthermore, the results highlight the importance of accounting for both selection and preparation effects in future CLIL studies to obtain unbiased CLIL effect estimates.

1. Introduction

CLIL is an approach to bilingual instruction widely implemented in Europe in which curricular content in non-language subjects is taught and learned through a foreign language (L2) (Dalton-Puffer, 2011).

Although promoting motivation in the L2 is a stated goal of CLIL implementation (Eurydice, 2006), research on the effectiveness of CLIL has rarely considered motivational characteristics such as students' academic self-concepts. The few studies explicitly examining academic self-concepts mostly showed that CLIL students had higher L2 self-concepts than non-CLIL students (e.g., Rumlich, 2017). This finding is particularly attributed to CLIL students' higher L2 achievement, which correlates strongly with self-concepts. However, recent studies found that CLIL students' higher L2 achievement was often not due to CLIL but to pre-existing differences between CLIL and non-CLIL students resulting from selection effects and enhanced L2 instruction prior to CLIL, referred to as preparation effects (e.g., Feddermann, Möller, & Baumert, 2021). For L2 self-concepts, comparable findings are conceivable but still largely unexplored.

Besides L2 self-concepts, examining CLIL effects on math self-concepts provides an interesting extension to both CLIL research and dimensional comparison theory (DCT; Möller & Marsh, 2013), the latter describing negative effects of verbal achievement on math self-concepts. Given the higher L2 achievement levels of CLIL students compared to

* Corresponding author.
E-mail addresses: mwunberg@ipl.uni-kiel.de (M. Wunberg), sekbaumert@mpib-berlin.mpg.de (J. Baumert), Maja.Feddermann@ifbq.hamburg.de (M. Feddermann), jlohmann@ipl.uni-kiel.de (J.F. Lohmann), jmoeller@ipl.uni-kiel.de (J. Möller).

https://doi.org/10.1016/j.learninstruc.2024.101923
Received 19 May 2023; Received in revised form 30 November 2023; Accepted 5 April 2024
Available online 18 April 2024

non-CLIL students, it would be interesting to explore whether attending CLIL might have detrimental side effects on math self-concept.

Against this background, the present study aimed to investigate the effects of English-language CLIL on students' English and math self-concepts based on a large German panel study. Using structural equation modeling, we explored more specifically whether participation in CLIL positively affected students' English self-concepts but came at the expense of their math self-concepts. In addition, we accounted for and disentangled potential selection and preparation effects by applying propensity score matching (PSM) and controlling for preparatory English instruction.

In this way, the study, on the one hand, complements and extends the sparse research on the effectiveness of CLIL regarding motivational student characteristics. On the other hand, it is also essential for educational practice and policy to learn more about whether CLIL creates a learning environment that positively or even negatively affects academic self-concepts in different domains.

1.1. CLIL

The acronym CLIL points to the frequently mentioned *dual focus* of this bilingual approach that integrates content and foreign language learning (Coyle, 2006). A common rule of thumb states that the proportion of L2 instruction in the content subjects taught in CLIL programs is 50% or less. If the proportion exceeds 50%, it is referred to as immersion (Dalton-Puffer, 2011).

CLIL programs typically use a widely spoken language like English or French as the language of instruction. The range of subjects taught in CLIL programs includes almost all non-language subjects, with subjects such as history, geography, and biology dominating in practice (Eurydice, 2006, 2017).

In Germany, where the present study's data originate, CLIL programs are predominantly English-language and start at the secondary level in Grade 7 (Breidbach & Viehbrock, 2012; Rumlich, 2017). While participating in CLIL, CLIL students continue to attend regular English instruction. However, in preparation for CLIL, later CLIL students usually receive enhanced English instruction in Grades 5 and 6, which can lead to corresponding advantages over non-CLIL students, known as preparation effects.

Another characteristic of CLIL in Germany is its selective access (Rumlich, 2017). For instance, CLIL programs concentrate on academic-track schools (*Gymnasium*) in German secondary education, which are inherently selective and tend to attract students with favorable learning prerequisites and socioeconomic backgrounds (Breidbach & Viehbrock, 2012). Even within academic-track schools, students with better grades and higher motivation are more likely to participate in CLIL due to internal school criteria for CLIL participation (Rumlich, 2017).

In response to these insights, CLIL research has increasingly considered selection and preparation effects, which is essential to avoid overestimating the impact of CLIL on the outcomes of interest. Indeed, particularly for English skills, pre-existing differences between CLIL and non-CLIL students due to selection and preparation were found to contribute substantially to the observed advantages favoring the CLIL students.

In the German context, several studies consistently showed that CLIL students outperformed non-CLIL students in general English proficiency and English listening comprehension. However, when controlling for selection and preparation effects, the isolated effect of CLIL on students' English skills reduced substantially and partly no longer reached statistical significance (e.g., Dallinger, Jonkmann, Hollm, & Fiege, 2016; Feddermann, Baumert, & Möller, 2022, 2023). For example, Feddermann et al. (2021) examined English skill development of CLIL and non-CLIL students between Grades 7 and 8 in academic-track schools using the same longitudinal data set and comparable statistical analyses as in the present study. The authors found that selection and preparation effects could explain CLIL students' significantly higher English skills in Grade 8. By participating in CLIL, students between Grades 7 and 8 could maintain, but not further extend, the advantage they had built up through selection and preparation.

Similar results regarding English skill development were obtained by Goris, Denessen, and Verhoeven (2019) in a recent systematic review of longitudinal experimental studies in Europe, revealing that most of the included studies found no significant CLIL effect on English skills.

The extensive findings on CLIL effects on English skills contrast with a much smaller number of studies on motivational outcomes, like English self-concepts–although enhancing students' language learning motivation is also a declared goal of CLIL implementation (Eurydice, 2017). However, CLIL studies on English skills demonstrate that, regardless of the outcome of interest, it is indispensable to consider selection and preparation effects if the effects of CLIL are not to be overestimated.

1.2. Academic self-concepts

Academic self-concepts, defined as students' self-beliefs about their abilities in different school subjects or academic domains, are a central and widely studied construct in educational research. Numerous studies have shown that academic self-concepts positively predict key educational outcomes, such as academic achievement, motivation, and educational choice behaviors (Trautwein & Möller, 2016).

Despite their obvious relevance for learning- and achievement-related behavior, academic self-concepts have hardly been examined in CLIL research to date. However, it seems plausible that participation in CLIL can influence how competent students consider themselves in different domains, especially in the L2, to which they are significantly more exposed than non-CLIL students.

Research on the structure of academic self-concepts revealed that they are domain specific and consist of two distinct, empirically almost uncorrelated higher-level factors: verbal self-concept and math self-concept (Marsh et al., 2015; Marsh & Shavelson, 1985). Verbal self-concept includes self-concepts in verbally oriented subjects, such as self-concepts in the first language (L1) or the L2. Math self-concept, in contrast, includes self-concepts in the math and science domain, such as math or physics self-concepts. The near-zero correlation between verbal and math self-concept suggests that students view themselves as either a 'language person' or a 'math person' and is surprising given the usually high positive correlation between verbal and math achievement (Marsh & Hau, 2004). This finding led to the development of the internal/-external frame of reference model (I/E model; Marsh, 1986; for a meta-analysis, see Möller, Zitzmann, Helm, Machts, & Wolff, 2020), one of the most influential models on self-concept formation.

According to the I/E model, students develop domain-specific academic self-concepts in the verbal and math domains by comparing their achievement in one domain to their classmates' achievement in the same domain (external or social comparisons) and to their own achievement in the other domain (internal or dimensional comparisons). Social comparisons are assumed to lead to higher (lower) self-concepts when students compare their achievement with the worse-off (better-off) achievement of their classmates. Dimensional comparisons, which are at the center of DCT (Möller & Marsh, 2013), are assumed to lead to higher self-concepts in the intraindividually better domain and lower self-concepts in the intraindividually worse domain. A basic understanding of underlying processes of academic self-concept formation is relevant for the present study, as it can help clarify whether and how CLIL participation may influence students' self-concepts in the L2 and potentially also in math, as outlined in the following section.

1.3. CLIL-effects on academic self-concepts

1.3.1. CLIL-effects on L2 self-concepts

From the perspective of educational policy, practice, and research, CLIL is usually expected to positively affect motivation-related variables in the respective L2, such as students' L2 self-concepts. In this context, D. Marsh (2002) exemplarily states that CLIL provides important prerequisites for "cultivating a *can-do* attitude towards language learning" (p. 175).

The assumed positive effects of CLIL on L2 self-concepts are attributed, in particular, to CLIL students' higher L2 achievement levels. Indeed, a widely replicated finding of educational research is that better achievement leads to higher self-concepts within particular school subjects or academic domains (e.g., Möller et al., 2020). Accordingly, the typically higher English achievement of students in English-language CLIL programs compared to non-CLIL students–-whether due to CLIL or selection and preparation effects–could lead to correspondingly higher English self-concepts. In addition, CLIL has high prestige in Germany. Therefore, if CLIL students perceive themselves as part of a privileged group known for their high linguistic skills, a *basking in reflected glory effect* (BIRGE, Marsh, 1984) might emerge and also positively influence their English self-concepts. However, it is conversely conceivable that CLIL students' English self-concepts suffer from belonging to a high-achieving group because of possible unfavorable upward comparisons, as shown by studies on the *big fish little pond effect* (BFLPE; Marsh, 1987).

Previous CLIL studies considering self-concepts mostly revealed higher English self-concepts favoring the CLIL students. For example, in two longitudinal studies with German academic-track school students, Rumlich (2017) and Dallinger et al. (2016) found significantly higher English self-concepts for CLIL students compared to non-CLIL students in Grade 8 after two years of CLIL participation. Similarly, for the more English-intense immersion, Zaunbauer, Gebauer, Retelsdorf, and Möller (2013) showed that the English self-concepts of the studied second, third, and fourth graders in the immersion program exceeded those of their peers without immersion. Furthermore, Lo and Lo's (2014) meta-analysis from Hong Kong confirmed significantly higher English self-concepts for secondary students in immersion schools compared to students in regular schools (combined mean effect size: $M = 0.28$, $SE = 0.01$, $p < .01$). For the opposite case, a systematic negative impact of CLIL on English self-concept in terms of a BFLPE, there is no evidence apart from a single study (see Seikkula-Leino, 2007).

However, just as with English skills, the clear advantages in English self-concepts favoring students in CLIL or immersion could not be unambiguously attributed to participation in the bilingual program. Rumlich (2017), for example, found higher English self-concepts favoring CLIL students as early as Grade 6 before the start of CLIL, suggesting pre-existing group differences. Consistent with this, CLIL had no significant effect on students' English self-concepts in Grade 8 when combined selection and preparation effects were accounted for ($\beta = 0.10$, $p = 0.12$). Furthermore, the English self-concepts in Rumlich's (2017) study did not change significantly for either CLIL or non-CLIL students between Grades 6 and 8, and those of the elementary school-aged immersion and non-immersion students in Zaunbauer et al. (2013) increased comparably over the study period.

In summary, students in CLIL or immersion programs usually showed higher English self-concepts than their peers without CLIL or immersion, with the cause of the lead being selection and preparation effects rather than the programs themselves. However, the isolated impact of selection, preparation, and CLIL on this lead is still insufficiently understood. In the few CLIL studies on English self-concepts, selection and preparation effects were only jointly controlled for, if at all, but not disentangled.

1.3.2. CLIL-effects on math self-concepts

The strong focus of CLIL research on outcomes in the respective L2, whose promotion is the undoubted core goal of CLIL implementation (Eurydice, 2006), is reflected in a significant underrepresentation of studies on the effects of CLIL on educational outcomes in other subjects or academic domains. This is especially true for academic self-concepts. In this regard, a closer look at math self-concepts could provide an interesting new perspective for CLIL research and DCT (Möller & Marsh, 2013), as studies on DCT typically found negative effects of verbal achievement on math self-concepts (for a recent meta-analysis see Möller et al., 2020). Given the higher English achievement of CLIL students compared to non-CLIL students, it is therefore conceivable that CLIL might negatively affect students' math self-concepts. Furthermore, participation in the language-intensive CLIL may contribute to or reinforce a strong identification of CLIL students as 'language persons', which, in turn, could result in a devaluation of their math abilities, leading to lower math self-concepts.

However, whereas two studies on immersion found no differences in math self-concepts between immersion and non-immersion students (Lo & Lo, 2014; Zaunbauer et al., 2013), studies on the effects of CLIL on math self-concepts are lacking, pointing to the need for further research.

1.4. The present research

In the present study, we examined the effects of English-language CLIL on students' academic self-concepts in English and math. Previous research revealed that CLIL students usually show higher English self-concepts than non-CLIL students. However, combined selection and preparation effects were found to contribute significantly to CLIL students' English self-concept leads and not (just) participation in CLIL. In this respect, it is still to be determined what role selection and preparation effects play in isolation, as they have not yet been disentangled.

Regarding math self-concepts, DCT suggests that the typically higher English achievements of CLIL students compared to non-CLIL students might negatively affect their math self-concepts, but studies supporting this assumption are lacking so far.

Against this background, we posed the following research questions (RQ).

1. Does CLIL positively affect students' English self-concepts?
2. Does CLIL negatively affect students' math self-concepts?

Based on prior findings, we expected CLIL students to have higher English self-concepts than non-CLIL students. Furthermore, we assumed that advantages favoring the CLIL students were not only due to CLIL but also to selection and preparation effects, which we therefore explicitly considered in the analyses and, as the first CLIL study on self-concepts, separated from each other. Because there has hardly been any research on the effects of CLIL on math self-concept, we had no prior assumptions regarding this research question.

2. Method

2.1. Sample

This study was based on data from the Competencies and Attitudes of Students Study (KESS; Bos & Pietsch, 2006), a panel study conducted in the northern German city state of Hamburg. The primary focus of KESS is the systematic and continuous assessment of an entire Hamburg student cohort's subject-specific achievements and attitudes. During the study period between 2003 and 2012, the student cohort was examined five times, from the end of elementary school in Grade 4 (KESS 4), through Grades 7 (KESS 7), 8 (KESS 8), and 10/11 (KESS 10/11), to

Grade 12/13 (KESS 12/13). KESS was commissioned by the Ministry of Schools and Vocational Training in Hamburg, which was also responsible for reviewing research ethics and privacy concerns. The Institutional Review Board of the Ministry of Schools and Vocational Training granted approval for the study. Parents were required to provide written informed consent. Student and parent questionnaires were voluntary. Students and parents suffered no disadvantage from nonparticipation. The achievement tests for students were obligatory. Different institutions were involved in data collection and processing at the five measurement waves, including the International Association for the Evaluation of Educational Achievement (IEA) and the State Institute for Teacher Education and School Development in Hamburg (Bos & Gröhlich, 2010). The testing usually took place on two consecutive days during regular instruction. Data were assessed using standardized measurement instruments successfully employed in prior large-scale educational monitoring studies. All instruments, along with detailed information regarding their reliability and validity, have been comprehensively documented and published for each measurement wave (see, e.g., Bos, Gröhlich, Dudas, Guill, & Scharenberg, 2011).

CLIL participation was assessed in KESS 7 and 8. Since CLIL was only offered at academic-track schools then, our analyses were limited to this school type. The students examined in KESS received regular English instruction starting in Grade 3. In the academic-track schools, later CLIL students received enhanced English instruction of at least 6 h weekly in Grades 5 and 6, typically between one and 2 h per week more than non-CLIL students. CLIL started in Grade 7 with at least 3 h per week and usually two to three bilingually taught subjects through Grade 10. Math was not taught in English.

This study used data from KESS 4, 7, and 8, thus comprising 6020 students examined at the end of Grade 4, at the beginning of Grade 7, immediately at the start of CLIL, and at the end of Grade 8, after almost two school years of CLIL. As an inclusion criterion, information on CLIL participation had to be available. After excluding CLIL students participating at only one measurement point ($n = 9$), non-English CLIL students ($n = 45$), and invalid cases ($n = 3$), we obtained a final sample of 5963 students from 256 elementary schools and 66 academic-track schools (51.2 % female, age at Grade 8: $M = 14.5$, $SD = 0.53$), including 385 CLIL students.

Of the academic-track schools, six had optional CLIL classes and two were purely bilingual schools in which all students participated in CLIL.

2.2. Measures

2.2.1. Academic self-concepts

English and math self-concepts in Grades 7 and 8 were measured with well-established scales (Jerusalem, 1984; Jopt, 1978). English self-concept was assessed with five items, and math self-concept with four items. Except for the target domain, the item wording for English and math was comparable (e.g., "I'm just not good at English/math"). A complete list of all items is presented in Table A1 the Appendix. Students responded to each item on a 4-point Likert scale ranging from 1 = *strongly agree* to 4 = *strongly disagree*. Higher scores indicated higher self-concepts for all items in English and math.

Reliabilities were good for both subjects in Grades 7 and 8 (English: $\omega = 0.92$; math: $\omega = 0.90$).

2.2.2. CLIL attendance

Participation in CLIL (0 = *no*, 1 = *yes*) was assessed with one item.

2.2.3. Academic achievement

Grade 7 school grades and standardized test scores in English and math served as achievement measures and were included as control variables. For the achievement tests, we used available person parameters obtained from weighted maximum likelihood estimators (WLE) that were estimated based on 2-PL item response theory models (Feddermann et al., 2019).

Grades. School grades in English and math were obtained from the student participation lists. Since KESS 7 took place at the beginning of Grade 7, when grades were not yet available, we used grades from the end of Grade 6. In addition, we formed cluster means to measure class achievement in both domains. The German grading system includes grades from 1 (*excellent*) to 6 (*failed*). For the analyses, all grades were reverse-coded so that higher scores indicated higher achievement.

English Test Scores. Students' general English proficiency was measured using C-tests. The C-Test consisted of 77 items respectively words to be completed. The reported WLE reliability was 0.94 (Feddermann et al., 2019).

Math Test Scores. Students' math achievement was assessed with 72 items covering different mathematical content areas. The reported WLE reliability was 0.90 (Feddermann et al., 2019).

2.2.4. Baseline covariates

For the successful implementation of PSM, we included a comprehensive set of covariates measured before the start of CLIL in Grade 4. Covariates comprised demographics, students' socioeconomic backgrounds, language-related variables, attendance at preschool educational institutions, cognitive abilities, achievement, and motivation-related variables (see Section S1 in the Supplemental Material for detailed descriptions of all covariates). Descriptive statistics are presented in Table A2 in the Appendix.

2.3. Statistical analysis

2.3.1. Missing data

Missing data ranged from 24.7% to 63.3% for the covariates in Grade 4 and from 8.8% to 81.6% for the central predictors and outcome variables in Grades 7 and 8. Thus, we used multiple imputation and generated 100 complete data sets. Subsequent analyses were performed with all 100 data sets, and the results were pooled according to Rubin's rules (Rubin, 1987) (see Section S2 in the Supplemental Material for further details).

2.3.2. Propensity score matching

We applied PSM to account for the non-randomized assignment of students to CLIL, thus adjusting CLIL effects on academic self-concepts for possible selection bias. Given the selective access to CLIL, it seemed reasonable to estimate CLIL effects only for the population typically participating in CLIL (i.e., the *average treatment effect on the treated*; ATT), rather than for the overall student population.

PSM was conducted based on 23 covariates associated with both CLIL assignment and students' English and math self-concepts (Caliendo & Kopeinig, 2008) (see Table A2 in the Appendix for further details on the covariates). Using nearest neighbor (NN) matching with replacement, we matched non-CLIL and CLIL students with a ratio of 5:1 (i.e., assigning five non-CLIL students to one CLIL student) and a caliper–the maximum tolerated distance between the matching partners–of 0.25. This matching procedure was convincing regarding the achieved balance of the covariate and propensity score distributions (for alternative matching procedures see Table S1 in the Supplemental Material).

To evaluate whether matching succeeded in creating balanced covariate distributions between the groups, we inspected standardized mean differences between CLIL and non-CLIL students for all 23 covariates and the propensity score before and after matching. We calculated the standardized mean differences like effect sizes by dividing the difference in means of each covariate in the CLIL and non-CLIL group before and after matching by the standard deviation in the entire unmatched sample. As criteria to evaluate covariate balance between the groups, we applied cut-offs for the standardized mean differences proposed by the What Works Clearinghouse (What Works Clearinghouse, 2022) and the PSM literature of >0.25 for non-equivalent groups, 0.05–0.25 for necessary statistical adjustment, and <0.05 for equivalent groups.

PSM was conducted using the R package MatchThem (Pishgar, Greifer, Leyrat, & Stuart, 2021), which is suitable for multiply imputed datasets.

2.3.3. Estimating CLIL effects on student' academic self-concepts

We used structural equation modeling to estimate the effects of CLIL on students' English and math self-concepts in Grade 8. Structural equation modeling allowed the simultaneous modeling of the effects on both dependent variables in a joint model and accounting for measurement error in English and math self-concepts. To separate the CLIL effects from and disentangle possible selection and preparation effects, we ran three models: First, we set up a model in which self-concepts in English and math were regressed on CLIL participation based on the unmatched sample (Model 1). Subsequently, we computed the same model based on the matched sample, thus controlling for selection effects (Model 2). Finally, Model 2 was extended by adding English and math self-concepts and achievement measured at the beginning of Grade 7 (Model 3) to control for preparation effects due to increased English instruction in Grades 5 and 6. Path diagrams representing the models are depicted in Fig. 1.

To address residual bias, we chose a *doubly robust* approach by additionally controlling for all covariates used in the PSM (Schafer & Kang, 2008).

Structural equation modeling was performed using Mplus 8.4 (Muthen and Muthen, 1998) via the R package MplusAutomation (Hallquist & Wiley, 2018). We used the *type = complex* option in Mplus to adjust the standard errors for the clustering of students within classes.

As preliminary analyses, we conducted tests of measurement invariance for English and math self-concept across groups (CLIL and non-CLIL) and time (Grades 7 and 8). Following Chen's (2007) recommendations, scalar measurement invariance across groups and partial measurement invariance across time was acceptable for English and math self-concept.

3. Results

3.1. Sample differences before and after PSM

Fig. 2 shows the differences between CLIL and non-CLIL students at the end of Grade 4 before and after matching. In the unmatched sample, the absolute standardized mean differences of the covariates between the groups ranged from 0.003 for gender to 0.425 for recommended type of secondary education, with an average absolute standardized mean difference of 0.244 (see also Table A2 in the Appendix).

Using the criterion of >0.25 for non-equivalent groups (What Works Clearinghouse, 2022), CLIL and non-CLIL students differed most regarding their family background, cognitive abilities, and achievement: Compared to non-CLIL students, CLIL students had more favorable socioeconomic backgrounds as reflected by higher family income and HISEI. In addition, CLIL students scored higher on verbal and figural intelligence tests, outperformed non-CLIL students on school grades and standardized test scores in English, math, German, and science, showed higher math interest and reading self-concepts, and were more likely to receive a academic-track school recommendation. These results point to a substantial level of positive selection favoring the CLIL students. In line with that, the propensity score, representing the probability of attending CLIL given the observed covariates, was significantly higher in the CLIL group than in the non-CLIL group, as expressed by a standardized mean difference of 0.736.

Although the propensity score distributions differed markedly between the groups before matching, they overlapped (see Fig. A1 in the Appendix). The overlapping area, the so-called area of common support, indicates the range of CLIL and non-CLIL students with comparable covariate distributions who could reasonably be included in the matching procedure and for whom it was justified to estimate CLIL effects (Thoemmes & Kim, 2011). In our case, the propensity score distribution of the non-CLIL students almost completely covered that of the CLIL students, meaning that for almost every CLIL student, there was a comparable non-CLIL student, which is a basic requirement for estimating the ATT (Stuart, 2010).

After PSM, CLIL and non-CLIL students no longer differed significantly on any of the covariates (see Fig. 2 and Table A3 in the Appendix). With a range from <0.001 to 0.010, the absolute standardized mean differences of the covariates were all well below the predefined threshold of 0.05 for equivalent groups, indicating excellent covariate balance (What Works Clearinghouse, 2022). The absolute standardized mean difference for the propensity score decreased from 0.736 before matching to <0.001 after matching, resulting in almost identical propensity score distributions in both groups (see Fig. A1 in the Appendix). The final matched sample size ranged from 1808 to 1920 students (M = 1862.7 SD = 24.3), including 379 to 385 CLIL students (M = 383.2, SD = 1.21), depending on the imputed data set.

3.2. CLIL effects on English and math self-concepts

To estimate the effect of CLIL on students' academic self-concepts, we used structural equation modeling with Grade 8 English and math self-concepts as dependent variables and CLIL attendance as the independent variable. Table 1 shows the results as standardized and unstandardized regression coefficients (see Table S3 in the Supplemental Material for regression coefficients of the control variables).

Model fit was good in the unmatched sample (χ^2 = 423.558, df = 33, CFI = .986, TLI = .981, RMSEA = .045, SRMR = .019) and in the matched sample (χ^2 = 344.414, df = 194, CFI = .984, TLI = .980, RMSEA = .011, SRMR = .012)

For English, structural equation modeling without prior matching (Model 1) revealed a significant positive CLIL effect of b = 0.34 (p < .001) on English self-concept. Accordingly and in line with our expectations, CLIL students showed significantly higher English self-concepts than non-CLIL students in Grade 8. After accounting for selection effects using the matched sample (Model 2), we found a smaller but still significant positive CLIL effect of b = 0.22 (p < .001) on students' English self-concepts. However, when additionally controlling for prior self-concepts and achievement in English and math measured in Grade 7 immediately after the end of the later CLIL students' preparatory English instruction (Model 3), the CLIL effect on Grade 8 English self-concept was no longer significant (b = 0.11, p = .090). In fact, even after PSM, CLIL students' English self-concept was significantly higher than that of the non-CLIL students at the beginning of Grade 7 (d = 0.31, p = .008), as illustrated in Fig. 3 (see also Table S2 in the Supplemental Material). Since CLIL did not start until Grade 7, this effect can be seen as the result of later CLIL students' enhanced English instruction in Grades 5 and 6, indicating preparation effects. In line with that, there were no differential developmental trajectories for the English self-concepts of CLIL and non-CLIL students in the matched sample, as both groups showed a comparable increase in English self-concept between Grades 7 and 8 (see Fig. 3).[1] Accordingly, CLIL students did not increase their English self-concept lead further but held it over the two school years.

In sum, CLIL attendance had no significant positive effect on Grade 8 English self-concepts beyond pre-existing group differences (RQ 1). In

[1] An attentive reviewer noted that ANCOVA models and change score models can lead to different results when comparing changes across groups - known as the Lord's paradox. To double check our claim that "there were no differential developmental trajectories for the English [and math] self-concepts of CLIL and non-CLIL students in the matched sample", we employed latent change score modeling as an additional robustness check. Latent change score modeling also implied no differential development between CLIL and non-CLIL students' English and math self-concepts in the matched sample (details can be found in Tables S4 and S5 and Fig. S1 the Supplemental Material).

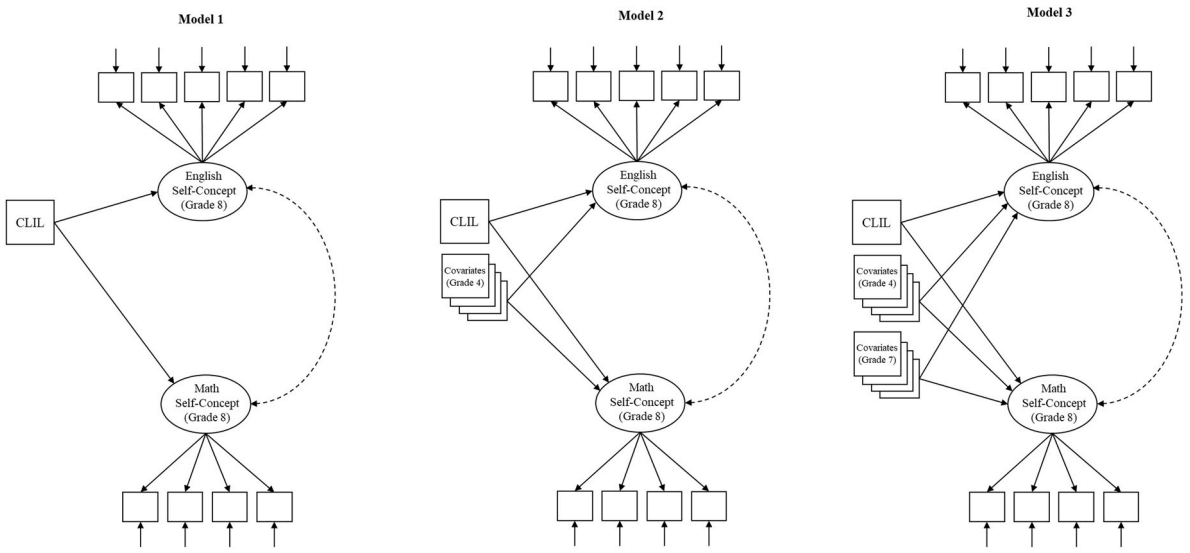


Fig. 1. Path diagrams representing the analysis models used to estimate CLIL effects on English and math self-concepts in Grade 8. Model 1: Regressing English and math self-concept on CLIL in the unmatched sample without considering covariates. Model 2: Regressing English and math self-concept on CLIL in the matched sample, thus considering selection effects, and additionally controlling for all covariates from Grade 4 used as predictors in the matching procedure to obtain doubly robust estimates. Model 3: Regressing English and math self-concept on CLIL in the matched sample, controlling for all covariates from Grade 4 to obtain doubly robust estimates, and additionally controlling for Grade 7 English and math self-concepts and achievements, thus considering selection and preparation effects. CLIL = Content and Language Integrated Learning

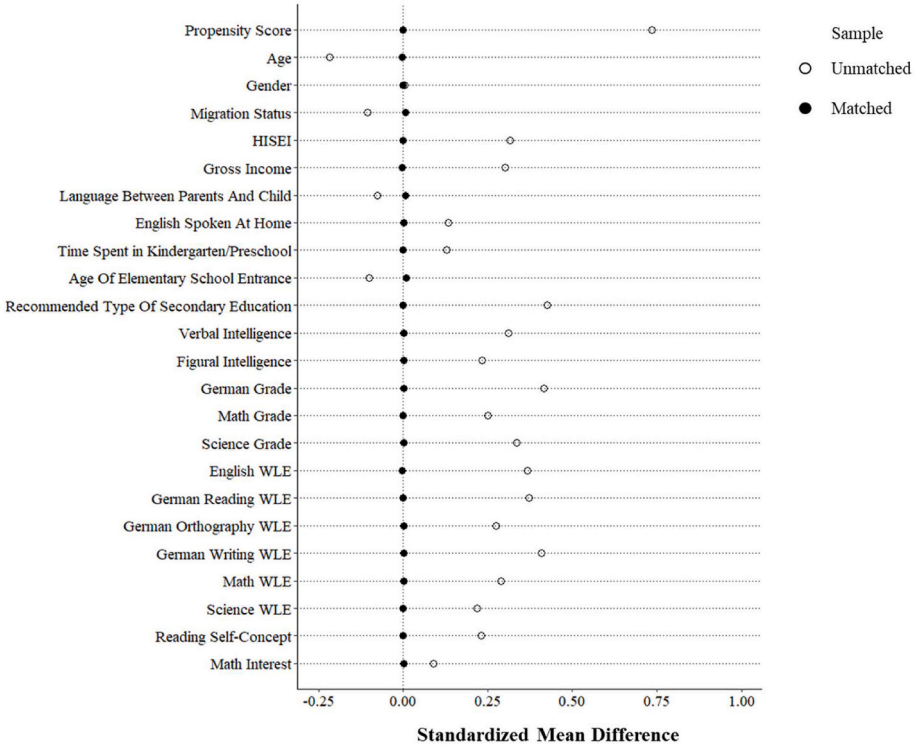


Fig. 2. Standardized mean differences between CLIL and Non-CLIL students before and after matching. Mean differences were standardized using the standard deviation of the entire unmatched sample and are presented as average across all 100 imputed data sets. CLIL = Content and Language Integrated Learning; HISEI = Highest International Socio-Economic Index of Occupational Status in the family; WLE = Weighted likelihood estimator.

this regard, it became evident that both selection and preparation effects contributed to CLIL students' significant advantage in English self-concept. Nevertheless, participating in CLIL helped students maintain their advantage built up through selection and preparation.

For math, a different pattern of results emerged. Here, CLIL was found to have no significant effect on students' math self-concepts either in the unmatched sample (Model 1) or after accounting for selection (Model 2) and preparation effects (Model 3) ($-0.04 \leq b \leq 0.06$, $ps \geq .125$) (see Table 1 and Table S3 in the Supplemental Material). Accordingly, contrary to our speculations, participation in CLIL was not at the expense of math self-concept (RQ 2).

Table 1
CLIL effects on English and math self-concepts in Grade 8.

Effects	Model 1				Model 2 [a]				Model 3 [a]			
	B (β)	*SE*	*p*	R^2	*B* (β)	*SE*	*p*	R^2	*B* (β)	*SE*	*p*	R^2
CLIL → ESC	**0.34 (0.12)**	0.05	<.001	.01	**0.22 (0.13)**	0.05	<.001	.17	0.11 (0.07)	0.07	.090	.39
CLIL → MSC	0.06 (0.02)	0.04	.125	.00	−0.04 (−0.02)	0.04	.312	.30	0.02 (0.01)	0.07	.761	.56

Note. All coefficients are averaged across the 100 imputed data sets. Significant results are printed in bold. Model 1: Estimates using the unmatched sample. Model 2: Estimates using the matched sample, thus considering selection effects. Model 3: Estimates using the matched sample and controlling for prior self-concepts and achievement in English and math, thus considering selection and preparation effects. CLIL = Content and Language Integrated Learning; ESC = English self-concept; MSC = Math self-concept.

[a] Double robust estimates by additionally considering all covariates used in the matching procedure as predictors in the regression model.

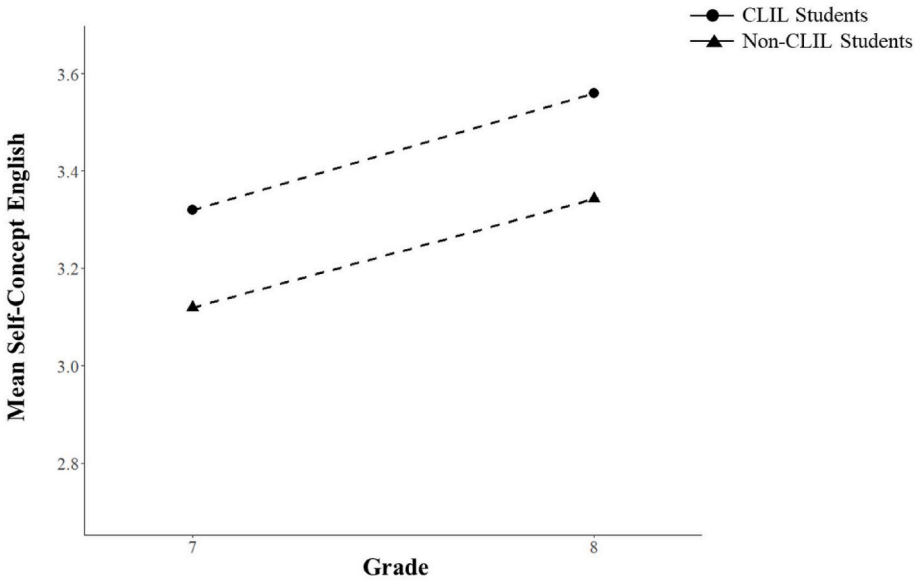


Fig. 3. CLIL and Non-CLIL students' English self-concept development between Grades 7 and 8 after matching. Displayed are the latent English self-concept means calculated based on the matched sample. CLIL = Content and Language Integrated Learning.

4. Discussion

The present study aimed to investigate the effects of CLIL on academic self-concepts in English and math. By examining academic self-concepts, we focused on a key motivational construct that has received little attention in CLIL research so far. Using data from a comprehensive panel study and applying PSM, we followed the call for longitudinal data and adequate statistical analysis methods to obtain unbiased estimates of CLIL effects.

The findings, their relations to previous research, and implications for future research, educational practice, and policy are discussed below.

4.1. Identifying the isolated effect of CLIL on academic self-concepts

Without accounting for selection and preparation effects, CLIL students, consistent with our assumptions, had significantly higher English self-concepts than non-CLIL students in Grade 8 after nearly two school years of CLIL participation. Their math self-concepts, however, were not lower than those of the non-CLIL students, as we had speculated.

The analysis of group differences before PSM in Grade 4 revealed that the later CLIL students, already by the end of elementary school, well before the start of CLIL, had more favorable socio-economic backgrounds and outperformed the later non-CLIL students in terms of cognitive abilities, achievement in English and other subjects, and motivation. This is in line with previous studies demonstrating the selectivity of CLIL programs (e.g. Dallinger, Jonkmann, & Hollm, 2018) and with criteria reported in the literature to be relevant for CLIL participation in Germany (e.g., Breidbach & Viehbrock, 2012).

After applying PSM to account for selection effects, we found a smaller but still significant positive effect of CLIL on the eighth graders' English self-concepts but no effect on their math self-concepts. However, we also had to consider CLIL students' increased English instruction in Grades 5 and 6, which could have caused the observed advantage in English self-concept. Consistent with this, CLIL students' English self-concept was significantly higher than that of the non-CLIL students as early as the beginning of Grade 7, immediately at the start of CLIL, which we attributed to the preparatory instruction. To adjust the CLIL effect for preparation effects, we additionally controlled for self-concepts and achievement in English and math measured at the beginning of Grade 7. In doing so, we found that CLIL no longer significantly affected students' English self-concepts in Grade 8. These results underscore that in future studies, it is not sufficient to consider only selection or preparation effects but both, as both contributed to CLIL students' significant English self-concept lead.

In summary, our study showed that CLIL had neither an additional effect on students' English self-concepts after accounting for selection and preparation effects (RQ 1) nor did it impact their math self-concepts (RQ 2). However, CLIL students could maintain their advantage in English self-concept during the first two school years of CLIL participation.

Regarding the English self-concept, our results correspond to those of Rumlich (2017), who also found no significant CLIL effects on academic-track school students' English self-concepts in Grade 8 when controlling for a priori differences between CLIL and non-CLIL students.

Furthermore, in line with our findings, the immersion study by Zaunbauer et al. (2013) showed similar developmental trajectories of English self-concept for elementary school-aged immersion and non-immersion students.

Apart from that, there are also noticeable parallels with studies examining CLIL effects on English skills. These studies repeatedly found that positive CLIL effects on English skills disappeared altogether or at least substantially diminished when a priori group differences were considered (e.g., Dallinger et al., 2016). Analogous to our results on English self-concept, Feddermann et al. (2021), who also examined KESS data with comparable statistical analyses, found that CLIL had no additional positive effect on global English proficiency after controlling for selection and preparation effects. Similarly, there was an increase in English proficiency for CLIL and non-CLIL students that did not differ significantly between the groups. Accordingly, CLIL here, as in our study, contributed to maintaining, but not improving, the English proficiency lead built up through selection and preparation.

Regarding math self-concept, there are no comparable CLIL studies so far. Therefore, our study makes a first important contribution here by suggesting that attending CLIL does not harm math self-concept but leaves it unaffected. These results are consistent with findings from the immersion study by Zaunbauer et al. (2013) and Lo and Lo's (2014) meta-analysis on immersion which found no differences between immersion and non-immersion students' math self-concepts. However, given our analyses' exploratory nature and the scarcity of comparable CLIL studies, it is important to examine the replicability of the results in future research.

While the lack of negative CLIL effects on math self-concept is undoubtedly desirable, the absence of positive effects on English self-concept points to a discrepancy between the widely postulated positive impact of CLIL on educational outcomes in the L2 and its empirically demonstrable effects–especially since this applies not only to self-concepts but also to L2 skills. This raises the question of why CLIL seems to fall short of the potential attributed to it by educational policy, practice, and research.

A straightforward answer in this context might be that CLIL, at least in the form currently implemented in Germany, is unsuitable for sufficiently promoting English self-concept and skills due to basic structural features of its implementation. In this context, the selective access to German CLIL programs, mainly offered at academic-track schools and tending to attract high-achieving and linguistically gifted students with already high levels of English self-concept and skills before the start of CLIL, could play a role by limiting the opportunities for further improvement. Accordingly, increasing the effectiveness of CLIL in terms of English self-concept may require a targeted and broader expansion of CLIL to students with lower English self-concept and skills in non-academic tracks. However, such an explanatory approach falls somewhat short and denies CLIL the potential to promote English self-concepts and skills within the student population typically participating in CLIL.

Another more differentiated explanation explicitly concerning English self-concept is provided by Rumlich (2018), who showed that CLIL students obtained lower English grades than their peers without CLIL for the same level of English achievement. A similar trend was found in a side analysis of the present study, suggesting that English grades did not adequately reflect the significant differences in English achievement favoring the CLIL students as measured by the achievement test in Grade 8. Suppose teachers tend to systematically underestimate CLIL students' English achievement due to a BFLPE and report this back to students as unjustifiably low grades. In this case, this could prevent a positive English self-concept development in CLIL programs and make it difficult for studies like ours to detect positive CLIL effects (Rumlich, 2018).

Furthermore, it is also conceivable that CLIL does not positively affect self-concept regarding global English proficiency, but it may positively affect specific sub-facets thereof. For example, some CLIL studies on English skills found significant positive effects of CLIL on English listening comprehension but not global English proficiency (e.g., Dallinger et al., 2016). Therefore, positive CLIL effects on English self-concept might likewise emerge when it is operationalized as self-evaluation of specific English skills rather than of global English proficiency, as in our case.

Finally, although CLIL had no additional positive effect on English self-concept, it should be acknowledged that CLIL students maintained their lead over non-CLIL students between Grades 7 and 8. That is remarkable in that, first, CLIL students already started with a high English self-concept level, which even increased over the two school years. Second, CLIL students could maintain their lead despite an expectable negative influence of a BFLPE, which we did not explicitly investigate but which is quite reasonable due to belonging to the high-performing CLIL group. Thus, it would be highly interesting to explicitly investigate BFLPE as well as a possible interplay of BFLPE and BIRGE in CLIL classes in future studies.

4.2. *Theoretical and practical implications*

By investigating the effects of CLIL on students' English and math self-concepts, this study provides new insights and extends the body of knowledge of both CLIL and self-concept research: Regarding CLIL research, the present study, firstly, makes a significant contribution to the existing knowledge on the effectiveness of CLIL concerning motivational student characteristics. Although CLIL is often described as a promising approach to increase student motivation in the respective L2 (Eurydice, 2017), motivational constructs, especially CLIL students' L2 self-concepts, have hardly been investigated so far. Instead, the focus of previous research has been primarily on CLIL students' achievement in the L2. Accordingly, our study is one of the first to explicitly examine and deepen the understanding of the effects of CLIL on English self-concept. Secondly, this study is, as far as we know, the very first to provide knowledge on the impact of CLIL on math self-concept, suggesting that participating in language-intensive CLIL is not associated with an increased devaluation of students' math abilities, as we found no differences in the math self-concept between CLIL and non-CLIL students. Thirdly, by considering and, for the first time, disentangling the effects of selection and preparation within a CLIL study on self-concepts, we were able to demonstrate that the significant advantages of CLIL students in English self-concept compared to non-CLIL students could be explained by pre-existing differences, as has already been shown for English achievement (e.g., Dallinger et al., 2016). This insight is crucial for future studies and underscores the necessity to control for both selection and preparation effects to avoid overestimating the true impact of CLIL, regardless of the outcome of interest.

Regarding self-concept research, this study's theoretical contribution lies particularly in providing insights into how a specific type of instruction, namely the bilingual approach CLIL, firmly established in most European countries (Eurydice, 2017), may influence students' self-concepts in the verbal and math domains. Our findings firstly indicate that CLIL can help maintain but not enhance existing advantages in English self-concept despite–or even due to–the high-achieving and prestigious environment of CLIL classes. Secondly, despite the strong language focus associated with CLIL, no disadvantages in math self-concept occurred. Therefore, this study links to different central theoretical models and phenomena in the field of self-concept research, including the I/E model, DCT, the BFLPE, and the BIRGE, which represent an interesting field of research for future CLIL studies.

Apart from theoretical contributions, the present study also provides different implications for educational practice and policy. Our results confirm the findings of other recent CLIL studies that have shown limited additional positive or no effects of CLIL on English self-concept and achievement after accounting for selection and preparation effects. While not questioning the usefulness of CLIL altogether, the results indicate that positive CLIL effects on educational outcomes in English are not automatic and that CLIL could benefit from improvements in current practice.

This insight has important implications for those involved in the practical implementation of CLIL, such as CLIL teachers and school administrators, as well as for educational policy makers, which set the overall goals and framework of CLIL implementation. Furthermore, as CLIL students continue to attend regular English instruction, our findings are also relevant for English teachers, who assign English grades and may also play a role in CLIL students' English self-concept development.

Regarding English teachers, a starting point to improve the status quo may be the findings of Rumlich (2018) on English grades. Rumlich (2018) revealed that CLIL students received lower grades in English than non-CLIL students for the same achievement level, indicating BFLPE in grading. Assuming further studies confirm these findings, an important step for educational practice, as suggested by the author, might be to raise special awareness of and emphasize BFLPE and achievement-adequate feedback in teacher training and in-service training to mitigate potentially unfavorable effects on CLIL students' English self-concepts.

Apart from that, BFLPE can also directly and negatively influence CLIL students' English self-concepts due to the possibility of unfavorable social upward comparisons within the high-achieving environment of CLIL classes. In this context, a concrete strategy that could prove effective in enhancing English self-concept is encouraging both English and CLIL teachers to use an individualized frame of reference (e.g., Lüdtke, Köller, Marsh, & Trautwein, 2005) when evaluating students' English skills and providing achievement-related feedback. Unlike teachers who use a social frame of reference, evaluating a student's achievement in comparison to other students, teachers applying an individualized frame of reference consider past achievement and effort and tend to provide feedback that emphasizes a student's intra-individual learning and achievement progress over time. Although not eliminating the negative impact of BFLPE, individualized teacher feedback is assumed to be highly beneficial for promoting students' self-concepts and motivation in general (Dickhäuser, Janke, Praetorius, & Dresel, 2017; Lüdtke et al., 2005). Concerning CLIL teachers, it is important to note that the dual focus of CLIL involves integrating L2 and content learning, so the improvement of L2 skills is one but not the sole focus. Consequently, another way for CLIL teachers to promote CLIL students' English self-concepts could be to place greater emphasis on progress in L2 learning than on content learning in CLIL subjects. However, such a measure is more of a normative nature and could undermine the concept of CLIL, which is not designed as traditional L2 instruction but rather as an instructional approach where content is learned by means of an L2 (Dalton-Puffer, 2011).

Furthermore, the effectiveness of CLIL with regard to English self-concept could benefit from more fundamental changes in current CLIL implementation, which includes an expansion of CLIL to other target groups besides the typically participating students. Since CLIL programs in Germany and related research focused primarily on linguistically gifted, high-achieving, and motivated students in selective CLIL programs in academic-track schools, it might prove promising to facilitate access to CLIL further and expand appropriate CLIL programs for students with lower English skills and motivation in non-academic tracks. For example, regarding English skills, Goris et al. (2019) showed in their systematic review that in Spain, where general English proficiency is relatively low and CLIL targets all students, not just those with already favorable learning conditions, significant positive CLIL effects on English proficiency were more common than in most other European countries. Consequently, English self-concept might benefit likewise if CLIL (also) targets students with low English proficiency and language learning motivation, for whom large gains are possible in both areas.

4.3. Limitations and Directions for future research

When interpreting the results of the present study, some limitations must be considered, one of which concerns the results' generalizability. Firstly, our analyses were limited to Hamburg students, so our results are not readily transferable to other federal states and, in particular, to other countries and education systems outside Germany, as the CLIL implementation varies considerably between countries and education systems (Eurydice, 2017). Secondly, we studied only academic-track school students. Hence, future research needs to determine to what extent our results are transferable to intermediate-track and low-track school students who have been rather underrepresented in previous CLIL studies. Thirdly, our study joins other German studies that have examined the effects of CLIL in lower secondary education, especially in Grade 8 (e.g., Rumlich, 2017), which typically corresponds to the second year of CLIL attendance. Since many German academic-track schools offer CLIL until graduation in Grade 12 or 13, it might be promising to investigate the effects of CLIL at a later stage of CLIL attendance. It is conceivable that CLIL unfolds effects on academic self-concepts–positive or negative–only after longer participation.

Furthermore, the replicability of the results should be tested with alternative self-concept scales for which strict measurement invariance across time can be assumed to draw valid conclusions about the change in the latent means of the self-concepts. However, recent simulation studies suggest that partial measurement invariance, as found in our study, is sufficient for robust comparisons of latent means and path coefficients (Pokropek, Davidov, & Schmidt, 2019).

Finally, there are also limitations concerning the PSM. The potential of PSM to eliminate selection bias depends on whether all covariates associated with the selection process and the outcomes have actually been measured and included in the matching procedure (Caliendo & Kopeinig, 2008). Particularly relevant here might be that we could not include pretreatment measures of English and math self-concepts as these were not assessed in Grade 4. Since accounting for pretreatment measures of the outcomes is often considered central to eliminating selection bias (see e.g., Cook & Steiner, 2010), future studies on the effects of CLIL on academic self-concepts should, whenever possible, rely on data in which self-concept was examined before CLIL began.

4.4. Conclusion

This study makes an important contribution to CLIL research by shedding light on the hitherto scarcely studied effects of CLIL on academic self-concepts. In examining the effects of CLIL on both English and, for the first time, math self-concepts, the present study provides insights into the extent to which participation in a specific bilingual program influences self-concepts in the verbal and math domains, thus also enriching self-concept research. The particular strengths of the study, including longitudinal data from an entire Hamburg student population, a large number of available covariates, and the use of PSM, allowed us to adjust the net effects of CLIL for selection and preparation effects and to disentangle them. Our results suggest that attending CLIL can help maintain existing advantages in English self-concept while not negatively affecting math self-concept. However, the advantages favoring CLIL students could be explained by selection and preparation effects, not by CLIL, which had no additional positive effect on English self-concept. In this context, the study also highlights the importance of accounting for selection *and* preparation effects in future CLIL studies, as both contributed to the significant a priori differences between CLIL and non-CLIL students.

These findings are not only relevant for educational research but also for teachers, teacher educators, and educational policy responsible for the practical implication and organization of CLIL. Since the CLIL students examined in our study could not further extend their advantages in English self-concept over the non-CLIL students, a central challenge and, simultaneously, a promising opportunity for both future research and educational practice lies in identifying the strengths and shortcomings of current CLIL implementation to pave the way for ongoing improvement of CLIL. This seems all the more significant as innovative approaches to bilingual instruction such as CLIL will continue to be of great importance

in the future given an increasingly globalized world, internationally oriented education systems and labor markets, and multilingual and multicultural societies, as exemplified in Europe (Eurydice, 2017).

Declaration of competing interest

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

Acknowledgements

This research was supported by German Research Foundation Grant MO 648/26–1 awarded to Jens Möller. The present paper used longitudinal data from the KESS study, which was commissioned by the Ministry of Schools and Vocational Training of the Free and Hanseatic City of Hamburg. The data set has been provided to the scientific consortium MILES (Methodological Issues in Longitudinal Educational Studies) for a limited period to conduct in-depth examinations of scientific questions. MILES is coordinated by the Leibniz Institute for Science and Mathematics Education.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2024.101923.

References

Bos, W., & Gröhlich, C. (Eds.). (2010). *KESS 8 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8 [KESS 8 - Competencies and attitudes of students at the end of grade 8]*. Münster: Waxmann.

Bos, W., Gröhlich, C., Dudas, D.-F., Guill, K., & Scharenberg, K. (2011). KESS 8 - Skalenhandbuch zur Dokumentation der Erhebungsinstrumente [KESS 8 - Scale manual for the documentation of the survey instruments]. *HANSE - Hamburger Schriften zur Qualität im Bildungswesen*. Münster: Waxmann.

Bos, W., & Pietsch, M. (Eds.). (2006). *KESS 4 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen [KESS 4 - Competencies and attitudes of students at the end of grade 4 in Hamburg elementary schools]*. Münster: Waxmann.

Breidbach, S., & Viehbrock, B. (2012). CLIL in Germany: Results from recent research in a contested field of education. *International CLIL Research Journal, 1*(4), 5–16.

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods, 15*(1), 56–68. https://doi.org/10.1037/a0018536

Coyle, D. (2006). Content and language integrated learning: Motivating learners and teachers. *Scottish Languages Review, 13*(5), 1–18.

Dallinger, S., Jonkmann, K., & Hollm, J. (2018). Selectivity of content and language integrated learning programmes in German secondary schools. *International Journal of Bilingual Education and Bilingualism, 21*(1), 93–104. https://doi.org/10.1080/13670050.2015.1130015

Dallinger, S., Jonkmann, K., Hollm, 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. https://doi.org/10.1016/j.learninstruc.2015.09.003

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

Dickhäuser, O., Janke, S., Praetorius, A.-K., & Dresel, M. (2017). The effects of teachers' reference norm orientations on students' implicit theories and academic self-concepts. *Zeitschrift für Padagogische Psychologie, 31*(3–4), 205–219. https://doi.org/10.1024/1010-0652/a000208

Eurydice. (2006). Content and language integrated learning (CLIL) at school in Europe. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/756ebdaa-f694-44e4-8409-21eef02c9b9b.

Eurydice. (2017). Key data on teaching languages at school in Europe. Retrieved from https://op.europa.eu/o/opportal-service/download-handler?identifier=ff10cc21-aef9-11e7-837e-01aa75ed71a1&format=pdf&language=en&productionSystem=cellar&part=.

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

Feddermann, M., Baumert, J., & Möller, J. (2023). A replication study to assess CLIL effects on second language learning in Germany: More than selection and preparation effects? *International Journal of Bilingual Education and Bilingualism*, 1–14. https://doi.org/10.1080/13670050.2022.2164174.

Feddermann, M., Guill, K., List, M. K., Matthießen, R., Ömeroğulları, M., Stallasch, S. E., … Nagy, N. (2019). KESS – Skalierung der Leistungstests [KESS - Scaling of the achievement tests]. Kiel: Leibniz-Institut für die. *Pädagogik der Naturwissenschaften und Mathematik*.

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

Goris, J. A., Denessen, E., & Verhoeven, L. T. (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

Hallquist, M. N., & Wiley, J. F. (2018). Mplusautomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Jerusalem, M. (1984). *Selbstbezogene Kognitionen in schulischen Bezugsgruppen—eine Längsschnittstudie [Self-related cognitions in scholastic reference groups—a longitudinal study]*. Berlin, Germany: Free University, Department of Psychology.

Jopt, U. J. (1978). *Selbstkonzept und Ursachenerklärung in der Schule [Self-concept and attribution at school]*. Bochum: Kamp.

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

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish–little-pond effect. *Contemporary Educational Psychology, 30*(3), 263–285. https://doi.org/10.1016/j.cedpsych.2004.10.002

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education, 28*(2), 165–181. https://doi.org/10.1177/000494418402800207

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*(1), 129–149. https://doi.org/10.3102/00028312023001129

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Marsh, D. (2002). CLIL/EMILE - the European dimension: Actions, trends and foresight potential. Retrieved from https://www.semanticscholar.org/paper/CLIL%2FEMILE-the-European-dimension-%3A-actions%2C-trends-Marsh/1607bf2b186bbd270776e3d88168d4740fd25b09.

Marsh, H. W., & Hau, K.-T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Education Psychology, 96*(1), 56–67. https://doi.org/10.1037/0022-0663.96.1.56.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Abduljabbar, A. S., Abdelfattah, F., & Jansen, M. (2015). Dimensional comparison theory: Paradoxical relations between self-beliefs and achievements in multiple domains. *Learning and Instruction, 35*, 16–32. https://doi.org/10.1016/j.learninstruc.2014.08.005.

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*(3), 107–123. https://doi.org/10.1207/s15326985ep2003_1.

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review, 120*(3), 544–560. https://doi.org/10.1037/a0032459.

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research, 90*(3), 376–419. https://doi.org/10.3102/0034654320919354.

Muthen, L. K., & Muthen, B. O. (1998). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Pishgar, F., Greifer, N., Leyrat, C., & Stuart, E. (2021). MatchThem:: Matching and weighting after multiple imputation. *The R Journal, 13*(2), 228. https://doi.org/10.32614/RJ-2021-073

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(5), 724–744. https://doi.org/10.1080/10705511.2018.1561293

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Rumlich, D. (2017). CLIL theory and empirical reality–two sides of the same coin? A quantitative-longitudinal evaluation of general efl proficiency and affective-motivational dispositions in CLIL students at German secondary schools. *Journal of Immersion and Content-Based Language Education, 5*(1), 110–134. https://doi.org/10.1075/jicb.5.1.05rum

Rumlich, D. (2018). Englischnoten und globale englische Sprachkompetenz in bilingualen Zweigen [English grades and global English language proficiency in bilingual streams]. *Zeitschrift für Erziehungswissenschaft, 21*(1), 29–48. https://doi.org/10.1007/s11618-017-0801-z

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279–313. https://doi.org/10.1037/a0014268

Seikkula-Leino, J. (2007). CLIL learning: Achievement levels and affective factors. *Language and Education, 21*(4), 328–341. https://doi.org/10.2167/le635.0

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21. https://doi.org/10.1214/09-STS313

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90–118. https://doi.org/10.1080/00273171.2011.540475.

Trautwein, U., & Möller, J. (2016). Self-concept: Determinants and consequences of academic self-concept in school contexts. In A. A. Lipnevich, F. Preckel, & R. D. Roberts (Eds.), *Psychosocial skills and school systems in the 21st century: Theory, research, and practice* (pp. 187–214). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-28606-8_8.

What Works Clearinghouse. (2022). Procedures and standards handbook version 5.0. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf.

Zaunbauer, A. C. M., Gebauer, S. K., Retelsdorf, J., & Möller, J. (2013). Motivationale Veränderung von Grundschulkindern in Englisch, Deutsch und Mathematik im Immersions- und Regelunterricht [Motivational change in English, German, and mathematics of children in immersion programs and regular classrooms]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 45*(2), 91–102. https://doi.org/10.1026/0049-8637/a000083.