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

Dominik Rumlich
University of Duisburg-Essen (Advisors: Bernd Rueschoff and Detlev Leutner)

This article summarizes the essential theoretical and empirical findings of a large-scale doctoral dissertation study on content and language integrated learning (CLIL) streams at German secondary schools (Gymnasium) with up to three content subjects taught in English (Rumlich, 2016). A theoretical account rooted in teaching English as a foreign language (EFL), language acquisition and educational psychology provides the basis for the development of a comprehensive longitudinal model of general EFL proficiency, which incorporates cognitive, affective-motivational, and further individual variables.

In a second step, the model is used to estimate the effects of CLIL on general EFL proficiency, EFL self-concept and interest over a span of two school years (Year 6 to Year 8). The statistical evaluation of the quasi-experimental data from 1,000 learners finds large initial differences prior to CLIL due to selection, preparation, and class composition effects brought about by the implementation of CLIL within streams. After two years, the analyses found no CLIL-related benefits for general EFL proficiency or interest in EFL classes and solely a minor increase in EFL self-concept that might be attributable to CLIL. The results make a strong claim for comprehensive longitudinal model-based evaluations and the inclusion of selection, preparation, and class composition effects when conducting research on CLIL programmes in similar settings. The findings also suggest that not all language competences and affective-motivational dispositions might benefit from CLIL (the way it is currently taught in Germany) to the same extent.
Keywords: Content and language integrated learning, bilingual education in Germany, EFL proficiency, academic/EFL self-concept, language learning interests, large-scale quantitative research

1. Introduction

Following an “explosion of interest” in teaching content through the medium of a foreign language in the 1990s (Coyle, 2006, p. 2), content and language integrated learning (CLIL) is currently realised on all levels of education from preschool to tertiary institutions in the majority of European countries.1 It is considered to be an “effective means of improving language learning provision” (Council of the European Union, 2008, p. 22), “adding a European dimension to the curriculum” (Fruhauf, 1996, p. 8). In 2007, Lorenzo (2007, p. 35) argued that “CLIL is bilingual education at a time when teaching through one single language is seen as second rate education” – a claim which García renewed in 2009 with her statement that “bilingual education is the only way to educate children in the twenty-first century” (p. 5; her emphasis). These quotations are only a few testimonies that capture the hype around a concept that has been endorsed internationally by a considerable number of academics, administrators, politicians, and teachers in a wealth of theoretical, practical, and empirical publications. It may hence be justifiably regarded as one of the central topics in the realm of present-day foreign language education.

In theory, CLIL is, inter alia, thought to implement fundamental constructivist (e.g., Vygotsky, 1978) as well as neuroscientific claims of language learning (e.g., Ting, 2010), turning it into an advantageous setting for intense cognitive activity. It is envisaged to tap into learners’ natural predispositions to learn language while creating a favourable learning environment with contextualised real-world language input and ample opportunities for content-driven, meaningful social interaction. At the same time, CLIL is said to encourage a functional pluriliteracies approach to language learning (Meyer, Coyle, Halbach, Schuck, & Ting, 2015) and helps to take into account the interrelatedness of content, culture, cognition, and communication (4Cs framework; Coyle, 2008). All of this is seen as a fundamental asset to creating rich and authentic settings facilitating (sub-)conscious, incidental, and deep learning in conjunction with the development of learner autonomy, content-, language-, strategy-, and culture-related competence. By means of enhancing learners’ exposure to and use of the target language without extra lessons,

1. This is not to say that CLIL does not exist outside of Europe, but European developments (together with research on the immersion programmes in Canada) were the major driving force behind CLIL in Germany.
they obtain additional opportunities to practice in challenging CLIL classes what they learnt in their regular English classes to the mutual benefit of both (Marsh, 2000). In short, CLIL is generally linked to great (language) learning expectations and enjoys the reputation of being high-quality (language) education.

A closer, more critical look at the existing literature reveals, however, that many might feel very strongly about CLIL, but their convictions about its promises and benefits are not backed up by corresponding research findings on a grand scale. This overall situation is summarized by Pérez-Cañado (2012, p. 329) in a critical appraisal of the literature on CLIL:

CLIL has engendered widespread discussion on the continent and spawned an inordinate – almost infinite – amount of publications on the topic. A series of key figures have spurred the latter on […] and have engaged in extensive theorizing on CLIL, its principles and models, recommendations for its implementation, or reviews of the research conducted on it. However, solid empirical studies have been sparse. As Navés (2010) underscores, in the last two decades, whereas North America has been busy researching the features and effects of successful bilingual programs, Europe has merely been occupied in describing their benefits.

The reasons for the lack of reliable and valid empirical research on CLIL in a European context are manifold and, inter alia, related to the fuzziness and overly inclusive nature of the concept of CLIL itself, the relative recency of the establishment of CLIL as a part of mainstream education in Europe, and the diversity of CLIL programmes realised in different educational contexts.

Against the background of this overall situation, it was decided to address the current research void on two levels: By means of a contextualized theory-based and at the same time large-scale empirical study on learning in EFL CLIL streams in Germany, the overarching goal was to explore two general research questions (RQs):

1. What is the effect of CLIL on general language proficiency?
2. If CLIL has an effect, how could it be explained?

The general nature of RQ #1 mirrors the fact that little is known “for sure” about the effects of CLIL on language competence. Hence, general language proficiency as the broadest, most general construct from this domain appears to be an appropriate starting point for foundational empirical research. RQ #2 is meant to address popular explanations for the oft-reported linguistic benefits of CLIL: It is often assumed that CLIL exerts positive influence on the affective-motivational dispositions of the participating learners (e.g., their self-concept and their motivation/interest), which might be one of the driving forces behind improvements in language proficiency.
The paper at hand follows the structure of the original doctoral dissertation thesis: Starting with a general illustration of the notion of CLIL, the subsequent section will provide a succinct depiction of the conceptual and institutional background of CLIL in Germany. This is crucial as the educational context and the way CLIL is implemented in Germany constitutes the foundation for all of the ensuing theoretical and empirical considerations. On the basis of existing theoretical and empirical literature, a theory of learning in EFL CLIL will be devised with a particular focus on general EFL proficiency, EFL self-concept, interest in EFL classes, verbal cognitive abilities, and spare-time English as essential variables of a model trying to explain language learning in CLIL. This theory together with a summary of empirical CLIL research from Germany supplied the basis for the study with which the existing research gaps were addressed. The subsequent sections will present the analyses and interpretation of the empirical data on the basis of descriptive and inferential statistics, followed by a critical conclusion and suggestions for further research.

2. Conceptual and institutional background of CLIL in Germany

CLIL is a broad and inclusive umbrella term denoting a multitude of teaching approaches that can be flexibly implemented with varying stress on content and language (Dalton-Puffer, Llinares, Lorenzo, & Nikula, 2014; Eurydice, 2006):

[CLIL] is a generic term and refers to any educational situation in which an additional language and therefore not the most widely used language of the environment is used for the teaching and learning of subjects other than the language itself.

(Marsh & Langé, 2000, p. iii)

From the expression itself, it follows that CLIL describes a general teaching principle which stipulates that content and language are to be taught in an integrated way. As a result, Marsh (2002) enumerates at least thirty-three different approaches across Europe that can be subsumed under the roof of CLIL. While this has facilitated and maximized Europe-wide identification with the underlying teaching approach/principle, its ubiquitous use happens at the expense of accuracy and precision: Two people could talk about CLIL and refer to entirely different realisations within different educational contexts (e.g., low-intensity projects or high-intensity streams, monolingual L2 lessons with a strong focus on content, or bilingual lessons with occasional focus on content, etc.). Hence, the chief strength of the term simultaneously represents its major weakness. Given the fact that national education systems across the world are highly diverse, cross-national transfer of research results is highly problematic and usually invalid. In any case, a detailed
illustration of what is meant by CLIL in a particular context needs to be part of theoretical and empirical accounts on CLIL.

In Germany, CLIL is realised in the form of *Bilingualer Sachfachunterricht* (“bilingual content subject teaching”). The German technical term points to a strong focus on content, which is also brought about by the non-existence of separate CLIL curricula; if at all, there is a set of additional recommendations that complement the existing content subject curriculum. *Bilingualer Sachfachunterricht* as it is known today entered general mainstream education in 1969 with the establishment of a French-German strand at Hegau-Gymnasium in Baden-Wuerttemberg (KMK, 2006, p. 8). It remained a niche phenomenon until the beginning of the 1990s when the movement gained momentum on a noticeable scale in Europe in general and Germany in particular. Together with the increasing dominance of English in CLIL programmes, this overall development also mirrored the growing influence of European educational policy on education systems across Europe as a direct result of the Maastricht Treaty on European Union signed in 1992.

North-Rhine Westphalia is the most populous German state with the largest number of CLIL schools. In 2014, 26.3% of all public and private Gymnasien, Realschulen, Gesamtschulen provided some form of CLIL (continuous and episodic forms of CLIL); at Gymnasium, this figure is close to 40% with 21% offering continuous forms of CLIL with a duration of at least one school year. In this context, the original realization of CLIL in streams with up to three content subjects taught in a foreign language continues to be the most intense form of CLIL implemented on a larger scale in mainstream education at Gymnasium even today. Yet, at the same time, only approximately 10% of the total student population at Gymnasium attends CLIL streams, which is an important figure to relativise the potential impression that high-intensity CLIL programmes are the rule rather than the exception.3

---

2. *Gymnasium*: The most academic type of secondary school in Germany (grades 5–12 or 13), aimed at preparing pupils to take up studies at a university.

*Realschule*: A medium-academic type of secondary school (Grades 5–10), sometimes translated as secondary modern school, and typically aimed at preparing students for vocational training/an apprenticeship (to become a nurse, for instance). Upon passing their final examinations, students are allowed to continue at a Gymnasium if their marks are above a certain threshold.

*Gesamtschule*: Literally translates to “comprehensive school”; a type of school where all general secondary qualifications can be obtained.

3. Unfortunately, there are no reliable figures on all forms of CLIL, continuous and episodic. While a substantial number of schools offer some form of CLIL and often characterize it as an outstanding feature of their school profile, it remains unclear as to how many students end up participating in these CLIL instalments.
In the high-intensity version, students receive additional language lessons in Years 5 and 6 before CLIL starts in Year 7 at the age of 12 or 13 years. In North-Rhine Westphalia, this signifies a 50% increase in English lessons (six instead of four lessons per week). These regular EFL classes continue on the usual level of four lessons per week when CLIL begins in Year 7 with one subject; it continues in Years 8 and 9 usually with two and up to a maximum of three CLIL subjects.

Apart from the special preparation for the linguistic challenges of CLIL, a threefold selection process is characteristic of voluntary CLIL streams in Germany:

1. Parents decide for or against a CLIL school when registering their child in a particular type of secondary school (more or less academic types) at the age of 9 or 10 years.
2. Together with a teacher/the headmaster and their child, parents decide in favor or against the registration of their child in a CLIL strand.
3. Grade repetition/retention and drop-out take place as an optional or mandatory measure if students do not wish to continue CLIL or do not achieve the set curricular aims.

These (self-)selection and screening processes are bound to induce educational creaming (Gamoran, 2008)\(^4\), which is meant to single out those students who are most likely to master the challenges posed by CLIL streams. The additional preparatory lessons in Years 5 and 6 represent the second key measure to increase the likelihood of students’ educational success in CLIL.

These measures related to student selection and preparation also bear important implications for research on CLIL in this specific educational context in Germany: CLIL students are bound to represent a positively selected group of students, who also receive the equivalent amount of one additional school year’s worth of EFL teaching. It appears safe to assume that this leads to higher EFL proficiency and more favourable (affective-motivational) learning dispositions in CLIL students even before CLIL, in comparison to unselected and unprepared students from schools without CLIL. In the context of empirical evaluation studies, these initial differences need to be taken into account by appropriate longitudinal research designs or complex statistical techniques such as propensity score matching. At the same time, it is important to note that the students in non-CLIL strands at CLIL schools might represent a negatively selected group of students with below-average EFL proficiency (assuming that most of the above-average students end up in a CLIL strand, the remaining group of students would be below

\(^4\) Gamoran (2008, p. 160) defines it as “the practice of serving a select client base in order to improve the efficiency or efficacy of a treatment.” In the context of education, creaming is hence often used to describe some sort of positive selection.
average by default). This would render the non-CLILs an inappropriate group of controls, which would be a further important aspect to be considered in comparative studies.

3. Modelling learning in CLIL

This section represents an abridged version of two major foundational chapters of the original dissertation and can therefore only provide a general theoretical and empirical overview of the influence of CLIL on language learning, academic EFL self-concept, and interest in EFL classes. See Rumlich (2016) for details and further theoretical as well as empirical accounts on spare-time English, verbal cognitive abilities, biological sex, age and L1, which are excluded in the following.

3.1 Language learning

The Canadian immersion programmes originated from a general dissatisfaction with the results of language education at school (Wesche, 2002) – a situation which is also evident in many countries today. In the eyes of a large number of academics, practitioners, politicians, and administrators, CLIL is regarded as a promising catalyst to significantly improve the outcomes of language education.

From an empirical point of view, several quantitative studies have come to the conclusion that CLIL students show enhanced levels of general English language proficiency (Bredenbröker, 2000; Dallinger, 2015; Nold, Hartig, Hinz & Rossa, 2008; Zydatiß, 2007), listening and reading skills (Köller, Leucht, & Pant, 2012), and writing and speaking skills, in comparison to students outside of CLIL strands.\(^5\) These findings, however, do not prove that CLIL is the reason for these outcomes. In many cases, the observed differences are based on composite measures of selection, preparation, class composition, and potential CLIL effects that could not be separated due to cross-sectional study designs. Studies on a potential head start of the CLIL students found moderate to high effect sizes with respect to selection (listening comprehension: Cohen’s \(d = 0.50\)), and selection, preparation, and class composition (composite measure of general EFL proficiency and reading comprehension: \(d = 0.98\)) (Bos et al., 2009).

There are two large-scale longitudinal studies that have examined CLIL and non-CLIL students’ development of language proficiency over time while

---

\(^5\) To account for the fact that direct comparisons of research results from various educational contexts are not possible, the article at hand concentrates on research from Germany. See, e.g., Pérez-Cañado (2012) for an overview of CLIL research in Europe.
considering initial differences: Both the DESI study (short for “Deutsch Englisch Schülerleistungen International” [student achievement in German and English]) in Year 9 with a matched control group design (Nold et al., 2008) as well as Dallinger’s model-based evaluation (2015) found no CLIL-related increase in general EFL proficiency. However, CLIL seemed to have contributed to an enhancement of CLIL students’ listening comprehension skills over the one-year period of the studies.

3.2 EFL self-concept

In empirical research, subject-specific academic self-concept has frequently emerged as the most influential affective-motivational determinant of achievement across all subjects (Helmke & Weinert, 1997). From a differential point of view, there is general agreement that self-concept denotes a somewhat stable, dispositional, or habitual tendency of self-perception “formed through experiences with and interpretations of one’s environment. It is influenced especially by significant others, reinforcements, and attributions of one’s own behaviour” (Marsh, 2006, p. 8). “It includes feelings of self-confidence, self-worth, self-acceptance, competence, and ability” (Marsh & Scalas, 2010, p. 660).

Academics in the field of language learning have frequently underlined that the self plays a pivotal role (Mercer, 2011), yet, the area of language learning has rarely concerned itself with academic self-concept. As a result, there is a scarcity of theoretical considerations and empirical research in general and in the niche of CLIL in particular (cf. Rumlich, 2015). The original dissertation contains a first theoretical attempt to sketch the potential influence of CLIL on students’ EFL self-concept, which can be summarized as follows:

- Increases in language competence, mastery experiences, judicious praise and positive feedback, encouragement, appropriate internal attributions for success (and failure), choice, and involvement are bound to exert positive influence on students’ EFL self-concept.
- If learners feel overwhelmed, experience trouble following the lessons and comprehending the learning materials, are not able to solve the tasks set by the teacher or (feel that they) receive worse grades due to CLIL, such classes could have a negative impact on students’ EFL self-concept.

As self-concepts are relational and determined by external frames of reference provided, inter alia, by classmates and teachers, effects such as big-fish-little-pond (Marsh, 1987) or reflected glory (Cialdini et al., 1976) might also prove to exert influence on students’ EFL self-concept.
Due to this mixture of mechanisms with potentially positive and negative influences, it is hard to predict from theory what effect CLIL is likely to have on EFL self-concept. In spite of the fact that there is almost no empirical research, self-related constructs such as self-concept and self-confidence are generally assumed to be positively affected by CLIL. Indications of a more positive EFL self-concept in CLIL students can be found in Burmeister (1994), Dallinger (2015), and Zydatiß (2007), even though the study by Dallinger is the only one that looked at the construct in a systematic way. She found medium-sized differences of approximately half a standard deviation (Cohen’s \( d \approx 0.50 \)) when comparing CLIL to non-CLIL students at the beginning of Year 8.

### 3.3 EFL interest

Interest is a motivational construct whose key feature is its content specificity; dispositional, individual, or personal interest denotes “a relatively stable tendency to occupy oneself with an object of interest” (Krapp, 2002, p. 388). This object may be a concrete item, but also a more abstract construct such as a topic or subject-matter. Interest goes hand in hand with high subjective value and positive emotions such as enjoyment and pleasure; it facilitates commitment, cognitive activation, and feelings of competence so that the quality of learning is enhanced. Interested learners are ready to invest more time in the learning process as they have the intrinsic desire to repeatedly or continually concern themselves with their object of interest. This signifies that individual interests “play a major role in a learner’s preference to engage in a task or activity over time” (Subramaniam, 2009, p. 12).

As for EFL self-concept, there is a dearth of theoretical and empirical literature on the connection of language learning and interest in general, not to mention the influence of CLIL on interest in particular (cf. Rumlich, 2015). Positive effects of CLIL on interest and motivation have frequently been assumed in theoretical literature on CLIL (e.g., Wolff, 1997), yet, corresponding remarks are often based on a rather non-technical, everyday understanding of both constructs. From the point of view of a language learning theoretician, there are several reasons why CLIL could have a positive impact on students’ interest in EFL: Rich and meaningful authentic language input and use of the language; focus on the successful transmission of relevant messages; product and process orientation in task- and project-based settings; student-centred learning with high degrees of learner autonomy; and enhanced authenticity, relevance and complexity of materials, content, tasks, and interaction. Furthermore, CLIL possesses the potential to enhance students’ interest by satisfying their basic needs for autonomy, mastery, and social relatedness (see self-determination theory; Deci & Ryan, 1985), by setting demanding intellectual standards and assisting teachers in implementing principles.
of constructivist teaching, by nurturing/strengthening (shared) intrinsic values while avoiding a strong focus on extrinsic values reinforced, for instance, by written class tests and grades.

However, as CLIL represents a fusion of content and language teaching, the importance of students’ interest in the content subject (e.g., geography) and the interaction of subject- and language-related interests is yet unclear: Can higher levels of interest in the content subject compensate for a lack of interest in the language subject or vice versa (compensation and/or enhancement effects; cf. Bonnet, 2012a)? What is the development of students’ interest in CLIL subjects given that their interest in science subjects usually displays a stark decrease over the course of secondary school while the interest in EFL stays on a higher level (Daniels, 2008)? Does CLIL compensate for differences in interest between girls and boys as the former are, on average, reported to be more interested in languages whereas the latter are more interested in the sciences (Lasagabaster, 2008)?

A handful of empirical studies provide indications of enhanced levels of EFL interest in CLIL students (e.g., Bredenbröker, 2000; Fehling, 2008; Zydatiß, 2007), which included constructs distantly related to interest in EFL classes. To date, the only German study that incorporated the construct of interest in a systematic way is the one by Dallinger (2015). In her study at the beginning of Year 8 after one year of CLIL, she found small to medium-sized differences when comparing CLIL and regular students at schools without CLIL (Cohen’s $d = 0.27$) and CLIL and non-CLIL students’ from the same schools (Cohen’s $d = 0.40$). Unfortunately, it is not possible to determine whether this difference was brought about by CLIL or whether it was a difference that already existed a priori.

4. Research gap, study design, and instruments

As has become clear from the above illustrations, the area of CLIL in Germany (and also in other parts of the world) is in dire need of further research. An almost ubiquitously strong belief in CLIL, which frequently surfaces in academic statements on what is allegedly known about CLIL and how (well) it works, could endanger the credibility of an entire area of language learning (Bonnet, 2012b). Therefore, it is crucial to critically take stock of what is actually known about the effects of CLIL and thus uncover the existing voids that require further academic attention.

The overall research basis in Germany is thin and fragile due to the fact that CLIL has only been part of mainstream education on a noticeable scale for approximately 25 years. Furthermore, the available studies are often not suitable for making more generalizable statements about the effects of current CLIL practice
and, in comparative evaluations, often fail to take \textit{a priori} differences into account. Statistical analyses often remain basic, i.e. they do not incorporate more than two or three variables, not to mention complex models – this is not only a problem in the field of CLIL, but also in the area of language learning (in Germany) as a whole. Hence, the study at hand pursued three major goals:

1. Development of a comprehensive cross-sectional model of general EFL proficiency including its major determinants (EFL self-concept, EFL interest, spare-time English, verbal cognitive abilities, biological sex, age, L1) to examine \textit{a priori} differences among CLIL, non-CLIL, and regular students (those without a CLIL option) (end of Year 6).

2. Application of the longitudinal model to examine the development of students’ general EFL proficiency and the contribution attributable to CLIL after two years when taking \textit{a priori} differences and a variety of interrelated variables into account.

3. Application of the longitudinal model to examine the development of affective-motivational dispositions (EFL self-concept and EFL interest) and the potential contribution attributable to CLIL after two years when taking \textit{a priori} differences and a variety of interrelated variables into account.

For this purpose, it was decided to conduct a longitudinal quasi-experimental (i.e. with non-randomised allocation of participants to CLIL treatment and control groups) study at Gymnasien in North-Rhine Westphalia starting at the end of Year 6 after the selection and preparation of the future CLIL students had taken place and prior to the beginning of CLIL in Year 7. In total, seventeen schools were invited to participate, thirteen of which consented. They are spread across central-and south-western North-Rhine Westphalia, the westernmost federal state in Germany bordering the Netherlands and Belgium in the west, Lower Saxony in the north (east), Hesse in the south (east) and Rhineland-Palatinate in the south; the schools were concentrated around Essen and the northern Ruhr area. They are situated in different areas, socioeconomically, geographically, and as regards their degree of urbanisation, which is meant to capture some of the diversity of the general student population.

The sample consisted of 993 students in 43 classes, who had obtained parental consent to participate in the 90-minute study sessions. They took place during

6. The study incorporated further groups of students for exploratory purposes (students from Realschule and one all-English immersion school; \( N_{\text{total}} = 1,398 \)) and additional data were gathered on constructs such as writing skills, general learning and achievement motivation, EFL-related learning goals, etc. All of these data will be analysed in the context of future publications and was not included in the original doctoral thesis.
regular lessons at the end of the school year and were conducted by members of a team of trained research assistants or lecturers on the basis of a strict session manual. The participating students belonged to three sub-groups: Those from CLIL strands (503; CLIL) and those from non-CLIL strands at the same CLIL schools (473; non-CLIL). A third group was comprised of students from schools without CLIL (182; regular), who were supposed to represent a neutral control group unaffected by the selection effects that are bound to occur at CLIL schools.

While CLIL-related creaming effects cannot be ruled out completely, geographical distance to CLIL schools and the non-existence of streaming was meant to reduce the likelihood of creaming. Overall, 23 students with English as L1 were excluded from all analyses (and are not included in the above figures). The mean age of the participants at the end of Year 6 was 11.9 years with 50.5 % of them being female.

General EFL proficiency was measured with 50 items (Year 6) and 100 items (Year 8) from a collection of validated C-tests from the *Hamburger Schulleistungstest* (Hamburg scholastic achievement test; Behörde für Schule, Jugend und Berufsbildung. Amt für Schule, Hamburg, 1998; 2000), which have previously been used in other large scale studies (e.g., KESS7 (Kompetenzen und Einstellungen von Schülerinnen und Schülern [Students’ competences and attitudes]): Bos et al., 2009). A C-test usually consists of a text, in which all but the first and last sentence contains words from which the second half is missing (e.g.: “An American friend of ours hired a car in London although he was inexperienced in driving on the left-hand side of the road. Soon h____ found him____ going i____ the wr____ direction…”; taken from: https://szdb.uni-kassel.de/ctest/ctest_probe.php?l=en). It is based on the idea of reduced redundancy testing requiring the test takers to exploit information from many different levels of language in an integrated way: “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). C-tests have been found to correlate substantially with overall TOEFL (Test of English as a Foreign Language) scores (Grotjahn, 2011, p. 135) as well as moderately to highly with different measures of language subskills, which tap into the same general dimension (Eckes & Grotjahn, 2006). Hence, they represent an economical way of obtaining a general estimate of people’s overall language proficiency in the context of large-scale quantitative studies:

Especially in cases where aggregate estimates of overall language competency are sufficient (e.g., exercises of educational system monitoring) and where testing time is scarce, this approach [C-test] appears to be a suitable alternative to the more fine-grained testing programs aimed at drawing competency profiles which demonstrate specific strengths and weaknesses of the respondents. (Lehmann, 2003, p. 157).
A one-dimensional Rasch model estimated with *Acer Conquest 3.0* (Wu, Adams, & Wilson, 2007) was used for the calculation of students’ proficiency scores in the form of weighted likelihood estimates (WLE) in order to facilitate longitudinal comparisons and analyse the measurement properties of the individual items and the test as a whole. Cronbach’s α was consistently high (> .90), weighted mean square values generally fulfilled the PISA criteria (0.80 ≤ MNSQ ≤ 1.20) and item discrimination, person separation reliability (EAP/PV), as well as item separation reliability were satisfactory.

The measurement of EFL self-concept and interest took place on the basis of a collection of eight or seven items, respectively, that originate from other large-scale studies (see Rumlich, 2016 for details). Students responded on a four-point Likert scale with the options “I fully agree”, “I (do) rather (not) agree”, “I fully disagree”. Due to students’ young age, only four options were used in accordance with similar studies in the same age group; the high number of items ensured sufficient differentiation among students. A one-factorial structure for each of the scales emerged from both confirmatory and exploratory factor analyses. At all times, Cronbach’s α is equal to or larger than .87 and the data fulfils the conditions for longitudinal analyses with structural equation models (SEM) (strong factorial invariance). These were calculated with *Mplus* (Muthén & Muthén, 1998–2012) using a maximum likelihood estimator robust to violations of multivariate normality (MLR); the setting type = complex was meant to account for the clustered nature of the sample and avoid the inflation of Type I error rates. After systematic missingness could be ruled out, missing values were dealt with by full information maximum likelihood (FIML) estimations.

5. **Model development and *a priori* differences between CLIL, non-CLIL and regular students**

On the basis of theoretical considerations from the area of expectancy(-x)-value theory (e.g., Eccles [Parsons], 1983, 2005) and existing empirical studies on EFL learning (e.g., Zaunbauer, Retelsdorf, & Möller, 2009), the model of language proficiency development below was devised (Figure 1). EFL self-concept represents the operationalized construct that corresponds to the expectancy component and EFL interest represents the value component with the former influencing the latter and both influencing language proficiency. In a longitudinal setting such as the one in the study at hand, this pattern occurs twice for both points in time with the additional influence of prior measures on subsequent measures. CLIL and non-CLIL environments, biological sex and students’ L1 represent covariates that potentially exert influence on all cognitive and affective-motivational variables.
Verbal cognitive abilities were assumed to influence EFL proficiency and self-concept, but not interest in EFL classes as this relation does not seem justified from a theoretical point of view.

The fit indices of the corresponding structural equation model (SEM) with EFL self-concept (SC) and interest as latent and general EFL proficiency as manifest outcome variables provide sufficient reason to assume the appropriateness of the hypothesized relationships among the above variables ($R^2 = .64$; $X^2(777) = 1505.60$, $p < .001$; CFI = .93; TLI = .94; RMSEA = .04, .03 < 90% CI < .04; SRMR = .04). Hence, the model was confirmed and henceforth employed to examine the contribution of CLIL to the three major outcome variables while statistically monitoring and controlling for the influence of other, potentially confounded variables.

The calculated Rasch scores of students’ general EFL proficiency indicate the existence of marked differences between CLIL (0.69), regular (−0.02) and non-CLIL students (−0.49) with their averages in the anticipated order (see Figure 2). With respect to EFL self-concept and interest in EFL classes, there are also substantial differences between CLIL students and both non-CLIL groups, but differences between the latter two groups are less pronounced and not statistically significant (see Figure 3). Concerning EFL self-concept, it is striking, however, that
the assumed order of means is reversed and non-CLIL students show, on average, a slightly stronger EFL self-concept than regular students; this implies that the hypothesis of negatively selected non-CLIL students does not hold true in this case.

**Figure 2.** Mean values of general EFL proficiency (Rasch WLE score) in Year 6 before CLIL according to groups

**Figure 3.** Mean values of EFL self-concept (raw score; 4 = max.) and interest in EFL classes (raw score; 4 = max.) in Year 6 before CLIL according to groups

Simple hypothesis tests (ANOVA) without the inclusion of covariates (Table 1) in accordance with more complex model-based evaluations with structural equation models confirm the above picture: CLIL students are significantly different from students outside of CLIL strands, even prior to CLIL, regarding all of the three constructs measured. With respect to general EFL proficiency, the differences
amount to a magnitude between one to two school years (average gains per school year at that stage between 0.55 and 0.80 of a standard deviation). The facilitative influence of selection, preparation, and class composition on students’ affective-motivational dispositions manifests itself in effect sizes that range between one third \((g = 0.34)\) and slightly more than half a standard deviation \((\Delta = 0.55)\) and can also be considered substantial. The figures also confirm the assumed influence of negative selection on general EFL proficiency in the order of three-quarters to one school year, while there is no conclusive evidence for such an effect regarding EFL self-concept or interest in EFL classes. Furthermore, there is a trend towards higher EFL self-concept in non-CLIL students when compared to regular students.

Table 1. Effect sizes of mean differences (without covariates) in Year 6 before CLIL \((N = 809)\); ✓ indicates \(p < .05\); “trend” indicates \(.10 < p \leq .05\); ---- indicates \(p \geq .10\); effect sizes (conceptually similar to Cohen’s \(d\): \(\Delta = \text{Glass’ } \Delta; g = \text{Hedges’ } g\).

<table>
<thead>
<tr>
<th>Comparison/Variable</th>
<th>CLIL vs. regular</th>
<th>CLIL vs. non-CLIL</th>
<th>Regular vs. non-CLIL</th>
<th>Order of means</th>
</tr>
</thead>
<tbody>
<tr>
<td>General EFL proficiency</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>non-CLIL &lt; regular &lt; CLIL</td>
</tr>
<tr>
<td>EFL self-concept</td>
<td>✓</td>
<td>✓</td>
<td>trend</td>
<td>regular &lt; non-CLIL &lt; CLIL</td>
</tr>
<tr>
<td>Interest in EFL classes</td>
<td>✓</td>
<td>✓</td>
<td>----</td>
<td>non-CLIL &lt; regular &lt; CLIL</td>
</tr>
</tbody>
</table>

Furthermore, CLIL students possess stronger verbal cognitive abilities (effect sizes in the order of one third to half a standard deviation), report higher amounts of spare-time English exposure and use \((0.24 \leq \text{Cohen’s } d \leq .54)\), and are significantly younger (effect sizes in the order of one quarter to one third of a standard deviation), which might have to do with grade repetition. In addition, females choose CLIL significantly more often than regular or non-CLIL strands (odds ratio of 1.54 and 2.46, respectively). With respect to students’ L1s, there are no significant overall differences.

6. The contribution of CLIL to general EFL proficiency, EFL self-concept and interest

Figure 4 displays the longitudinal development of mean general EFL proficiency and group mean differences from Year 6 to Year 8 according to groups. First of all, it becomes evident that all groups advance significantly over time as indicated by
The effect size of mean gains is in the range of 1.5 standard deviations over the two-year span of the study, which equals 0.75 SD per school year and hence corresponds with the advancement observed in other studies in this age group in Germany (e.g., Lehmann, Gänfsus, & Husfeldt, 2011).

With regards to general EFL proficiency, CLIL and regular students develop rather similarly, which is a first indication that CLIL might not play a significant role. At the same time, non-CLIL learners show slightly weaker advancement from Year 7 to Year 8, increasing the differences between non-CLIL students and the other two groups; in complex structural equation models, this is indicated by a negative contribution (β = −.12, p = .002) of non-CLIL environments to students’ general EFL proficiency in Year 8 (Table 2).

With the help of the baseline data from the end of Year 6, it is possible to take a priori differences into account and estimate the potential contribution of CLIL to general EFL proficiency and the other two constructs under scrutiny. Table 2 displays the standardized coefficients obtained from structural equation model (SEM) calculations, which also serve as effect size estimates (they can be considered regression weights). It becomes evident that there is no detectable influence of CLIL on general EFL proficiency (β = .07, p = .69). General EFL proficiency is primarily determined by prior proficiency (β = .40), EFL self-concept (β = .21), verbal cognitive abilities (β = .19), and sex with advantages of females over males (β = −.14), each of which make their individual contribution even when the other variables are taken into account. Also, there is evidence of a negative influence of non-CLIL environments, i.e. students in these strands progress more slowly than regular students (as suggested by Figure 4 above). Contrary to model expectations, interest
does not make a significant contribution to general EFL proficiency, but there is a clear trend that it becomes increasingly more important over time (Year 6: $\beta = -.01$, $p = .91$; Year 8: $\beta = .09$, $p = .09$), approaching statistical significance in Year 8.

Table 2. $\beta$ coefficients for predictors of key constructs in Year 8 considering prior measures/baseline data from Year 6 (bold print indicates significant influences, i.e. $p < .05$).

<table>
<thead>
<tr>
<th></th>
<th>Proficiency Yr 8 $\beta$ ($p$ value)</th>
<th>EFL SC Yr 8 $\beta$ ($p$ value)</th>
<th>EFL int. Yr 8 $\beta$ ($p$ value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior prof. (Yr 6)</td>
<td>.40 ($&lt; .001$)</td>
<td>.59 ($&lt; .001$)</td>
<td>.19 ($&lt; .001$)</td>
</tr>
<tr>
<td>CLIL</td>
<td>.07 (.69)</td>
<td>.10 (.12)</td>
<td>-.02 (39)</td>
</tr>
<tr>
<td>Non-CLIL</td>
<td>-.12 (.002)</td>
<td>-.01 (.91)</td>
<td>-.07 (.23)</td>
</tr>
<tr>
<td>EFL SC</td>
<td>.21 ($&lt; .001$)</td>
<td>EFL SC</td>
<td>.60 ($&lt; .001$)</td>
</tr>
<tr>
<td>Interest</td>
<td>.09 (.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.19 ($&lt; .001$)</td>
<td>.14 ($&lt; .001$)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spare-time Engl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.05 (.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sex (0 = female)</td>
<td>Sex (0 = female)</td>
<td>Sex (0 = female)</td>
</tr>
<tr>
<td></td>
<td>-.14 ($&lt; .001$)</td>
<td>-.02 (.52)</td>
<td>-.11 (.003)</td>
</tr>
<tr>
<td></td>
<td>Ll: no German</td>
<td>Ll: no German</td>
<td>Ll: no German</td>
</tr>
<tr>
<td></td>
<td>.01 (.56)</td>
<td>.06 (.15)</td>
<td>.02 (.67)</td>
</tr>
<tr>
<td></td>
<td>Ll: German+</td>
<td>Ll: German+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.02 (.33)</td>
<td>.09 (.003)</td>
<td>.02 (.49)</td>
</tr>
</tbody>
</table>

Looking at the longitudinal development of students’ EFL self-concept scores (Table 3), one can see slight enhancement in CLIL students and a slight decrease in non-CLIL students, while regular students remain on the same level. Corresponding dependent-samples $t$-tests indicate trends in the former two cases that approach statistical significance (CLIL students: $p = .09$; non-CLIL students: $p = .06$). SEM analyses (Table 2, middle columns) confirms the mere tendency of CLIL inducing slightly higher EFL self-concept over time ($\beta = .10$, $p = .12$), while they do clearly not confirm the supposed negative influence of non-CLIL environments ($\beta = -.01$, $p = .91$).

Table 3. Dependent-samples $t$ tests for longitudinal changes in EFL self-concept scores from Years 6 to 8 (adapted from Rumlich, 2016, p. 395).

<table>
<thead>
<tr>
<th></th>
<th>Year 6</th>
<th>Year 8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>$M$</td>
</tr>
<tr>
<td>CLIL</td>
<td>321</td>
<td>3.32</td>
</tr>
<tr>
<td>Regular</td>
<td>134</td>
<td>3.06</td>
</tr>
<tr>
<td>Non-CLIL</td>
<td>221</td>
<td>3.12</td>
</tr>
</tbody>
</table>

SEM calculations suggest that the major determinant of EFL self-concept in Year 8 is prior EFL self-concept ($\beta = .59$), with cognitive abilities ($\beta = .14$) and multilingualism (if German is one of their L1s) exerting minor influence ($\beta = .09$; see Table 2).
Contrary to students’ interest in other subjects, their interest in EFL classes does not diminish over time (see Table 4). Longitudinal analyses identify a negligible increase in non-CLIL students and a small increase of a similar magnitude in CLIL and regular students ($0.11 \leq \text{Cohen’s } d \leq 0.13$); due to their obvious parallel development, the small rise in EFL interest in CLIL students cannot be attributed to CLIL, but seems to appear independently of CLIL ($\beta = -0.02, p = .79$; see Table 2, two columns on the right).

Table 4. Dependent-samples $t$ tests for longitudinal changes in EFL interest scores from Years 6 to 8 (adapted from Rumlich, 2016, p. 404).

<table>
<thead>
<tr>
<th>Group</th>
<th>$N$</th>
<th>$M$</th>
<th>$SE$</th>
<th>$M$</th>
<th>$SE$</th>
<th>$\Delta M$</th>
<th>Results of dependent-samples $t$ test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIL</td>
<td>321</td>
<td>2.98</td>
<td>0.03</td>
<td>3.06</td>
<td>0.03</td>
<td>0.08</td>
<td>$t(320) = 2.30, p = .02, \text{Cohen’s } d = 0.13$</td>
</tr>
<tr>
<td>Regular</td>
<td>134</td>
<td>2.71</td>
<td>0.04</td>
<td>2.79</td>
<td>0.04</td>
<td>0.08</td>
<td>$t(133) = 1.33, p = .19, \text{Cohen’s } d = 0.11$</td>
</tr>
<tr>
<td>Non-CLIL</td>
<td>220</td>
<td>2.63</td>
<td>0.05</td>
<td>2.66</td>
<td>0.05</td>
<td>0.03</td>
<td>$t(219) = 0.61, p = .54, \text{Cohen’s } d = 0.04$</td>
</tr>
</tbody>
</table>

With respect to interest in EFL classes, it becomes evident that EFL self-concept represents its major determinant ($\beta = .60$) next to prior interest ($\beta = .19$) and biological sex with advantages of females over males ($\beta = -.11$; see Table 2). Neither CLIL nor non-CLIL environments appear to play a role in interest development.

7. Discussion and conclusion

First of all, the results underline the importance of more complex and more realistic (model-based) evaluations to explore the individual influence of variables on a construct rather than many discrete analyses with less than a handful of variables, as these procedures are prone to overestimating the relevance of individual variables and inflate Type I error rates. Looking at the data in a cross-sectional manner, there is no doubt about the overwhelmingly favourable language learning characteristics that CLIL students display in comparison to both regular and non-CLIL students. In line with results from previous longitudinal research on general EFL proficiency and in contrast to the theory of language learning in CLIL outlined above, the observable differences with regard to general EFL proficiency cannot be attributed to CLIL, however. Rather, they are a direct consequence of CLIL-related selection, preparation, and class composition intended to help students master the challenges of CLIL. While the head start of the CLIL students with regards to general EFL proficiency appears impressive at first sight, a second look reveals that
their advantage of approximately (slightly more than) one school year in comparison to regular EFL students is in the range of what could be expected given that they have had an additional year’s worth of language teaching during Years 5 and 6. Considering their selection-induced advantage, such a difference could even be regarded as smaller than expected. In comparison to negatively selected non-CLIL students, CLIL students’ advantage appears disproportionately larger, which underlines that regular students should be used as controls in quasi-experimental settings (even though their recruitment might be less convenient and their demographic setup might be different than that of the non-CLILs from the same school).

A second major insight constitutes the fact that the high-ability environment of CLIL streams is obviously not at all detrimental to students’ EFL self-concept, but allows them to develop slightly higher levels under the influence of CLIL than their peers outside of CLIL streams. In the German/North-Rhine Westphalian CLIL context, the detrimental influence of the big-fish-little-pond effect (Marsh, 1987) and external comparisons does not seem to be stronger than positive effects potentially induced by the prestige of CLIL, the obvious selection process and teaching – as well as competence-related influences.

The absence of gains in EFL interest due to CLIL might have to do with the fact that school-related interests are usually highest at primary school and decrease over time. Hence, significant increases in interest would generally be counterintuitive and the absence of a loss of interest might be a significant finding in itself (at the same time, it has to be conceded that the regular students did also not display decreasing interest and EFL classes appear to have a special status among school subjects; cf. Daniels, 2008).

The existence of higher EFL self-concept in non-CLIL students in comparison to regular students at the end of Year 6 is counterintuitive to the otherwise confirmed assumption of negative selection. Initially higher self-concept in non-CLIL students, which decreases over time and is then on a similar level to that of regular students, might be induced by reflected glory effects (Cialdini et al., 1976) of the CLIL school. The fact that they attend a CLIL school might make them slightly more self-confident about their EFL skills and foster thoughts along the lines of “I am at a CLIL school, where EFL plays a pivotal role; obviously, I must be quite a good EFL student.” Adverse experiences might be detrimental to their “above-average” EFL self-concept and, over time, it approaches the level observable in regular students, who remain on the same level from the end of Year 6 to the end of Year 8.

It goes without saying that the current study also has shortcomings and blind spots: The unequal size of the sub-samples reduced statistical power. Treatment integrity could not be guaranteed, i.e. CLIL teaching might have happened very differently across schools and classes, and it remains unclear if the absence of effects
might be related to less than optimal teaching. While it might appear reasonable to conduct research on general language competence first before examining more specific language skills, it could also be argued that the study did not tap into the strengths of the CLIL students, such as subject-specific language competences, receptive skills and/or oral productive skills (an expectation due to the fact that CLIL teaching is dominated by reading, listening, and oral interaction). These are just three of a multitude of limitations that could be brought forward (see Rumlich, 2016 for details) and need to be addressed in further studies. Yet, together with results from existing research, the insights provided by the study at hand contribute one important piece to the puzzle of the effects of CLIL – with many more missing.

All in all, it has become clear that there is a dire need for appropriately complex longitudinal research on all competences and their determinants to be fostered by CLIL (e.g., affective-motivational or volitional dispositions). At the same time, qualitative research is required to look into current CLIL practice and, together with quantitative as well as action research, work towards examples of evidence-based, best practice CLIL also to inform and propel theories of learning in CLIL.

References


[English in preparatory CLIL classes: Pilot study on the efficiency of additional language classes in Year 5 of German-English bilingual streams in Schleswig-Holstein]. Kiel, Germany: l&f.


Author’s address

Dominik Rumlich
Teaching English as a foreign language (TEFL)
Westfaelische Wilhelms-Universitaet Muenster
Department of English
Johannisstr. 12–20
48143 Muenster
Germany

dominik.rumlich@uni-muenster.de