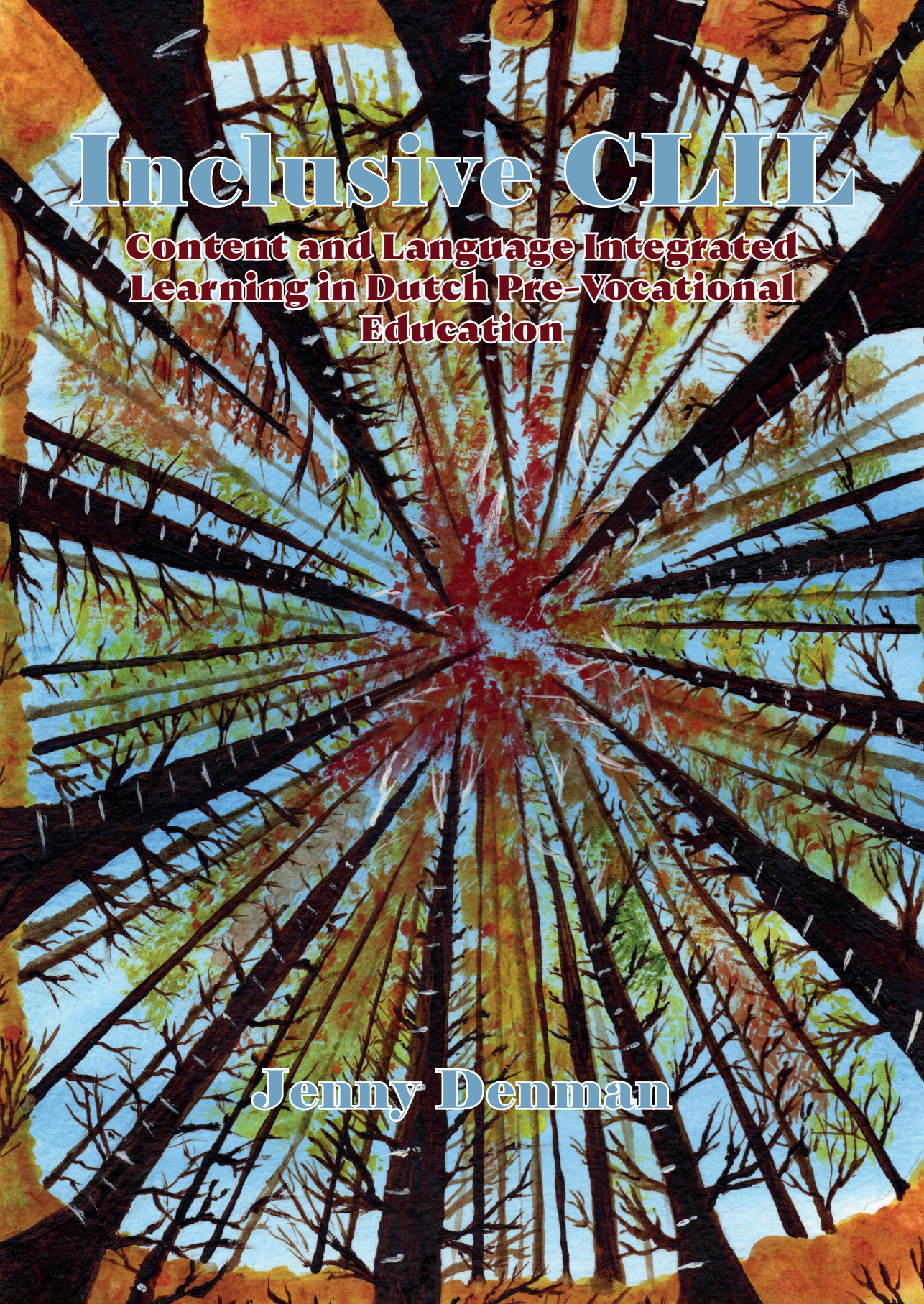

Inclusive CLIL
Content and Language Integrated Learning in Dutch Pre-Vocational Education
Jenny Denman

Inclusive CLIL

Content and Language Integrated Learning
in Dutch Pre-Vocational Education

Jenny Denman



Universiteit Utrecht



ROTTERDAM UNIVERSITY
OF APPLIED SCIENCES

Layout and printing of this dissertation were financed by Rotterdam University of Applied Sciences (Hogeschool Rotterdam).

Title: Inclusive CLIL: Content and Language Integrated Learning in Dutch Pre-Vocational Education

Titel: Inclusief CLIL: Content and Language Integrated Learning in het Nederlandse voorbereidend middelbaar beroepsonderwijs (vmbo)

Layout and Printing:	ProefschriftMaken www.proefschriftmaken.nl
Printed on	Recycled White Zero.
Cover image	© Sander Kruit
ISBN/EAN:	978-94-6534-472-0
DOI:	http:/ /doi.org/10.33540/3562

© Jennifer Love Denman, 2026

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the author or the copyright-owning journals for previous published chapters.

Inclusive CLIL

Content and Language Integrated Learning
in Dutch Pre-Vocational Education

Inclusief CLIL

Content and Language Integrated Learning
in het Nederlandse voorbereidend middelbaar beroepsonderwijs (vmbo)

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. ir. W. Hazeleger,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

woensdag 8 juli 2026 des middags te 4.15 uur

door

Jennifer Love Denman

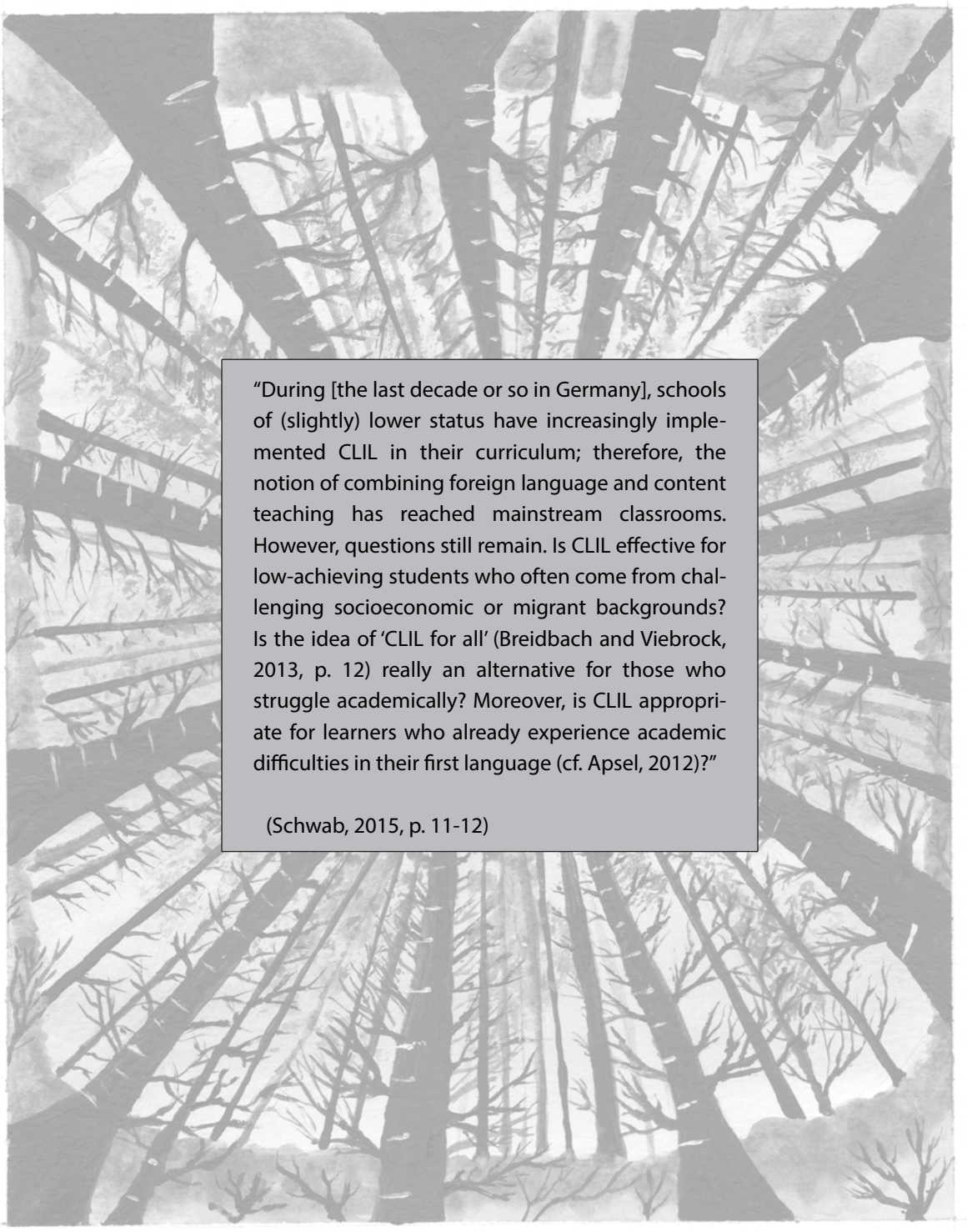
geboren op 23 februari 1957
te West Point, Verenigde Staten

Promotor:
Prof. dr. H. C. J. de Graaff

Copromotor:
Dr. E. J. van Schooten

Beoordelingscommissie:
Dr. J. Evers-Vermeul
Prof. dr. C. J. Jenks (voorzitter)
Dr. T. L. Mearns
Prof. dr. D. Rumlich
Prof. dr. M. H. Verspoor

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van Rotterdam University of Applied Sciences (Hogeschool Rotterdam).



"During [the last decade or so in Germany], schools of (slightly) lower status have increasingly implemented CLIL in their curriculum; therefore, the notion of combining foreign language and content teaching has reached mainstream classrooms. However, questions still remain. Is CLIL effective for low-achieving students who often come from challenging socioeconomic or migrant backgrounds? Is the idea of 'CLIL for all' (Breidbach and Viebrock, 2013, p. 12) really an alternative for those who struggle academically? Moreover, is CLIL appropriate for learners who already experience academic difficulties in their first language (cf. Apsel, 2012)?"

(Schwab, 2015, p. 11-12)

Table of Contents

Chapter 1: introduction 9
1. Overview 11
1.1 The Dutch educational context and CLIL development 12
1.2 Historical and conceptual background of CLIL 13
1.3 Theoretical foundations of CLIL 14
1.4 Research on CLIL outcomes and the question of selectivity 15
1.5. The research gap 18
1.6. The Dutch *vmbo* CLIL context and the present dissertation 21

Chapter 2: Attitudinal factors and the intention to learn English in pre-vocational secondary bilingual and mainstream education 29
2. Overview 31
2.1 Theoretical and conceptual framework 32
2.2 Methodology: research design and instrumentation 35
2.3 Data analysis and results 37
2.4 Discussion and conclusions 45

Chapter 3: CLIL in pre-vocational secondary education: gains in target language vocabulary and grammar 51
3.1 Introduction 53
3.2 Literature review: selection effects and language development in CLIL 53
3.3 Methodology and research design 57
3.4 Results: vocabulary and grammar development 66
3.5 Discussion, limitations and recommendations 71
3.6. Conclusion 74

Chapter 4: Inclusive CLIL: Pre-vocational pupils' target language writing and speaking proficiency, fluency, and Willingness to Communicate **77**
4.1 Introduction 79
4.2 CLIL selectivity and learners with lower scholastic attainment 82
4.3 The present study: research context and questions 84
4.4 Method and instruments 85
4.5 Analyses 90
4.6 Results: writing, speaking, and Willingness to Communicate 94
4.7 Discussion and conclusion 98

Chapter 5: "They underestimate us": Pupil and teacher perceptions of CLIL in pre-vocational secondary education **105**
5.1 Introduction and rationale 107
5.2 Theoretical and empirical background on CLIL perceptions 108
5.3 The Dutch pre-vocational CLIL context 116
5.4 Research questions and methodology 116
5.5 Results: teacher and pupil perspectives on CLIL 122
5.6 Discussion and conclusion 166
5.7 Limitations, recommendations for future research, and conclusion 183

Chapter 6: CLIL at vmbo: Conditions, Constraints, and Conclusions **187**
6.1. Background 189
6.2 Chapter core findings 194
6.3 Overarching analytical themes 198
6.4 Cross-context comparison 203
6.5 Implications for practice and policy 205
6.6 Current situation and future outlook 210
6.7 Limitations and future research 213

References 217

Appendix 2 238
Appendix 3 244
Appendix 4 264
Appendix 5 320
Appendix 6 326

Nederlandse Samenvatting 329

Acknowledgements 346

Curriculum Vitae 348



INTRODUCTION

1. Overview

Content and Language Integrated Learning, or CLIL, is the name given in the 1990s to the originally European approach to bilingual education in which some subjects in the school curriculum are taught mainly or partly through the medium of a foreign language. As the name indicates, the content of the specific school subject is integrated with the teaching of the foreign language (also called the L2[1], or second language, or the target language) so that attention is paid to both, with a fundamental goal being increased proficiency in the target language. It differs from some other well-known forms of language-intensive bilingual education, such as Canadian immersion, English Medium of Instruction, and Content-Based Instruction, in that learning goals are incorporated for both the subject content and the foreign language (Brown & Bradford, 2017). There are other typical characteristics of CLIL (Nikula, 2017a; van Kampen, 2019): generally, less than 50% of the curriculum is taught using the foreign language; the subject and language teachers are mostly non-native speakers of the foreign (target) language; foreign language lessons support subject lessons, rather than being replaced by them; and the target language is generally not a national or regional language, but an international lingua franca.

In the early years, European CLIL was seen as a way to provide highly motivated academic achievers with an extra challenge, and thus was highly selective. In countries and contexts where the educational system was streamed and pupils channeled into different academic levels, CLIL was offered only to pupils at the highest academic levels. Where pupils were not streamed into different levels, there were usually various selection processes in place, such as academic marks, language tests, or motivation letters. To some degree the selection of pupils with favorable learning prerequisites probably contributed to the generally positive results for CLIL pupils regarding the target language gains, as compared to their peers in the regular, non-CLIL streams[2]. Research and anecdotal evidence pointed in some contexts to other gains as well: more positive attitudes towards learning, intercultural sensitivity and international awareness, and sometimes even gains in the subject matter or mother tongue proficiency. Despite such profound potential or realized gains, only relatively recently have CLIL programs in various countries started to become more inclusive and open to learners with more divergent academic aptitude.

1 The abbreviation 'L2', or 'second language' is used hereafter to indicate English, the CLIL target language and main foreign language learned at school by the pre-vocational pupils in this study. We use it here regardless of whether English is the second, third, or even fourth language of the pupils in this sample, in keeping with the definition of an L2 as a language learned later than early childhood (Mitchell et al., 2019). Accordingly, we used L1 to indicate Dutch (the majority language and the language of school), still recognizing that approximately one-third of the pupils do not have Dutch as a first or home language.

2 In this dissertation, *track* refers to an educational pathway at a particular academic level (e.g., *vmbo*, *havo*, *vwo*), and the vocational levels within *vmbo*, whereas *stream* refers to a curricular approach that may occur across levels (e.g., CLIL or non-CLIL).

1.1 The Dutch educational context and CLIL development

1.1A Structure of Dutch secondary education

In the Netherlands, divergent aptitude is signaled at the end of primary school (age 11-12). Pupils are assessed by their primary school teachers and sit a standardized test which aims to measure cognitive development and mastery of basic academic skills. Based on the assessment and test result, pupils are streamed into one of several types of secondary schools, two pre-academic tracks (*vwo,* or pre-university, and *havo,* or pre-professional) and a less academic pre-vocational level (*voorbereidend middelbaar beroepsonderwijs*, or *vmbo*). Within this *vmbo* track there are four subtracks, from more theoretical to the most practical, serving approximately 50% of the total secondary pupil population. (For a more detailed overview, see Mearns et al., 2023.)

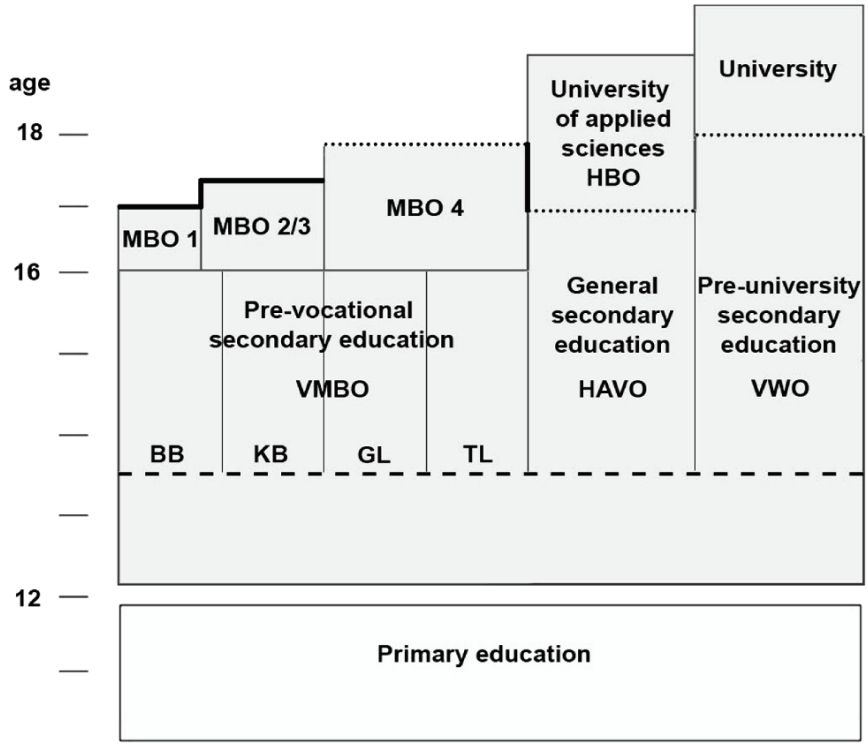


Figure 1.1: Overview of the Dutch educational system including the four levels of pre-vocational secondary education (*vmbo*) and typical progression routes to senior secondary vocational education (*mbo*), senior general secondary education (*havo*) and pre-university education (*vwo*). Note. From Testing in mathematics education in the Netherlands, by Scheltens et al., 2020, in M. van den Heuvel-Panhuizen (Ed.), *National Reflections on the Netherlands Didactics of Mathematics* (pp. 303-330). Open access chapter published by Springer under a CC license. *https://doi.org/10.1007/978-3-030-33824-4_16*

1.1B Introduction of CLIL at pre-vocational level

In Dutch education it took nearly twenty years for the CLIL approach to trickle down to the lower academic *havo* level, and even longer to *vmbo* level. This gradual expansion of CLIL to pre-vocational level meant that the Dutch national standards for bilingual education (Europees Platform, n.d.)—to which all CLIL school programs must subscribe—had to be adapted to the pre-vocational context in several ways, including the target levels for foreign language skills. Not only do the pre-vocational CLIL programs cater to a less academically privileged pupil population, but pre-vocational secondary education is shorter: one or two years less than *havo* and vwo, respectively.

Thus, one of the fundamental principles in the first Standard for pre-vocational CLIL programs was the deliberately modest target level for foreign language skills, so as to make the CLIL stream as accessible and inclusive as possible, while still aiming for a higher L2 level than the regular non-CLIL stream without compromising the L1 level or content learning.

1.1C Adaptation of the CLIL Standard and accessibility principles

The number of CLIL programs at Dutch pre-vocational level has more than doubled since the early 2010s, but the current study is the first to examine the growth of target language proficiency of Dutch pre-vocational CLIL pupils as compared to their non-CLIL peers, in addition to exploring their attitudes towards learning the target language and their perceptions of the CLIL approach as implemented in Dutch education. It is also, to the best of our knowledge, the first study worldwide to measure the target language proficiency gains, attitudes, and perceptions of average- to lower-attaining pupils learning schools subjects in a foreign language through the CLIL approach.

1.2 Historical and conceptual background of CLIL

Education through other languages than pupils' L1 has existed for thousands of years in different forms and in greatly different contexts, from the Roman Empire to British colonies to the 19th-century assimilation of immigrants in the United States. The second half of the 20th century saw the emergence of bilingual immersion programs in Canada, in which English-speaking pupils were 'immersed' in the French language for over 50% of the school curriculum, rather than only being exposed to the foreign language as a separate school subject. The foreign language gains reported in research on Canadian immersion (Genesee, 1986, 1987; Harley, 1991) stimulated interest in Europe, and in various countries grass-roots movements began to implement some immersion principles, expanding pupils' exposure to a target foreign language through its use in other school subjects. The 1995 European Commission proposed that every European Union citizen should achieve functional linguistic proficiency in two European Community foreign

languages—in addition to the mother tongue—in order to achieve the lofty aims of combatting unemployment, reducing social and gender inequality, contributing to a European identity, making the benefits of the border-free single market more accessible, and generally expanding citizens' cultural horizons (European Commission, 1995). It advised innovation in language-teaching methods and suggested that secondary school pupils could study content subjects in a foreign language. The encouraging research results from Canadian immersion, coupled with local and national interest in foreign language instruction innovation and the support of the European Commission, sparked the spread of CLIL in various European contexts.

1.2A Conceptualizing CLIL: variety and scope

In the mid-1990s several grass-roots approaches to bilingual education in Europe were brought together and the acronym CLIL—Content and Language Integrated Learning—was coined (Marsh et al., 2001) to refer to bilingual education in emerging European contexts, where foreign language instruction still took place in foreign language classes but was also used as a target language in subject classes. The CLIL approach was "a dual-focused form of instruction where attention is given both to the language and the content" (Coyle et al., 2010, p. 3), with varying degrees of focus, but in most contexts less intensive than in Canadian immersion. Given the plethora of contextual variables in European countries and in education in general, it is no surprise that no one CLIL size fits all. There are many different forms of CLIL (Coyle, 2007; Ruiz de Zarobe, 2023), depending on variables such as level, age, intensity, and duration, as well as the relative emphasis on language or content. While bilingual education is still the general organizing principle (Mearns & de Graaff, 2018a), CLIL can be seen as an umbrella term covering the various approaches to the goal of the integration of subject content and language learning at school.

1.3 Theoretical foundations of CLIL

1.3A The 4Cs Framework

The underlying pedagogical rationale of CLIL is perhaps most clearly articulated in Coyle's (2007) and Coyle et al.'s (2010) 4Cs Framework, which conceptualizes CLIL as the dynamic integration of *Content, Communication, Cognition,* and *Culture* within a given educational *Context*. These interrelated dimensions represent the dual focus of learning subject matter and developing target language competence, while also engaging learners cognitively and fostering intercultural awareness. *Content* encompasses both disciplinary knowledge and the subject-specific thinking skills learners must develop; *Communication* emphasizes interaction and negotiation of meaning in the target language, linking language learning to authentic use; *Cognition* concerns the development of thinking processes at appropriate levels of complexity, from remembering and understanding to analyzing and

evaluating; and *Culture* involves recognizing and valuing diverse perspectives, cultivating intercultural understanding, and situating learning in broader social contexts. As Coyle (2006b, 2007) argues, these dimensions cannot be addressed in isolation: effective CLIL pedagogy emerges from their integration and from the contextual conditions that enable such interaction. More recently, the centrality of teacher collaboration in CLIL implementation has been foregrounded, with Dale and Mearns (2023) proposing *collaboration* as a fifth 'C', complementing Coyle's original framework by explicitly recognizing the role of coordinated planning and shared pedagogical responsibility.

1.3B The Language Triptych

Complementing the 4Cs model, Coyle's Language Triptych (Coyle et al., 2010) offers a linguistic lens for planning and analyzing CLIL instruction. It distinguishes between *language of learning* (the disciplinary language needed to access and process new concepts), *language for learning* (the functional language required to participate in classroom interaction and carry out tasks), and *language through learning* (the emergent language that develops as learners articulate new ideas and communicate about content). The Triptych underscores that language development in CLIL is not confined to pre-taught structures but arises continuously through engagement with subject matter and through communicative, task-based pedagogy.

1.3C Application in Dutch TTO (bilingual education with CLIL) classrooms

In the Dutch *tweetalig onderwijs* (TTO) context, the 4Cs principles are operationalized through lessons conducted predominantly in the target language, attention to linguistic features of subject content, and communicative activities both in and beyond the classroom. Cross-curricular projects, reflective portfolios, and European and International Orientation (EIO) initiatives exemplify how theoretical principles are translated into classroom practice (see Dale & Tanner, 2012). Together, the 4Cs Framework and the Language Triptych provide a coherent theoretical and pedagogical foundation for the design, analysis, and evaluation of CLIL instruction, while the specific balance between language and content varies across national and institutional contexts—a diversity reflected in the range of approaches that have emerged in recent decades.

1.4 Research on CLIL outcomes and the question of selectivity

1.4A Evidence of language gains and conflicting findings

Whatever the approach, CLIL provision in schools in Europe and further afield has expanded dramatically in the past decades. Similarly, the body of CLIL research, first dependent on research on immersion programs in Canada (Morton, 2016), has grown steadily. As one of the main goals of CLIL is increased competence in the target language (Coyle

et al., 2010; Marsh, 2002), much of this research has investigated the effects of CLIL on the target language proficiency, though there is increasing attention paid to other aspects of CLIL effects as well: attitude and motivation, mother-tongue proficiency, subject content knowledge, teaching pedagogy, and other added value effects (Bonnet & Dalton-Puffer, 2013).

Research investigating target language proficiency has tended to focus on comparison between CLIL learners and their non-CLIL peers. Results in a variety of contexts examined different aspects of CLIL pupils' learning, including gains in the target language skills of speaking, writing, reading, and listening and the sub-skills of vocabulary, grammar, and 'willingness to communicate' (MacIntyre et al., 1998). Since the umbrella term CLIL encompasses a great variety of programs in terms of duration, intensity, age, language level, selectivity and methodology (Coyle, 2007), this variety in the definition and interpretation of CLIL makes program comparison problematic (Cenoz et al., 2014). This is also the case with comparisons of research studies which employ different research designs and methodologies. In particular, some early CLIL studies failed to take a baseline measurement, so that comparisons of CLIL and non-CLIL groups were at best questionable (Dalton-Puffer, 2017). It is not surprising that there have been some conflicting results in target language proficiency gains, and contrary interpretations of some of those results (c.f. Bruton, 2011b, 2015). Some research evidence suggests that the CLIL advantage is minimal or even nonexistent, even when explicit preselection and extra resources have been taken in to account (Bredenbröker, 2002; Dallinger et al., 2016; Madrid & Barrios, 2018; Roquet & Pérez Vidal, 2015; Rumlich, 2017; Saladrigues & Llanes, 2010).

However, most studies in a variety of European countries showed that CLIL pupils indeed achieved a higher level of foreign language proficiency than their peers in regular foreign language instruction (Admiraal et al., 2006; Alonso et al., 2008; Huibregtse, 2001; Jexenflicker & Dalton-Puffer, 2010; Jiménez Catalán & Ruiz de Zarobe, 2009; Lasagabaster, 2008; San Isidro, 2010; Várkuti, 2010; Verspoor et al., 2013; Zydatiß, 2012; to name but a few). It seemed to be assumed that these larger gains were due to the CLIL approach, and variously attributed to some of its characteristics such as constructivist learning strategies and neurobiology (Ting, 2010), authenticity of context, making learning meaningful (Coyle et al., 2010; Escobar Urmeneta, 2004), and substantial additional target language input and variety (Coyle, 2006a; Dalton-Puffer, 2008a; Lasagabaster, 2008).

1.4B Selection, elitism, and ethical considerations

For several decades, however, CLIL programs remained only the prerogative of the highest achievers, and the success of CLIL has sometimes also been attributed to the learning prerequisites of pre-selected, academically skilled and motivated participants (Bruton, 2011a, 2011b; Küppers & Trautmann, 2013). Whether through explicit selection in the form of entry tests or scores (Bruton, 2011b; Goris et al., 2022; Möller, 2016; Rumlich, 2013;

Zink, 2016), implicit selection in the form of school streaming where CLIL is only offered to the higher tracks (Admiraal et al., 2006; Weenink, 2005; Zydatiß, 2007), or self-selection through which pupils or their parents are warned of the difficulties or discouraged from enrolling (Dallinger et al., 2018; Somers, 2018; Van Mensel et al., 2020), "a 'selection of the best'-approach seems to be something almost inherent in CLIL" (Küppers & Trautmann, 2013, p. 293).

The question remains whether the greater learning gains found can be attributed to the CLIL approach, or whether they are due to the high academic level and motivation of the participants. If the former—and if CLIL can also 'work' with lower-achieving learners—then why would it be withheld from them if they also stand to benefit and perhaps need it even more to raise their foreign language proficiency level? Garcia-López and Bruton question the prevalence of studies with highly motivated, academically high-performing CLIL groups, and the lack of research into non-selected, non-privileged groups of learners: "Is CLIL something that only works with specific types of students?" (2013, p. 267). In other words, are average and lower achievers excluded from CLIL programs because they cannot cope with the challenge of CLIL, or because the elitist, selective, high-status nature of CLIL ensures its success?

This issue has caused particular controversy in CLIL research. If learners are selected for CLIL programs based on their higher pre-existing English proficiency, more positive attitudes and motivation, better general academic performance, and/or stronger cognitive abilities, this could indicate not only that results comparing CLIL and non-CLIL learners are not measuring equal groups, but also that there is a creaming effect which could negatively impact the composition of mainstream, non-CLIL classes, as suggested by Rumlich (2013) and Küppers and Trautmann (2013). And if CLIL is indeed an effective approach to achieving affective, interpersonal, and intercultural goals in addition to enhanced language proficiency (Coyle, 2006a) then it seems unethical to offer these gains—particularly in public education—only to those who are already the most capable. As Huibrechtse et al. stated already in 1994, in the nacent years of bilingual education in the Netherlands, "De vraag naar geschiktheid van immersion voor alle kinderen is onder andere van belang omdat het gevaar bestaat dat immersion elitair wordt als men denkt dat alleen goede leerlingen sucessvol kunnen zijn....Wanneer een groep leerlingen de mogelijkheid wordt onthouden een hoog taalvaardigheidsniveau in de tweede taal te ontwikkelen is dat ethisch gezien minstens twijfelachtig" (p. 147)[3].

3 [The question whether immersion is suitable for all children is important because the danger exists that immersion will become elitist if people perceive that only talented pupils can be successful...Denying a group of pupils the possibility to achieve a high level of language proficiency in a second language is at best ethically questionable.]

1.5. The research gap

In order to explore the extent to which CLIL could also be successful for a non-elite population, there must be CLIL programs which include these learners, and research which includes these programs. However, there is a dearth of CLIL research in mixed-ability settings and non-selected students (Küppers & Trautmann, 2013), non-elite populations (Schwab, 2014) and in vocational contexts (Kontio & Sylvén, 2015), and generally "which does not ignore the less privileged" (Bruton, 2011a, p. 236). As described by Schwab et al. (2014, p. 4) : "Mit wenigen Ausnahmen … werden CLIL-Angebote in der Regel nur in positiv selektierten Klassen gemacht und dort auch wissenschaftlich untersucht."[4]

A range of variables—including the variety of approaches under the CLIL umbrella, differences in formal and informal exposure to a target foreign language, the structure of a country's educational system, teacher training, and socio-economic influences—complicate predicting how the average-to-lower academic achievers might fare in a CLIL program. In CLIL research there is a distinct lack of empirical evidence regarding lower-ability CLIL learners, as CLIL programs have not widely been available to them.

1.5A Empirical studies with lower-aptitude learners

The few existing studies involving lower-achieving learners in CLIL yield some conflicting results. 'Less gifted' CLIL pupils showed greater gains than a non-CLIL control group, but in a Swedish context, some L2 skills only show significantly higher results after three years (Dentler, 2002). The weaker pupils in mixed-ability CLIL history lessons in a Spanish study had shorter attention spans, difficulty with vocabulary memorization, and poorer-quality written and oral production (Escobar Urmeneta, 2004), though an increase in confidence and positive attitude was observed. In another context (Norway) with mixed-ability CLIL history lessons, the weaker pupils showed no gains in the 'willingness to communicate' construct and also struggled with comprehension, anxiety, and low oral fluency (Lialikhova, 2018). Oral proficiency was also the focus of a four-school Austrian study (Mewald, 2007) in which CLIL and non-CLIL learners were divided into three levels based on their academic achievement. The high and middle CLIL groups' oral skills were significantly enhanced, but the oral proficiency performance of the 'low' CLIL group lagged behind the 'low' non-CLIL group for fluency, sentence length, providing factual information, and creative thinking, although the low CLIL group performed better regarding grammatical accuracy, lexical precision and lexical range. Target language vocabulary acquisition by pupils with a low-achieving language learning background was the focus of another Austrian study, which showed that although the CLIL group scored higher on vocabulary in the pretest, there was no significant advantage in overall vocabulary growth (Gierlinger & Wagner, 2016). These authors suggest that the positive results for CLIL pupils' vocabu-

4 With few exceptions, as a rule CLIL is only offered to positively-selected classes, which is also where scientific research is conducted.

lary development in other studies may be from contexts with school-based selection procedures whereby pupils with high language proficiency, advantageous parental background, and academic ability gain access to CLIL (see also Bruton, 2011b; Rumlich, 2016).

In the CLIL vocational sphere in Sweden, a practical, functional context and a communicative approach seem to help make oral use of the L2 more authentic and relaxed than in more text-oriented academic subjects (Kontio & Sylvén, 2015), as well as increasing general motivation and better general school results (Van de Craen et al., 2013).

Finally, two of the most relevant studies on the effects of CLIL for learners with lower academic aptitude are respectively from the Netherlands (Denman et al., 2013), and Germany (Schwab, 2011, 2013; Schwab et al. 2014). Denman et al. reported on a qualitative study of teachers' and pupils' opinions of good practices in five schools which had recently started with continuous-curriculum pre-vocational CLIL programs. On the whole, both groups found the CLIL approach very motivating and relevant for the pupils' future education and careers. Teachers noted that pre-vocational pupils generally have short attention spans, hence the need for variety in activities and plenty of 'hands-on' work. According to the pupils, by far the most important teacher behavior is speaking English nearly all the time in class and encouraging pupils to do the same. Pupils' real-life L2 contact with international peers is perceived as particularly motivating, as it gives opportunities for authentic communication.

Schwab's studies followed a CLIL group in a German *Hauptschule*—often associated with "slow learners" (Schwab, 2011, p. 8)—for two years and found that the greatest growth in language proficiency occurred in the first year, with more improvement in the English language and communicative confidence for CLIL pupils compared to their non-CLIL peers. The CLIL approach was offered in a modular structure, with just one short CLIL module at a time. Particularly interesting in this study was the inclusion of migration background as a variable; results showed that there was no difference between pupils with a migrant background and those with an L1 German background on the L2 test results or for motivation. Pupil and teacher perspectives formed a qualitative part of the study; pupils stated that it was easier to understand the subject when it was taught in English, probably because of the intensive efforts made to scaffold the material and explain it stepwise. These pupils had already had four years of English-language instruction in primary school, and teachers expressed the opinion that pupils should have an adequate foundational level in the target language prior to commencing with CLIL.

1.5B Theoretical underpinnings

Some theoretical underpinnings which may explain the conflicting results for language proficiency development in CLIL are particularly relevant for non-privileged, lower-achieving pupils in CLIL. The linguistic threshold hypothesis (Cummins, 1979b) holds that

there are threshold language levels that must be reached in a learner's first language in order to benefit from bilingual education: the first level being essential to gain access to the educational content, and a second level necessary to allow cognitive growth. This can be particularly relevant in CLIL contexts with a high percentage of minority-language or literacy-challenged, lower-ability learners (Schwab, 2014; Zink, 2016; Zydatiß, 2012). Cummins' distinction between Basic Interpersonal Communication Strategies, or BICS, and Cognitive Academic Language Proficiency, or CALP (Cummins, 1979a, 1984b), is said to relate to these threshold levels in that a BICS threshold is necessary in order for the learner to cope with the linguistic challenges of CLIL, whereas a CALP threshold is needed in order to deal with academic content (Zydatiß, 2012). In contexts where learners have limited prior L2 instruction or exposure, the linguistic and academic challenges of CLIL will be difficult to overcome.

Another relevant theoretical concept is that of the L2 proficiency gap (Johnson & Swain, 1994), where there is "a difference between the level and type of L2 proficiency students have and the target of 'threshold' level they require in order to engage effectively with the curriculum they are required to study" (p. 211). The ramifications of this proficiency gap can be severe for CLIL pupils with lower academic aptitude and low L2 proficiency. A yawning proficiency gap can result in "incomprehensible input", leading to intellectual helplessness and resistance to learning (Otwinowska, & Foryś, 2017, p. 457).

Finally, cognitive load theory (Sweller, 2011, 2017), and particularly the "narrow limits of change principle" (Sweller, 2011, p. 54), which posits that instructional procedures that overload working memory constrict learning, may also help explain some of the difficulties of learning language and content simultaneously. Learning both a foreign language and subject content may place too many demands on working memory and thus inhibit learning, the transfer of information to long-term memory. For academically-challenged CLIL pupils this may mean that either L2 learning or content learning—or both—will be impeded. Studies have noted evidence of a possible cognitive overload in CLIL learners (Gierlinger, 2017; Mewald, 2007; Otwinowska & Foryś, 2017; Piesche et al., 2016; Prieto-Arranz et al., 2015). For this reason, Sweller (2017) warns against the integration of language and content instruction in second language learning: "Learning a second language and learning other curriculum areas should be kept separate whenever possible" (2017, p. 8). The cognitive load theory is also invoked by Ball et al. (2015, p. 11), who state that "there must clearly be a level of learner ability in the MoI [Medium of Instruction] below which teaching and learning subjects becomes ineffective. In this case, learners have to learn both new language and new subject concepts; when too much of both is unfamiliar, the cognitive load is simply too heavy." They emphasize that excellent CLIL skills and strategies on the part of the teacher are essential if learners with low L2 ability are to acquire subject knowledge.

The scarcity of empirical research and the limited application of existing theoretical perspectives to lower-aptitude CLIL contexts provide the rationale for the present study, which investigates whether concerns about the suitability of CLIL for these learners are warranted.

1.6 The Dutch *vmbo* CLIL context and the present dissertation

More research is needed to explore whether these criticisms of and concerns about the CLIL approach for pupils with lower academic aptitude and little L2 proficiency are justified. Can they cope with the extra challenges of learning content and language through the target language?

1.6A The Dutch *vmbo* CLIL context

The Dutch educational system offers an ideal context to study the progress of CLIL pupils with lower academic aptitude. As described in Section 1.1, the Netherlands has an early-streamed system in which pupils are placed into *vwo*, *havo*, or one of the four *vmbo* sub-levels on the basis of a commonly administered tests of general academic aptitude (often referred to as the *Doorstroomtoets* or *Cito-toets*) and teacher assessments. About 50% of secondary pupils attend the pre-vocational track, characterized by a shorter program length and a pupil population that is, on average, academically weaker than those in the higher tracks.

The pre-vocational track also has a higher proportion of pupils with a migrant background or non-Dutch home language than the more academic track; more pupils with parents with a low level of education or lower socio-economic status; more boys than girls; and more mixed-level classes (Inspectie van het Onderwijs, 2019). Learning disabilities, motivational issues, and behavioral problems are more frequently reported (Kneyber, 2009). Many pupils do not like to read (Broekhof et al., 2019; Land et al., 2007) and are more likely to have reading difficulties (Feskens et al., 2016; Hacquebord, 2007), which can be especially acute in the lower *vmbo* sub-levels (Du Bois-Reymond & Altinyelken, 2011; Schijf, 2009). They tend to prefer practical, hands-on activities (Denman et al., 2013), particularly during internships that form part of their curriculum (Hamstra & van den Ende, 2006).

Bilingual education in the Netherlands—almost always with English as the target language—began as a grassroots initiative in 1989 and is now firmly established in public education. A CLIL stream is currently (2025) offered at about 130 *vwo* schools and about 85 *havo* schools. A network of pre-vocational schools began in 2009; by 2012 there were 15 *vmbo* schools offering CLIL, and by 2024 this number had grown to 32, predominantly

at the highest *vmbo* level (*theoretische leerweg*) (I. Swaan, personal communication, July 4, 2024). A first national Standard for pre-vocational bilingual education (Europees Platform, 2009) specified requirements for CLIL pedagogy, curriculum, international exchanges, achievement targets, teacher qualifications, and program assessment[5]. This includes target Common European Framework of Reference for Languages (CEFR) levels (A2 for productive skills in the lower *vmbo* tracks and B2 for at least one productive skill in the higher *vmbo* tracks), teacher proficiency expectations (B2 for CLIL subject teachers—that is, teachers of non-language school subjects taught through CLIL—and C2 for English teachers, the latter required for teacher certification), and curriculum requirements (at least 30% CLIL content across the four years). The CLIL curriculum should comprise at least 30% of the total curriculum (including English as a foreign language) averaged over the four years, or about 275 curriculum hours per year, or more in some cases. In contrast, the mainstream (non-CLIL) pupils have about 120 hours of English foreign language instruction per year, although in contrast to most other countries, schools may decide how much time to allocate (European Education and Culture Executive Agency, 2017), and the target language may be used very little in the classroom (European Commission/EACEA/Eurydice, 2012; Fasoglio et al., 2015; Kordes & Gille, 2012). At the time of this research, the young pre-vocational CLIL programs were still too new to undergo national certification.

In contrast to the selective nature of many CLIL programs in *havo* and *vwo*, pupils in *vmbo* are *negatively* selected by virtue of their lower Cito scores and teacher assessments; they do not have access to these higher academic tracks and their CLIL programs. The only CLIL selection measure at these *vmbo* schools during the period of this research was parental/pupil self-selection, although a few schools have since introduced selection criteria (van Dongen, 2017) or formative English tests.

5 A new, more compact Standard for pre-vocational CLIL schools was adopted in 2018. Though the content remains largely the same, the earlier Standard also included an explanatory document (Toelichting bij de Standaard Europees Platform tweetalig *vmbo*) and a document detailing teacher qualifications. The 'Toelichting' document explicitly stated the inclusive character of pre-vocational CLIL: "*Het uitgangspunt van t-vmbo is dat alle vmbo-leerlingen, in alle leerwegen, er aan deel moeten kunnen nemen. Daarom wordt er gedifferentieerd in het gewenste eindresultaat op het gebied van de taalvaardigheid. Het minimumniveau voor de BBL en KBL is niet hoog, juist om het tweetalig onderwijs, met behoud van uniforme kwaliteitseisen, voor alle leerlingen binnen bereik te brengen.*" *(Europees Platform, 2009; document produced within the framework of the Landelijk Netwerk voor tweetalig onderwijs)*. [The guiding principle for pre-vocational CLIL is that all pre-vocational pupils, at all levels, should be able to take part. This is the reason for the differentiation in the final attainment levels for the target language. The minimum attainment levels are deliberately not high for the lower academic tracks so that the CLIL streams, with their quality control requirements, remain feasible for all pupils]. Since the newer Standard includes this differentiation in final attainment levels, the guiding principle of inclusion is built into the Standard and stating it explicitly was deemed superfluous. Another difference in the newer Standard is the inclusion of specific requirements for CLIL, including lesson pedagogy, continuous professional development, and collaboration between subject and English teachers.

Together, the structure of the Dutch system and the existence of *vmbo*-specific CLIL guidelines create favorable conditions for examining whether CLIL is effective for a non-elite, non-selected academically challenged pupil population.

1.6B Research setting and scope

This dissertation is comprised of four studies, all of which focus on Dutch pre-vocational CLIL programs and were carried out in authentic settings, with no treatment or intervention other than the CLIL approach for approximately half of the participating pupils. Data were collected between 2012 and 2014, when pre-vocational CLIL programs in the Netherlands were still in their infancy. Studies 1–3 (Chapters 2–4) used an observational longitudinal design based on a convenience sample of six pre-vocational schools with a CLIL program and one without a CLIL program. Chapter 5 involved only the CLIL pupils and teachers from those schools.

Across the studies, a total of 603 pupils participated (313 CLIL and 290 non-CLIL). The sample reflected the linguistic diversity characteristic of many *vmbo* contexts: 38.4% of pupils reported using at least one home language other than Dutch, with more than 40 non-Dutch home languages represented and Turkish and Moroccan Arabic the most common. CLIL classes included slightly higher proportions of pupils with linguistically diverse backgrounds than the non-CLIL classes. Participant numbers vary slightly across analyses due to pupil absence or incomplete data.

Although all participating schools served the pre-vocational pupil population, they differed from each other in several respects. Three schools were CLIL-only and served a combination of the lower tracks of the *vmbo* pupils (as did the one non-CLIL school, included for balance). The others ran a CLIL stream parallel to a regular Dutch stream and served only the most academic *vmbo* track. Schools also varied in their organizational structures for CLIL, their emphasis on projects or vocabulary, their alternation of English/Dutch modules, their geographical settings, their proximity to another strong CLIL program, and their teachers' experience and proficiency. No two schools were comparable even within the general *vmbo* CLIL context.

1.6C Value and aims of the present research

This is the first study in the Netherlands focusing bilingual education with the CLIL approach and its effects on pre-vocational pupils' foreign language proficiency gains, attitudes, and perceptions. It is also one of the few studies internationally to explore these goals with lower academic achieving pupils. The idea of inclusive bilingual education, whereby learners of all abilities are offered access to a CLIL program, is increasingly attracting scholarly interest but still remains woefully under-researched (Pérez Cañado, 2023). This research aims to contribute to the general body of knowledge about CLIL

and bilingual education for pupils with less privileged learning prerequisites than have dominated past research studies, by:

- using three cohorts (three different class years) to compare pupils' attitudes, language gains, and perceptions over time
- reporting longitudinal results of an attitude study comparing CLIL pupils and their non-CLIL peers
- comparing English language gains of CLIL and non-CLIL pupils for different skills: vocabulary, grammar, speaking, writing, and 'willingness to communicate'
- presenting and comparing pupils' and teachers' positive and negative perceptions of their CLIL experience
- making recommendations for pre-vocational CLIL programs and teachers in the Netherlands, also applicable to or adaptable for any CLIL context with average or lower-achieving pupils

Pérez Cañado (2016) stipulates a number of characteristics of necessary and robust CLIL research related to equity and inclusion in bilingual education. These include research design fundamentals: quantitative quasi-experimental research should use a longitudinal pre-test - post-test design, and qualitative research should use surveys with multiple triangulation. Regarding variables, studies should ensure the homogeneity of CLIL and non-CLIL groups, probe the existence of statistically significant differences, and control for potentially moderating variables. The studies in this dissertation aim to satisfy these conditions.

1.6D Overview of chapters and research questions

Chapter 2 presents the outcomes of a quantitative longitudinal study using the Model of Planned Behavior (MPB) (Ajzen, 2005; Ajzen & Fishbein, 1980) and the MPB's five attitudinal constructs—*cognition*, *subjective norm*, *perceived behavioral control*, *affect*, and *intention* to compare the attitudes towards learning English of 488 pre-vocational CLIL (n=268) and non-CLIL pupils (n=220) in three cohorts (years 1, 2, and 3) in 25 class groups across seven pre-vocational schools, including all four *vmbo* subtracks. A questionnaire using statements based on the MPB constructs was administered to pupils three times over the course of two years. Confirmative factor analysis and multilevel regression analysis were used to answer the research questions:

RQ1 Which of three competing theoretical MPB models fits better, and to what extent?

RQ2 How strong are the relationships between the intention to learn English and the other constructs of the MPB in the Dutch pre-vocational context?

RQ3 To what degree is the best fitting model, and thus the instrument based on the MPB, measurement invariant for the CLIL / non-CLIL groups? And when measurement invariance holds,

RQ4 Do these CLIL pupils score higher on the MPB components than non-CLIL pupils?

RQ5 How large are the differences in scores for CLIL and non-CLIL pupils on the MPB components in years 1, 2, and 3 and do these differences increase as pupils gain more CLIL experience?

Chapter 3 focuses on receptive English vocabulary and grammar proficiency and reports on the results of a quantitative longitudinal study comparing the English language test results of largely the same 603 pupils in the same cohorts, class groups, and schools. The test instruments developed for this study included a modified Peabody vocabulary test (Dunn & Dunn, 1997) and a lexico-grammar gap-fill test in the form of a humorous story, both conducted three times over the course of two school years. The research questions guiding this study were:

RQ1 What differences in English receptive vocabulary and grammar proficiency are there between pre-vocational CLIL pupils and non-CLIL pupils at the outset of pre-vocational CLIL?

RQ2 What is the effect of CLIL on the growth in English receptive vocabulary and grammar proficiency for pre-vocational pupils?

RQ3 Is there a differential effect of CLIL for pre-vocational pupils in different school years and at higher or lower pre-vocational levels?

RQ4 Is there a moderating influence of potentially relevant background characteristics (gender, prior academic performance, home language, or prior target language instruction) on effects of CLIL on the development of English vocabulary and English grammar proficiency?

Chapter 4 reports the outcomes of a quantitative longitudinal study focusing on English productive skills: writing, speaking and willingness to communicate (WtC) of nearly the same group of pre-vocational CLIL and non-CLIL pupils (approximately 600), who completed writing tasks and WtC questionnaires three times over two academic years. The writing test was a letter based on prompts written in Dutch. Nearly 600 pre-vocational CLIL and non-CLIL pupils participated in the writing test and WtC questionnaires, which were conducted three times over two academic years. In addition, 289 CLIL (n=147) and non-CLIL pupils (n=142) participated in two individual speaking tests, at the beginning

and end of the two-year study. A total of 2,167 writing and speaking tests were assessed for both proficiency and fluency by a team of trained assessors. WtC was measured using a self-assessment questionnaire. The research questions guiding this study were:

RQ1 What differences are there between pre-vocational CLIL and non-CLIL pupils' English writing proficiency and fluency, speaking proficiency and fluency, and WtC at the start of CLIL in secondary education (grade 7)?

RQ2 What is the effect of CLIL on pre-vocational pupils' growth in English writing and speaking proficiency and fluency, and WtC?

RQ3 What are the differential effects of CLIL on growth in English writing, speaking and WtC, dependent on pupils' grade, pre-vocational level, gender, scholastic aptitude test score, home language, and prior English-language instruction at primary school?

Chapter 5 presents the results of a qualitative study exploring the perspectives of the same pre-vocational CLIL pupils as in the preceding studies (267 pupils) and 53 of their teachers from the six schools offering a CLIL program. Both pupils and teachers completed a questionnaire asking about the most and least positive aspects of CLIL. The teachers were additionally instructed to indicate when and how much they used English for different kinds of pupil contact, and they were asked to give suggestions for improving the CLIL program at their school. To elicit more in-depth perceptions from the 'student voice', semi-structured interviews were held with 15 pupil focus groups (70 pupils total) from each of the 15 CLIL classes. The research questions guiding this study were:

RQ1 How do pre-vocational CLIL teachers in the Netherlands (RQ1a) and their CLIL pupils (RQ1b) perceive CLIL classroom practice?

RQ2 What recommendations do these teachers and their pupils make for the improvement of CLIL?

Finally, **Chapter 6** offers a summary of the results, a general discussion that includes overarching analytical themes among the four studies, a description of the present situation and future outlook, and a conclusion that includes recommendations and implications for research and practice.



2

Attitudinal factors and the intention to learn English in pre-vocational secondary bilingual and mainstream education

This chapter has been published as:
Denman, J., van Schooten, E., & de Graaff, R. (2018). Attitudinal factors and the intention to learn English in pre-vocational secondary bilingual and mainstream education. *Dutch Journal of Applied Linguistics, 7*(2), 203–226. https://doi.org/10.1075/dujal.18005.den

Note.
The original published version of this chapter uses the term bilingual education (BE) to refer to Content and Language Integrated Learning (CLIL). Throughout the remainder of this dissertation, the term CLIL is used as the preferred label for this educational approach in the Dutch pre-vocational context. In tables and statistical models, however, the labels BE and non-BE are retained where they reflect original variable names or grouping conventions used at the time of data collection and analysis; these labels are functionally equivalent to CLIL and non-CLIL.

Abstract

The effect of bilingual education (CLIL) on the attitude towards learning English of pupils in the first three years of pre-vocational secondary education in the Netherlands (n=488) was investigated. Contrary to several other CLIL studies, in the present study pupils choosing for a bilingual stream are not preselected based on their attitude or motivation. Attitude was measured using the Model of Planned Behavior (MPB). The best-fitting attitude model was one in which the perceived importance of learning English was a direct predictor of the *intention* to make an effort, and not mediated by *affect*. At the start of CLIL in year 1 (age 12), attitudinal differences between bilingual and mainstream pupils were non-significant, but after one or more years of a bilingual program, CLIL pupils scored significantly higher on four of the five MPB attitudinal constructs. CLIL appears to positively influence the attitude towards learning English of pre-vocational students.

2. Overview

Research from a wide variety of countries and contexts has shown that pupils in some form of bilingual educational programs exhibit more positive attitudes and stronger motivation towards learning English than those in monolingual streams (Doiz et al., 2014a; Mearns et al., 2020; Meriso-Storm, 2011; Sultan et al., 2012; Sylvén & Thompson, 2015; Verspoor et al., 2015). These studies have focused almost exclusively on selected, higher-achieving pupil populations (Bruton, 2015; Rumlich, 2014), and the results may not be generalizable to a non-selective, lower-achieving population. Furthermore, several studies have shown that pupils enter bilingual education with a more positive attitude towards learning English than pupils in regular educational programs (Mearns, 2015; Rumlich, 2014), though this has not yet been shown for the pre-vocational tracks. The question is whether the more positive attitudes found in these studies can also be found—and maintained over time—in a non-selective, less academically-skilled population. Such a population can be found in the Netherlands because of the highly-streamed nature of secondary education and the existence of bilingual education programs with a Content and Language Integrated Learning (CLIL) pedagogical approach for all tracks. The least academic track in Dutch secondary education is called *vmbo,* or pre-vocational secondary, and is further divided into four cognitive ability levels. In total, approximately 50% of the secondary pupil population follows a four-year pre-vocational secondary track. The four subtracks thus approximately represent 'average' to 'far below average' for academic ability. Once the pupil's subtrack has been determined, pupils and their parents are free to choose any school in that subtrack, as well as choose for either a bilingual education program (hereafter called 'CLIL program' and 'CLIL pupils') or a regular Dutch-language program ('non-CLIL'), if there is one nearby.

CLIL programs in the Netherlands date from 1989, but until recently these were limited to the university track and general secondary schools. In 2009 the first pre-vocational school started with CLIL and at the time of this research there were 27 pre-vocational secondary schools with a CLIL program, none of which yet had national bilingual (*TTO*) accreditation. Nearly all of these offer English-Dutch CLIL (de Graaff & van Wilgenburg, 2015), which is the focus of the research reported here.

The Dutch pre-vocational CLIL approach is set out in the national Standard for bilingual pre-vocational education (first version, Europees Platform, 2009; updated version, Nuffic, n.d.-b). It entails a less intensive bilingual program than the higher tracks, with subjects taught using CLIL pedagogy and English comprising approximately 30% of the curriculum. It is pertinent that there is a higher percentage of English-medium provision in the first two years (age 12–14), and a smaller amount in the last two years, when the pupils receive regular Dutch-language instruction for most of the previously English-medium subjects (age 14–17) since all final national examinations are in Dutch. The pupil popula-

tion is characterized by a higher percentage of first- and second-generation immigrants than in higher tracks, particularly in the urban regions (Hartgers, 2007), and the lower the pre-vocational track, the higher the percentage of non-Western immigrants and a related higher prevalence of home languages other than Dutch (Centraal Bureau voor de Statistiek, 2014). Clearly this population differs from the academically-inclined population comprising most CLIL programs and research. If CLIL has a positive effect on pupils' attitude towards learning English, it could be even more beneficial for this lower-achieving group. Furthermore, since secondary school pupils' attitudes towards learning English (Cenoz, 2001)—or towards school in general (Lasagabaster & Sierra, 2009; Pels et al., 2011; van Nuland, 2011)—tend to become more negative over time, it is relevant to explore whether bilingual pre-vocational programs have a mitigating effect on this downward trend.

2.1 Theoretical and conceptual framework

2.1A Attitude theory and applications in education

In a review of CLIL research in Europe, Pérez Cañado (2012) calls for studies which address "attitudinal and affective factors" and use "solid empirical evidence" (2012, p. 330). Attitudes, as defined by Keil in the International Encyclopedia of Education (1991), are positive or negative feelings of an individual towards objects, persons or ideas, and thought to be learned, changeable, and to influence behavior. Attitude theory has developed into an attitude model, the Model of Planned Behavior (MPB) (Ajzen, 2005; Ajzen & Fishbein, 1980), validated in hundreds of studies in fields such as behavioral psychology (Armitage & Conner, 2001; Conner & Sparks, 2005; Godin & Kok, 1996) and education (Davis et al., 2002; Sideridis & Padeliadu, 2001; van Schooten et al., 2004) and has shown strong predictive validity (Ajzen, 1991). In the Model of Planned Behavior (MPB), *intention*—in this study, the intention to make an effort to learn English, such as paying attention, actively participating in class, doing homework, asking questions, and heeding corrective feedback—causes the actual behavior (not measured in this study) and is in turn determined by three constructs: *affect*, *subjective norm*, and *perceived behavioral control* (*PBC*). In this chapter, key constructs from the MPB are italicized to distinguish their technical use within the model from their everyday meanings. *Affect* represents the individual's feelings about performing the behavior and is thought to be influenced by evaluative beliefs about its expected outcomes. *Subjective norm* stands for the person's own moral ideas about the desirability of the behavior, and *PBC* is the degree to which the person perceives impediments to performing the behavior (Ajzen, 2002). The MPB and its postulated causal structure are shown in Figure 2.1:

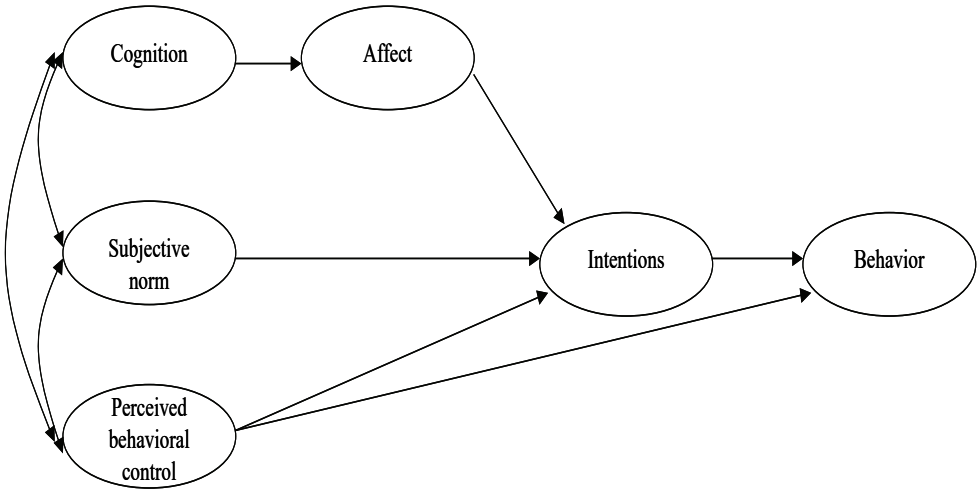


Figure 2.1: MBP as operationalized by van Schooten et al., 2004

2.1B Alternative structural models

The literature also mentions two other models, so it is relevant to consider these. Some researchers (Ajzen, 2005; Bagozzi & Burnkrant, 1980; Dillon & Kumar, 1985) suggest that for some domains, *affect* and *cognition* might be one and the same construct. Other studies show that a model in which *affect* and *cognition* each have a direct path to *intention* fits better (Bagozzi & Burnkrant, 1979, 1985; Brinberg, 1981; Fen & Sabaruddin, 2009; Triandis, 1977; Valois et al., 1988).

2.1C Adaptations of the MPB in CLIL and L2 learning research

The first question is which of these three aforementioned models is most suitable for measuring CLIL and non-CLIL pupils' attitudes towards learning English. Using the best fitting model, we can answer the following questions: how strong the relationships are between the various constructs of the MPB, and which constructs of the MPB show the strongest relationships with the *intention* to learn English. Subsequently, we also want to compare scores of CLIL and non-CLIL pupils on the MPB constructs in order to verify whether CLIL at this pre-vocational level is beneficial for the attitude towards making an effort to learn English. Before this comparison however, it is crucial to verify whether the designed instrument measures the same constructs on the same scale for the CLIL and the non-CLIL pupils (Koomen et al., 2012).

If CLIL changes the way in which pupils perceive the meaning of the MPB questionnaire items, the relationships between item-scores and latent traits might change, resulting in a difference in the constructs measured in both groups. In research the assumption is often made that a test measures the same constructs in different groups (metric invariance) and/or on the same scale in different groups (scalar invariance); in other words, is measurement invariant (Cheung & Rensvold, 2002; Meredith, 1993; Meredith & Teresi, 2006). The results of a comparison may be invalid if measurement invariance does not hold (Vandenberg & Lance, 2000); however, if both metric and scalar invariance for the CLIL and non-CLIL pupils hold, then we have 'strong factorial invariance' (Meredith & Teresi, 2006). If the measurement instrument is measurement invariant, then we will be able to explore differences in MPB scores between CLIL and non-CLIL pupils and determine whether the CLIL pupils in this study score higher on the MPB constructs than the non-CLIL pupils, even after controlling for other individual characteristics such as gender and language background.

Recalling the more positive attitude of more academically-inclined CLIL pupils in the previously-mentioned research results, we expect a causal effect of CLIL on the MPB constructs (higher for the CLIL pupils). Given that this cross-sectional measurement includes three different class years, we predict that if there is a positive causal effect of CLIL on the different MPB constructs, we will find higher scores on these constructs for CLIL pupils. We also expect that the longer pupils follow a CLIL program, the larger the differences between CLIL and non-CLIL students will be. Prior research has shown that CLIL pupils in more academic tracks have a more favorable attitude towards English from the outset than their non-CLIL peers (Mearns, 2015; Rumlich, 2014), likely caused by a twofold selection: the CLIL stream selecting pupils with a positive attitude, and by positive pupils also selecting CLIL. The present study examines if this is also the case for our non-selective, non-elite stream sample: whether pupils at the beginning of pre-vocational CLIL also have a more positive attitude towards learning English than their non-CLIL counterparts. If we do not find this difference at the start of CLIL in our sample (year 1), but do find it in year 2, a causal interpretation of the positive effects of CLIL on pupil attitude is more plausible.

The final research questions are therefore: whether the hypothesis holds that CLIL pupils score higher on the MPB constructs, and whether the difference between CLIL and non-CLIL pupils is larger in year 2 than in year 1. For differences in year 3 we had no clear expectations, as the intensity of the CLIL program diminishes starting in the third year. Because of the cross-sectional design, differences in scores between CLIL and non-CLIL pupils could be caused by factors other than CLIL participation. Therefore, differences between CLIL and non-CLIL pupils on MPB scores were also tested after correcting for some pupil characteristics.

The research questions for this article, then, are:

RQ1 Which of the three theoretical MPB models fits better, and to what extent?

RQ2 How strong are the relationships between the intention to learn English and the other constructs of the MPB in the Dutch pre-vocational context?

RQ3 To what degree is the best fitting model, and thus the instrument based on the MPB, measurement invariant for the CLIL / non-CLIL groups? And when measurement invariance holds,

RQ4 Do these CLIL pupils score higher on the MPB components than non-CLIL pupils?

RQ5 How large are the differences in scores for CLIL and non-CLIL pupils on the MPB components in years 1, 2, and 3 and do these differences increase as pupils gain more CLIL experience?

2.2 Methodology: research design and instrumentation

2.2A Context and participants

All 14 Dutch pre-vocational secondary schools with an English-language CLIL program were invited to participate and all six responding schools were included. None of these schools select pupils for CLIL based on previous motivation; four of the schools had no selection criteria whatsoever and two administered an English placement test and a brief interview[6], so the only true selection criteria is pupil self-selection. There is only minimal attrition from the CLIL streams: schools report that once first-year pupils choose CLIL, they continue for the 4-year duration of pre-vocational secondary education and almost no pupils drop out of the CLIL program. Three schools were CLIL-only, with no non-CLIL streams; these schools provided CLIL classes for the experimental group. The other three schools had parallel CLIL and non-CLIL streams; these three schools provided both CLIL (experimental group) and non-CLIL (control group) classes. One additional non-CLIL school was included in the control group. Twenty-five entire class groups of pupils from the seven schools participated (242 boys and 278 girls, mean age 13 years and 3.2 months). Each school had between two and six classes participating, for a total of 15 CLIL classes (n=293) and 10 non-CLIL classes (n=242). CLIL/non-CLIL class size was approximately the same; some CLIL pupils were logistically unable to participate in the research because of a slightly different vocational profile. The participants were in year 1 (n=251), 2 (n=168) and 3 (n=116).

6 The BE coordinators at these schools reported that this is generally a formality, and in practice almost no pupil is excluded from the BE program.

Although our sample was non-randomly selected, it does not seem very biased. Firstly, the average CITO score (a standardized national primary education exit test), the main tool used to stream pupils into the pre-vocational secondary levels, was within the range of the national CITO ranges for 24 of the 25 classes. Secondly, the different CLIL organizational approaches found in the experimental group schools reflect the different CLIL approaches found nationally. Thirdly, the percentage of pre-vocational secondary pupils with a non-Dutch parental background is 38.4% in our study, and 33.6% nationally (Centraal Bureau voor de Statistiek, 2016). All of the schools in our sample are located in the urban west, the most densely populated area of the country, where also the majority of pre-vocational secondary CLIL programs can be found.

There was no exclusion of participants on the basis of previous scores, learning difficulties, or behavioral issues. Interviews with the CLIL coordinators, teachers, and focus groups of pupils confirmed the low threshold for participating in the CLIL stream, thus further supporting the claim of non-selectivity of CLIL for these pre-vocational secondary groups. The CLIL segment of this population, therefore, cannot be called selective, unlike the CLIL streams at higher types of secondary education in the Netherlands and elsewhere (Bruton, 2011b, 2015; Cenoz et al., 2014; Rumlich, 2014; Verspoor et al., 2015).

2.2B Instruments and measures

Data were gathered by means of two questionnaires, a biodata questionnaire[7] for personal information (gender, age, parents' first language and pupils' language spoken with parents; see Appendix 2, Attachment 2.1) and an MPB questionnaire with items representing all MPB constructs except *behavior*. As there was no extant MPB instrument for measuring attitude towards learning English, the MPB items were specially constructed based on MPB questionnaire guidelines (Ajzen, n.d.), written in statement form in Dutch and evaluated on a 5-point Likert scale (1=completely disagree; 5= completely agree). It was piloted with a group of first-year pre-vocational secondary pupils from a school not participating in the research project. In a 'think-aloud' session with the head researcher, the pilot pupils told what they thought each item meant and indicated if there was any confusion. Subsequently, several of the items were re-worded or dropped. The final questionnaire contained 40 MPB items (see Attachment 2.2). Example statements for each MPB construct are:

7 The biodata questionnaire included additional background variables intended to capture socioeconomic background and extracurricular exposure to English (e.g., parental occupation, household resources, media use, gaming, and online interaction). Several of these items yielded highly incomplete, ambiguous, or internally inconsistent responses and were therefore excluded from the analyses.

- *Cognition*: People who are good at English have a better future.
- *Affect*: I like learning English.
- *Subjective norm*: I would be ashamed later if I didn't know English well.
- *Perceived Behavioral Control*: I don't have enough time to learn English (rescored).
- *Intention*: I intend to work on my English every day.

2.2C Procedure

The cross-sectional sample contains pupils in years 1–3 in class groups participating in either a CLIL program, or a regular non-CLIL program with Dutch as the only language of instruction. The questionnaires were administered at the beginning of the school year under the supervision of a teacher and a researcher. On several occasions either the teacher or the researcher had to read the statements aloud to pupils hindered by reading difficulties in Dutch.

2.3 Data analysis and results

2.3A Analytical approach and statistical procedures

The first research question (To what extent does each MPB model fit the data, and which of the three fits best?) was answered through confirmative factor analyses (Muthén & Muthén, 2006), using maximum likelihood. Exact fit is reflected by a non-significant χ^2 test. Since the power of χ^2 becomes very large with large sample sizes, and because exact fit is a very strict criterion for social science research (MacCallum et al., 1996), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA) and the Root Mean Square Residual (SRMR) were also used. Values of CFI and TLI higher than .90 (Bentler, 1992) indicate good fit. Values of RMSEA below .05 are indicative of close fit, .05–.08 fair fit, .08–.10 mediocre fit and values > .10 misfit (MacCallum et al., 1996). For SRMR, values below .08 indicate good fit (Hu & Bentler, 1999). Modification indices were used to identify items causing misfit. Nine of the 40 MPB items were removed, after which 31 items remained: 9 for *cognition*, 5 for *affect*, 5 for *subjective norm*, 6 for *PBC*, and 6 for *intention*. To determine the best-fitting model, the fits of these three nested models were compared by subtracting the values of χ^2 and the numbers of degrees of freedom of both models to create a new χ^2 test to verify the significance of fit improvement.

To answer the second research question, concerning which components of the MPB best predict the intention to learn English, the structural model of the best-fitting model (see above) is interpreted. Because suppressor effects are possible in a structural model, the zero-order correlations between sums of the items per MPB construct are also examined. The power of the structural equation modeling is influenced among other things by the number of indicators per factor and the magnitude of the factor loadings (Wolf et al.,

2013). Having five to nine indicators per factor we intended to sample at least 400 pupils, so our sample of 488 pupils used for the structural equation modeling is adequate.

To answer the question concerning measurement invariance, the resulting best-fitting model was refitted as a two-group model (CLIL-pupils vs non-CLIL pupils). Three nested two-group models were fitted: one implying strong factorial invariance (factor loadings and intercepts constrained to be equal in both groups), one implying metric, but no scalar invariance (factor loadings constrained to be equal in both groups, intercepts free to vary over groups) and one implying no measurement invariance (factor loadings and intercepts free to vary over groups). Fit improvement between nested models was verified through a χ^2 test based on the difference in χ^2 and in degrees of freedom between both models.

The fourth research question (Do CLIL pupils score higher on the MPB components than non-CLIL pupils?) is answered by means of regression analyses. Dependent variables are the sums of the items for each construct of the MPB. The reliability (Cronbach's alpha) of each of the sums is also calculated. Regression analyses were conducted separately for each sum representing an MPB construct. These analyses are also used to verify whether first-year CLIL and non-CLIL pupils differ in their scores on the MPB sums at the outset.

Since we have a clustered sample (pupils within classes within schools), analyses were conducted multilevel. Significance of fit improvement by adding a variance component to the model was verified by means of the χ^2 distributed difference in deviance of both nested models. Since variances cannot be negative, the probability found should be divided by 2 (Hox, 2010). Multilevel regression analyses were conducted using the program ML-Win (Rasbash et al., 2009). The difference between scores on MPB constructs of CLIL and non-CLIL pupils was evaluated both with and without correcting for pupil characteristics (covariates). The covariates used were the dummies indicating the pupil's pre-vocational stream level, gender, the native tongue of the father and mother (Dutch or not), and the language spoken with the father and mother (Dutch or not). Covariates were added to eliminate alternative explanations and to increase power; only covariates that significantly predict the dependent variable were included.

The significance of independent variables was calculated in two ways: first, by the χ^2 test calculated as the difference between both deviances of a model with and a model without the predictor(s) with as degrees of freedom the difference in numbers of parameters estimated in both models. Secondly, significance of predictors was evaluated by dividing the regression coefficient of the predictor by its standard error, resulting in a t-score with a number of degrees of freedom equaling the sample size minus the number of predictors minus 1 (N-p-1). For pupil level variables, the sample size is the number of pupils; for

class level variables, the number of classes; and for school level variables, the number of schools (Hox, 2010).

The last research question (How large are the differences in scores for CLIL and non-CLIL pupils on the MPB components in the three years, and do these differences increase as pupils gain more CLIL experience?) was answered by testing the effect of year and the interaction effect between CLIL and the pupil's year for each MPB construct. To estimate this interaction effect, two interaction terms for each construct were added to the model: the interaction between CLIL and year 2 and between CLIL and year 3, implying that year 1 is the reference group. These analyses were conducted both with and without correcting for significant covariates. For testing contrasts between years 2 and 3, the analyses were also repeated with year 3 as reference group.

2.3B Results: model fit and predictors

Confirmatory factor analyses to answer the first question show that the model as proposed by Ajzen (1991) shows a good fit (χ^2=811.197; df=427; p<.001; RMSEA=.043; CFI=.90; TLI=.89; SRMR=.054). All items have significant factor loadings on the factor they are supposed to measure (all p<.001). The next model is that in which *cognition* and *affect* are combined into one construct (Dillon & Kumar, 1985). This model also has a good fit (χ^2=896.255; df=428; p<.001; RMSEA=.047; CFI=.88; TLI=.87; SRMR=.057); again, all factor loadings of the items on their factor are significant (all p<.001). The difference in fit between both models is significant (χ^2=85.058, df=1, p<.001); therefore, *cognition* and *affect* should not be combined into one construct. The third model fitted is the modified MPB in which *Cognition* has a path directly onto *intention*, not mediated by *affect*. The model again shows a good fit (χ^2=755.175; df=424; p<.001; RMSEA=.040; CFI=.91; TLI=.91; SRMR=.050) and all item factor loadings are significant (all p<.001). The difference in fit between both models is significant (χ^2=56.022; df=3; p<.001). This last model, with *cognition* loading directly onto *intention*, is thus the preferred (less parsimonious) model.

To answer the second research question, the structural model of this best-fitting model was inspected. Since the structural model can be interpreted as a regression model (albeit after removing error) which means that highly correlated predictors (here: latent traits) might result in suppressor effects, we inspected not only the standardized coefficients of the structural model, but also the zero-order correlations between the item sums of all constructs measured in the instrument, taking into account that the zero-order correlations are lower because the sums contain error and the latent traits in the structural model do not. The structural model is presented in Figure 2.2 and the zero-order correlations are presented in Table 2.1.

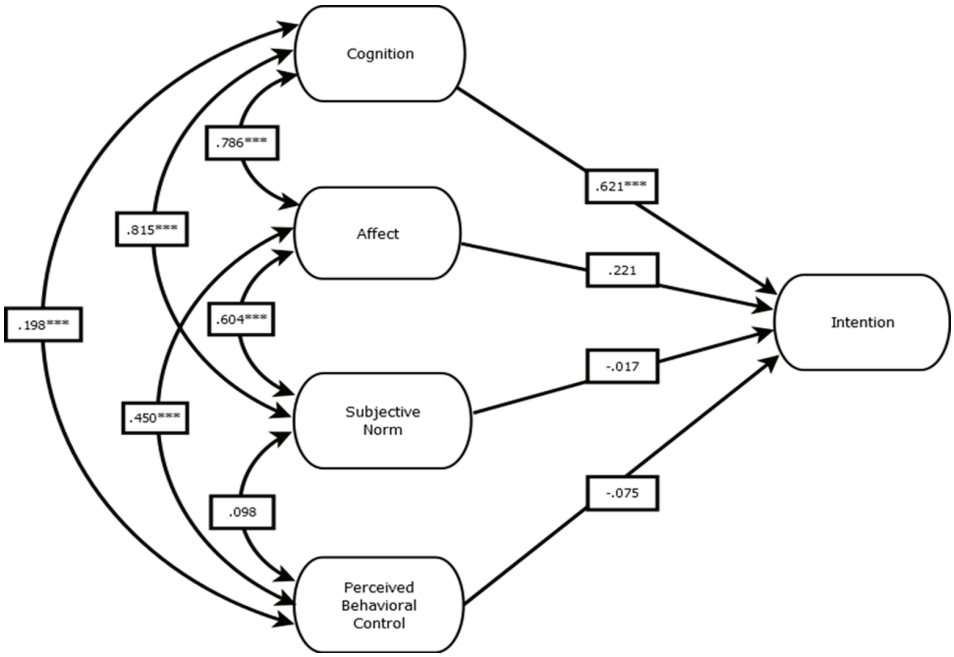


Figure 2.2: Structural part of the third MPB model tested (best fitting model) (*p<.05; **p<.01; *** p<.001). Two-directional arrows (dotted lines) indicate correlations; one-directional arrows indicate paths (loadings).

Table 2.1: Zero order correlations between sums scores MPB (higher scores = more positive attitude)

	Affect	***Sub. Norm***	***PBC***	***Intention***
Cognition	.608** (487)	.576** (485)	.074 (486)	.564** (486)
Affect		.481** (485)	.287** (485)	.553** (486)
Subjective norm			-.004 (484)	.470** (485)
PBC				.035 (486)

**. Correlation is significant at the 0.01 level (2-tailed).

In the structural model (Figure 2.2), it appears that only *cognition* has a significant loading on *intention* (standardized loading=.621). *Cognition* also correlates significantly and highly with *subjective norm* (r=.815) and *affect* (r=.786). The correlation between *cognition* and *perceived behavioral control* (*PBC*) is rather low (r=.198), albeit significant. Significant correlations are also found between *affect* and *subjective norm* (r=.604) and between *affect* and *PBC* (r=.450). The result of the model fit therefore leads to the conclusion that for the pupils in the sample, *cognition* (perceived importance) is the most important precursor of the *intention* to work hard to learn English. Furthermore, those who score higher on *cognition* also tend to score higher on *subjective norm* and *affect*, and those who score

higher on *affect* also tend to score higher on *PBC*. In order to also inspect the zero-order correlations, we first inspected the reliabilities of the sums (see Table 2.2).

Table 2.2: Reliabilities (Cronbach's alpha) of sum scores MPB

Variable	Number of items	Cronbach's Alpha	N	Range item-test correlations
Cognition	9	.69	488	.236-.437
Affect	5	.85	488	.475-.751
Sub. Norm	5	.72	486	.658-.716
PBC	6	.64	486	.239-.521
Intention	6	.76	487	.324-.589

The Cronbach's alphas for *cognition* and *PBC* are lower than for the other constructs. It is likely that these relatively low alphas were found because they measure different, non-interchangeable aspects of 'usefulness' (*cognition*) and 'practicability' (*PBC*) which do not necessarily have to show high intercorrelations in order to present a valid sum score indicative of their respective constructs (Davis et al., 2002).

Comparing the standardized regression coefficients and correlations from the structural model (Figure 2.2) with the correlations in Table 2.1, we can see that with the exception of *PBC*, all constructs show relatively high correlations with *intention*. Only *cognition* is a highly significant predictor of *intention* in the structural model, but the strong zero-order correlations between the *cognition*, *affect*, and *subjective norm* on the one hand and *intention* on the other mean that causal relationships between these constructs and the *intention* to make an effort to learn English cannot be excluded. *PBC* also correlates significantly with *affect* (zero-order correlation of .287 and in the structural model .450) but not with the other constructs.

Next, measurement invariance of the instrument for CLIL and non-CLIL pupils was verified (research question 3). To do so, first the best-fitting model was fitted separately in both groups (CLIL pupils and non-CLIL pupils). The model fit is reasonably good (the non-CLIL group χ^2=642.509; df=424; p<.001; RMSEA=.048; CFI=.87; TLI=.86; SRMR=.063: the CLIL group χ^2=669.066; df=424; p<.001; RMSEA=.046; CFI=.89; TLI=.87; SRMR=.063). Then three nested two group models were fitted (see Appendix 2, Table 2.3). Since none of the differences in fit between the models is significant, the most parsimonious model is accepted (c.f. Bagozzi & Yi, 1988). This implies that the measurement based on the third model, the modified MPB, shows strong factorial invariance over the CLIL- and the non-CLIL-group: the MPB scores of both groups of pupils can be validly compared.

It appears that all MPB sums contain significant proportions of class level variance, but no significant proportions of school level variance. Regression analyses are therefore performed multilevel with a pupil and a class level.

In Table 2.4, we can see that CLIL pupils score significantly higher than non-CLIL pupils on four of the five MPB constructs: *cognition*, *affect*, *PBC* and *intention*. Being a CLIL- or a non-CLIL pupil explains about 40% of the differences between classes for *cognition*, *affect* and *intention*; for *PBC* this is 25%. The differences in mean scores for CLIL- and non-CLIL-pupils for *subjective norm* are non-significant. The regression models also show that there is a non-significant difference between the scores of first-year CLIL and non-CLIL pupils for all attitude constructs (*cognition* t=-1.908, df=19, p=n.s.; *affect* t=-0.868, df=19, p=n.s.; *subjective norm* t=-1.341, df=19, p=n.s.; *PBC* t=0.243, df=19, p=n.s.; *intention* t=-1.673, df=19, p=n.s.).

Table 2.4: Results (effect sizes) of regression analyses with MPB-sums as dependent variables and variable indicating CLIL/non-CLIL as predictor (N varies between 483 and 488).

Dependent variable	Percentage of class level variance	Regression coefficient variable CLIL/non-CLIL (SE in parentheses)	Percentage of explained class level variance by CLIL/ non-CLIL	Percentage explained total variance
Cognition	7.7%***	.218** (0.070)	44.0%	3.4%
Affect	11.9%***	.500** (0.156)	36.4%	6.5%
Subjective norm	5.1%**	.100 (0.105)	-	-
PBC	4.1%**	.145# (0.081)	25.0%	1.0%
Intention	7.1%***	.282** (0.092)	43.9%	3.1%

#p<.10 (=5% one sided), *p < .05, **p < .01, ***p < .001
Percentages of explained variance only included if significant.

Correction for the significant covariates (year, home language, gender; Table 2.5), results in roughly the same picture, except that the differences between class means for *intention* are now explained for 66.7% by whether the class is a CLIL class or not. Likewise, after correcting for significant covariates (only for *cognition*, *affect* and *intention*) the differences in sums between CLIL and non-CLIL pupils in year 1 are again non-significant (*cognition* t=-1.822, df=16, p=n.s.; *affect* t=-0.741, df=17, p=n.s.; *intention* t=-1.673, df=19, p=n.s.).

Table 2.5: Results (effect sizes) of regression analyses with MPB-sums as dependent variables and variable indicating CLIL/non-CLIL as predictor after correcting for significant covariates.

Dependent variable	% class level variance with covariates in the model	Unstandardized regression coefficient variable CLIL/non-CLIL (SE in parentheses) with covariates in the model	% of explained class level variance by CLIL/non-CLIL after correcting for effect covariates	% explained total variance by CLIL/non-CLIL after correcting for effect covariates	Significant covariates added to the model
Cognition	8.5%***	-0.212** (0.073)	37%	3.1%	First language mother, language spoken by pupil with father, gender
Affect	17.4%***	-0.460** (0.158)	31.9%	5.6%	First language mother, first language father
Subjective norm	2.9%	-0.089 (0.091)	-	-	Year 2; year 3 (year 1 = ref. group)
PBC	-	-	-	-	No covariates significant for *PBC*
Intention	3.8%*	-0.252** (0.076)	66.7%	2.9%	Year, first language mother, language spoken by pupil with father, language spoken by pupil with mother

*p < .05, **p < .01, ***p < .001

To verify whether the difference between CLIL-pupils and non-CLIL-pupils increases for year 2 compared to year 1 (the fifth research question), the interaction between year and the CLIL variable was inspected with separate analyses for each MPB-construct as dependent variable: once with and once without correcting for significant pupil characteristics.

Appendix 2, Table 2.6 presents two models for each MPB-construct: one containing the main effects of year and CLIL and one with both interaction terms (year2*CLIL and year3*CLIL) added, enabling comparison of the nested models. Adding both interaction terms significantly improves model fit only for *affect*, explaining 37.0% of the class bound variance and 4.2% of total variance. No pupil variance is explained, as expected, since both year and CLIL are class-related variables.

Inspection of the regression coefficients reveals that for *affect* the interaction between CLIL and year 2 has a significant coefficient of -.870 (*affect* (year2*CLIL) t=2.890; df=20; p<.01 (1-sided); non-CLIL = '1', CLIL= '0'). The interaction term consisting of year 3 and CLIL is not significant (see Figure 2.3). An additional analysis with year 3 as reference group (dummy year 3 replaced by dummy year 1 in the regression model) shows that the difference for *affect* between CLIL and non-CLIL pupils in year 3 does not differ significantly

from this difference in year 2 (without covariates: t = 1.693, df=19; p=n.s.: with covariates: t= 1.910; df=17; p=n.s.). This implies that the difference in *affect* between CLIL and non-CLIL pupils is significantly larger in the second year than in the first, but that difference in year 3 does not differ significantly from the same differences in years 1 or 2.

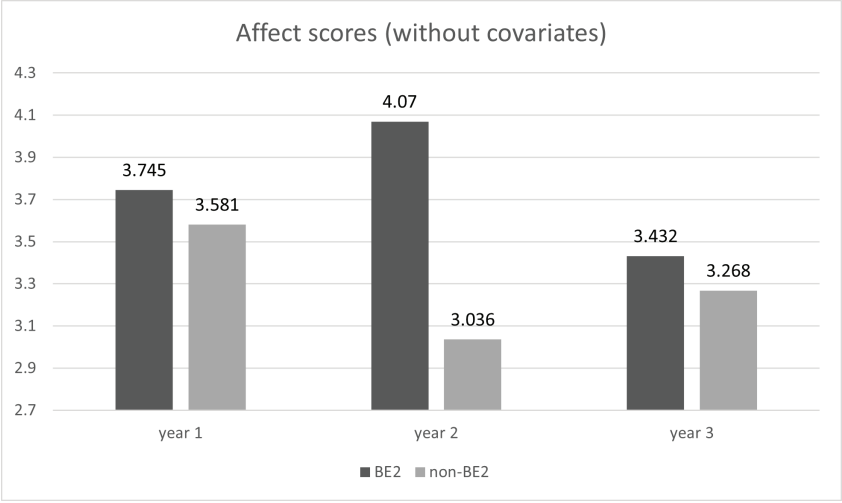


Figure 2.3: *Affect* scores for CLIL (BE) and non-CLIL (non-BE) pupils per year (without controlling for covariates)

In the analyses with *PBC* as dependent variable, both interaction contrasts in Table 2.6 (year 1 vs. year 2; year 1 vs. year 3) are positively significant for CLIL ($p<.05$), but the difference between models with and without the interaction terms is only significant at 10%. The proportion of explained variance at class level for *PBC* is larger than for *affect*, 53.8%. The proportion of class level variance in *PBC* scores is very small, however, resulting in only 1.4% explained total variance. An additional analysis with year 3 as reference group shows that for *PBC* the contrast for the difference between year 2*CLIL vs. year 3*CLIL is not significant (no covariates: t=-0.026/.200=0.130; df=19; p=n.s.; see Figure 2.4).

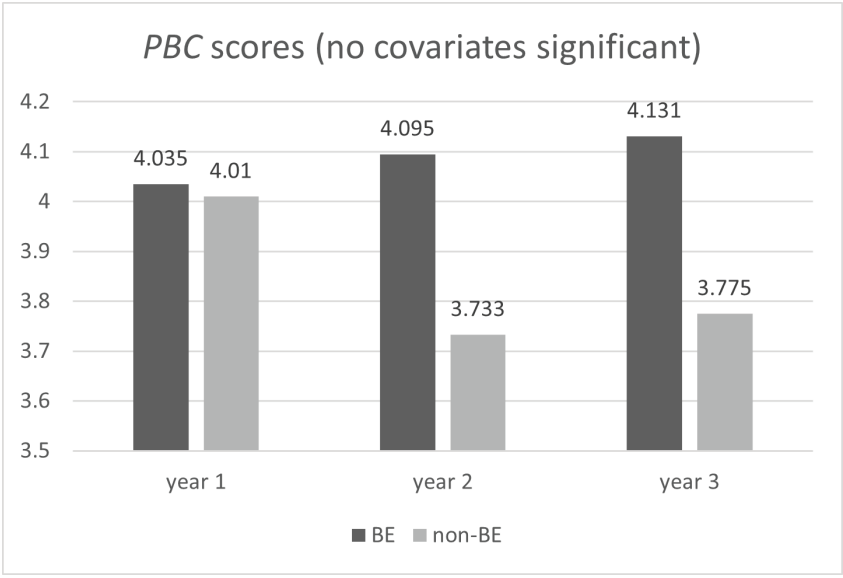


Figure 2.4: *PBC* scores for CLIL (BE) and non-CLIL (non-BE) pupils per year

2.4 Discussion and conclusions

2.4A Interpretation of findings

The data collected from over 500 pre-vocational secondary pupils yielded revelatory results concerning the prediction of their intention to learn English and the difference between CLIL and non-CLIL pupils' attitudes in years 1, 2, and 3 towards learning English. The first research question concerned the choice of the best structural model. Analyses showed that the best-fitting model was one in which *cognition* and *affect* served independently as predictors of *intention*. Although highly correlated (correlation in the structural model .786, $p<.001$), it appears that these two constructs represent distinct aspects of the attitude towards learning English (Breckler & Wiggins, 1989; Trafimow & Sheeran, 1998): broadly stated, utilitarian goals versus emotional factors. For instance, a pupil might intend to work hard because he or she enjoys learning English (*affect*), but also because he or she believes that a good command of English will be useful at senior vocational school (*cognition*). This is noteworthy, as many other studies (reported in Ajzen, 1991) have found that the effect of *cognition* on *intention* is mediated by *affect*, implying that *cognition* does not influence *intention* directly.

The second research question asks which components of the MPB best predict *intention*. Although the regression coefficient from *cognition* to *intention* shows the strongest relationship, the zero-order correlations between two of the other three predictors (*affect* and *subjective norm*) are also high, implying they are suppressed to a certain extent by *cognition*. Still, *cognition* predicts best, and generally explains the same variance which

the other two predictors share with *intention*; only *PBC* shows a minimal relationship with *intention*, both in the model and in the zero-order correlations. Apparently, the perceived importance and usefulness of learning English are quite relevant for teenagers who are generally not academically-inclined.

The best-fitting structural model results show that there are high correlations between believing learning English is important (*cognition*), and liking it (*affect*), and between believing learning English is important and perceiving social pressure to learn it (*subjective norm*). There is a medium-strength correlation between enjoying learning English and perceiving few impediments to learning it (*perceived behavioral control*); this correlation might be lower because pupils like learning English for various non-academic reasons, such as gaming, television, films, music, and social media, and perceive few difficulties in learning it, which may be partly due to the prevalence and accessibility of English in daily life in the Netherlands. Finally, there is a weak but still significant correlation between believing English is important and perceiving few impediments to learning it, as was also found by Greene et al. (2004). Importantly, though, the structural equation modelling showed that in our sample *PBC* does not contribute to the prediction of *intention*: a perception of impediments to learning English does not predict the intention to make more or less effort. *PBC* is in relation to the other constructs the 'odd one out'.

One strength of this kind of behavioral model result is its possible contribution to designing effective interventions for producing behavioral change (Fishbein, in Conner & Armitage, 1998), so the results concerning the strength of the predictors may be useful in designing interventions to elicit the desired behavior at school. Since the structural model results are for all pupils in the study (CLIL and non-CLIL together), these possible interventions are not limited to a CLIL population. For instance, CLIL programs in the Netherlands include a 'European and International Orientation' (EIO) aspect which may indirectly emphasize the importance of English globally; this could be strengthened and extended to non-CLIL pupils in order to tap into the strength of the *cognition* construct.

The nested two-group models (CLIL and non-CLIL pupils) were fitted in order to determine measurement invariance (research question three). The resulting analysis showed strong factorial invariance, so that the results from the two groups could validly be compared in order to answer the fourth research question, concerning the comparison of attitudes of CLIL and non-CLIL pupils. The results here confirm the results of other studies (Doiz et al., 2014a; Lasagabaster & Sierra, 2009; Mearns et al., 2020; Merisuo-Storm, 2011; Rumlich, 2014) regarding a more positive attitude among CLIL pupils: the CLIL pupils scored significantly higher on the model constructs *affect*, *cognition*, *intention* and *PBC* than non-CLIL pupils. The results show that this more positive attitude was not yet present in first-year pupils at the start of CLIL, as none of the MPB constructs showed a significant difference between the CLIL and non-CLIL pupils at the outset; the two groups

were comparable in terms of attitude towards making an effort to learn English at the start of pre-vocational secondary education. This is in contrast to results for more challenging secondary streams, as was found in Mearns (2015) and Rumlich (2014), or highly selective programs (Doiz et al., 2014a), and supports the schools' assertion that there was no pre-selection of the CLIL pupils based on attitude.

The final research question concerned the difference in MPB construct scores between CLIL and non-CLIL pupils in years 1, 2, and 3. The interaction effects for *affect* and *PBC* between CLIL/non-CLIL and the pupil's year show that the differences between CLIL and non-CLIL pupils are significantly larger in second-year than in first-year pupils. This suggests that CLIL may have contributed to a growth in positive attitude, which is a salient finding, as Mearns et al. (2020) found that attitude differences in a more academic secondary stream were lower in third-year than first-year pupils. In year 3 in the current study—when CLIL is substantially reduced—the difference in *affect* between CLIL and non-CLIL pupils disappears and is about the same as in year 1. However, for *PBC* the CLIL advantage seems to be maintained in year 3.

As noted in the Introduction, the CLIL programs at pre-vocational secondary level in the Netherlands form a much more substantial part of the curriculum in years 1 and 2 than in years 3 and 4 due to Dutch-language vocational profiling and Dutch final examination preparation in the latter two years. This greater focus on CLIL in the first two years may explain some of the results in this study: as the programs shrink and third-year CLIL pupils change to regular Dutch-language schooling, the positive effects of CLIL on *affect* seem to diminish. Another explanation could be the general trend towards less enthusiasm for learning as pupils age and progress to higher years, as reported by Mearns (2015) and Mearns et al. (2020).

Regarding the increase in *perceived behavioral control*, which is related to autonomy and self-efficacy, it may be that the extra exposure to English in the CLIL group, the heterogeneous input from different teachers, and a wider variety of language strategies give pupils the tools and skills to exert more control over their own learning, and that these tools and skills do not diminish when the CLIL program ends. For *PBC*, the benefits of CLIL for pre-vocational secondary pupils seem to support the purported "sustainable advantage[8]" of pre-university bilingual education in the Netherlands that was found in Verspoor and Edelenbos (2011, p. 3) and Verspoor et al. (2015). For *affect*, however, this sustainable advantage seems to dwindle after the initial period of more intensive CLIL. Put simply, learning English may not be as enjoyable once the amount of CLIL decreases. *Affect* is, after all, the second-best predictor of *intention*, and sustained enjoyment of learning English indicates a sustained intention to make an effort to learn it. Clearly, longitudinal

8 [*een duurzame voorsprong*]

data analyses are needed to further support these claims. For this study, at least, CLIL and its related CLIL pedagogy seem to contribute to an increase in enjoyment and motivation of a wide variety of lower-achieving language learners, appearing to confirm some of the side effects of CLIL stated by Coyle (2006a, 2013): learner engagement and involvement, enjoyment, confidence, self-esteem, motivation, and autonomy.

2.4B Limitations and recommendations

Since the data are cross-sectional, definitive assumptions cannot be made about pupils' attitude development over time. Because it was not possible to obtain a measurement of actual behavior, we can only predict behavior from the *intention* construct; a measurement of behavior could be included in order to verify or disprove the prediction of behavior from intention. Further research could replicate this study using a longitudinal design to establish whether the expected positive causal effect of CLIL on the constructs of the attitude model can be verified. Such a study might also explore the correlation between the attitudes of CLIL/non-CLIL pupils and their language proficiency development over time. Considering suggestions in the literature regarding potential benefits of CLIL for boys, a similar study with emphasis on gender differences might also be relevant (see Mearns & de Graaff, 2018b). Finally, a similar study could be carried out in another national context, although outside the Netherlands this is difficult, as CLIL is seldom offered specifically for negatively-selected 'average to below average' secondary pupils. It is hoped that the present study will encourage more research into bilingual education and CLIL for a non-selective population.



3

CLIL in pre-vocational secondary education: gains in target language vocabulary and grammar

Abstract

Research has shown that pupils in Content and Language Integrated Learning (CLIL) programs in academic secondary schools usually attain significantly higher levels of English language proficiency than their non-CLIL peers. In the present study we explore whether this is also true at pre-vocational level, given the lower academic level in the population, absence of selection criteria, and less intensive CLIL program aims.

This two-year longitudinal quasi-experimental study followed three cohorts of CLIL (n=313) and non-CLIL (n=290) pupils, from 7 different schools. Three repeated measurements of language proficiency (English as second/third language) were analyzed using multilevel linear growth models. CLIL pupils in most cohorts showed a small but significantly positive growth advantage over the non-CLIL pupils for target language vocabulary and grammar. The greatest growth and greatest difference in growth between the CLIL and non-CLIL pupils occurred in the first two years of the four-year pre-vocational program, implying that the initial growth rate for CLIL pupils was not sustained.

3.1 Introduction

Bilingual education (BE) with the Content and Language Integrated Learning (CLIL) approach, in which the foreign language "is used as a tool in the learning of a non-language subject in which both language and subject have a joint role" (Marsh, 2002, p. 58), is generally the preferred bilingual educational approach in Europe nowadays (Nikula, 2017a). Europe also holds a dominant position in CLIL research (Pokrivčáková, 2013), and studies to date show generally positive results for CLIL learners' foreign language proficiency development (Gálvez Gómez, 2021; Navés, 2011; Pérez Cañado & Lancaster, 2017; Pittas & Tompkins, 2024); in Europe, this foreign language is most commonly English (Dalton-Puffer, 2011), as it is in the current study. However, CLIL provision has also been called discriminatory and elitist, giving CLIL pupils a "selection-based advantage" (Rumlich, 2017, p. 129) and excluding less socio-economically or educationally favored learners (Bruton, 2013; Paran, 2013). Whether these less privileged CLIL learners can also show positive results in the proficiency development in the foreign language is the focus of the current study.

3.2 Literature review: selection effects and language development in CLIL

3.2A Selection effects in CLIL

Most CLIL programs are selective, either overtly (Alonso et al., 2008; Breidbach & Viebrock, 2012; Broca, 2016; Goris et al., 2013; Huibregtse, 2001; Möller, 2017; Sieben & van Ginderen, 2014) or covertly (Garcia López & Bruton, 2013), which Apsel (2012, p. 49) calls "a silent agenda of selectivity", or both (Van Mensel et al., 2020). This results in educational creaming and a population which has been called elitist (Weenink, 2005), as pupils with less academic ability and lower SES tend to be excluded (Küppers & Trautmann, 2013). Hence most CLIL research is conducted with higher-achieving students, with a gap in CLIL research for less academically-talented or less motivated learners (Dalton-Puffer & Smit, 2013). What scant research there is seems to indicate that bilingual education with a CLIL approach also has the potential to be effective for average- to lower-achieving populations (Brevik & Moe, 2012; Gierlinger & Wagner, 2016; Merisuo-Storm, 2007; Mewald, 2007).

3.2B Initial differences between CLIL and non-CLIL pupils

Whether CLIL programs select the more academically gifted learners, or more cognitively able or more motivated learners choose for CLIL, the question arises whether a comparison of beginning CLIL and non-CLIL pupils' language proficiency might already be skewed from the start. This question is all the more important since in a quasi-experimental evaluation of CLIL, the absence of an initial difference between experimental

and control groups enhances the internal validity of the quasi-experiment (van Breukelen, 2013). Such a baseline comparison has been done by using school marks in English (Broca, 2016) or a pre-test measurement instrument (Admiraal et al., 2006; Bredenbröker, 2002; Küppers & Trautmann, 2013; Sylvén & Ohlander, 2014; Verspoor et al., 2010). In all cases the CLIL pupils in these explicitly or implicitly-selective CLIL streams outscored their non-CLIL peers at the outset; would this also be true for pre-vocational CLIL pupils with no program-based selection procedures?

3.2C Vocabulary development in CLIL

A second question is whether such non-selected, less academically-inclined CLIL pupils would also make greater gains in the target language than non-CLIL pupils. Research on foreign language proficiency gains in bilingual education with the CLIL approach has focused on a wide range of language skills, including various aspects of vocabulary and grammar acquisition. Results from studies on receptive vocabulary have generally found a clear advantage for the CLIL learners as compared to their non-CLIL peers (Dalton-Puffer, 2017). Cross-sectional studies comparing a CLIL group with a non-CLIL group in primary education (Canga Alonso, 2015; Jiménez Catalán & Ruiz de Zarobe, 2009) and at secondary level (Sylvén & Ohlander, 2014; Verspoor et al., 2013), or both (Madrid & Barrios, 2018; Merikivi & Pietilä, 2014) show that the CLIL pupils scored significantly higher than their non-CLIL peers on vocabulary tests. Similarly, longitudinal studies from a variety of contexts report an advantage for CLIL pupils on growth in target language vocabulary (Goris et al., 2013; Iglesias Diéguez & Martínez Adrián, 2017; Pérez Cañado, 2018; Sylvén, 2004; Sylvén & Ohlander, 2014; Verspoor et al., 2010). These positive results are ascribed not only to CLIL, but also, for instance, to the amount of time spent in a CLIL program (Pérez Cañado, 2018), amount of exposure to the target language (Jiménez Catalán & Ruiz de Zarobe, 2009), and the near-exclusive use of the target language in the classroom (Verspoor et al., 2010).

On the other hand, some research reports less favorable effects of CLIL on vocabulary development (Admiraal et al., 2006; Arribas, 2016; Gierlinger & Wagner, 2016; Rumlich 2017; Saladrigues & Llanes, 2010). The lack of a significant CLIL growth advantage can have different causes in different contexts. Admiraal et al. (2006) point to the need for improvement in CLIL teachers' own proficiency and skill in providing comprehensible input and corrective feedback. Arribas (2016) attributes it in part to an irregular implementation of CLIL methodology. Gerlinger and Wagner (2016) note that positive vocabulary growth results for CLIL pupils seem to be found in contexts with a school-based selection procedure, which their sample does not have. In Rumlich's (2017) extensive study these results seemed to be partly due to controlling for a large number of confounding variables.

Except for the study by Gerlinger and Wagner (2016), none of the aforementioned studies explicitly included lower achievers in their sample, so it is unclear if effects of CLIL would be found with an exclusively lower-achieving group of CLIL pupils.

3.2D Grammar development in CLIL

The effects of CLIL on grammar proficiency have also been studied. Cross-sectional studies found significant differences in favor of the CLIL groups, both in primary (Jiménez Catalán et al., 2006) and secondary education (Lasagabaster, 2008; Martinez Adrian & Guttierez Mangado, 2015; Navés, 2011; Navés & Victori, 2010; Verspoor et al., 2013; Zydatiß, 2012). Longitudinal studies show both significant growth differences in favor of the CLIL pupils (Pérez Cañado, 2018; Pérez Vidal & Roquet, 2015) and non-significant differences (Bredenbröker, 2002; Roquet & Pérez Vidal, 2015). Goris et al. (2013) found significant positive effects of CLIL in a Dutch sample, but non-significant effects in a German sample. As with the vocabulary results noted above, the positive CLIL advantage for grammar is speculated to be not only a result of CLIL itself, particularly in the long run (Pérez Cañado, 2018), but also of some kind of selectivity (Goris et al., 2013). As with the vocabulary studies mentioned above, none of these grammar results were obtained using a specifically lower-achieving sample of CLIL pupils, and similarly, no generalization can be made about the performance of these pupils in the area of grammar knowledge.

Considering these results for vocabulary and grammar, it appears that there are different results in different school years and contexts, though most studies indicate a positive effect of CLIL on growth of target language vocabulary and grammar. A relevant question, thus, is whether positive effects of CLIL can also be found in a less academically-privileged secondary education population.

3.2E Learner variables and moderating effects: gender, prior achievement, home language, prior English instruction

Some studies have examined whether specific personal characteristics such as gender, prior general academic achievement, home language, and prior L2 instruction moderate the effects of CLIL. In other words, a number of background characteristics of pupils may enhance or hinder the effect of CLIL on learners' progress in foreign language learning, and CLIL research should include these variables in order to more specifically estimate the effect of CLIL. In the following we elucidate the choice of moderator variables in this study.

Regarding gender, in a meta-analysis (Voyer & Voyer, 2014) an advantage in general scholastic achievement for girls over boys was found, with the highest effect sizes for foreign language subjects. Research has also shown that girls generally have more motivation to learn foreign languages than boys (Dörnyei et al., 2006); some studies suggest that

CLIL may help to reduce these gender-related differences (Heras & Lasagabaster, 2015; Lahuerta Martínez, 2020; Merisuo-Storm, 2007, Roquet et al., 2016; San Isidro, 2010).

Some studies report significant larger effects of CLIL on target language proficiency for students having a higher prior academic performance (Admiraal et al., 2006; Huibregtse, 2001; Verspoor et al., 2015). Other studies (Dallinger et al., 2016; Sylvén & Ohlander, 2014) show that controlling for prior achievement renders the differences in language proficiency gains between CLIL and non-CLIL students non-significant. In studies from Broca (2016), and Rumlich (2017) conducted respectively in Spain and Germany, no correlation between language proficiency growth and prior academic level is found, but higher marks for CLIL students prior to the start of the CLIL program and teacher recommendations for CLIL based on pupils' prior academic performance have been noted.

In studies on CLIL (Dallinger et al., 2016; Schwab et al., 2014; Sierra & Lasagabaster, 2008; Somers, 2018), home language, specifically for pupils with a migration background, has not been found to influence the effects of CLIL. Somers (2017, p.210) concludes that CLIL seems to 'level the playing field' between migration-background and native-background pupils: pupils with a migration background can perform equally well in learning the foreign language as their majority-language peers. Dallinger et al. (2016) and Sylvén and Ohlander (2014) report that the percentage of migration background pupils choosing CLIL is higher than the percentage of native pupils choosing CLIL, although this was only the case in one of the three 'Gymnasium' schools in Zydatiß' (2007) study, which might be more of an effect of selection than choice. Where choice is offered, the higher percentage of migration background pupils may indicate that they feel empowered by learning part of the curriculum through a language which is also a foreign language for the majority-language pupils.

Finally, research has considered the effects of prior target language instruction at primary level on proficiency gains in CLIL at secondary level. In some federal states in Germany, preselected CLIL pupils receive extra CLIL instruction at primary school in preparation for CLIL at secondary school (Dallinger et al., 2016; Rumlich, 2016; Zydatiß, 2007). Dallinger et al. (2016) found no significant effect of CLIL after controlling for prior proficiency. Investigating the effect of this prior instruction, Rumlich (2017) found it to have twice the effect size of any other variable, including CLIL. However, in some other European contexts, mainstream primary school English instruction uses an informal, communicative, oral approach, which may not prepare pupils for a more grammar-based, teacher-centered approach in English lessons at secondary level (de Kraay, 2016; Lasagabaster, 2011).

3.2F Summary and research gap

In short, a relevant question is whether the aforementioned individual characteristics moderate the effects of CLIL on the development of target language proficiency of aca-

demically non-elite CLIL pupils compared to their non-CLIL peers. Considering Bruton's view that most CLIL programs are inherently discriminatory, "especially for the less academically able, the less linguistically proficient, or the less economically privileged" (2013, p. 596), and Somers' (2017) examples of immigrant minority-language background pupils being explicitly or implicitly discouraged from participation in a CLIL stream, the question arises whether bilingual education with a CLIL approach contributes to increased target language proficiency of a self-selected, academically average-to-far-below-average pupil population. The highly-streamed educational system in the Netherlands, together with the presence of national guidelines and standards for CLIL at pre-vocational secondary level, contribute to ideal circumstances for researching this question.

3.3 Methodology and research design

3.3A Research questions and hypotheses

Against this background, and in response to the identified research gap concerning the effects of CLIL for academically non-elite pupils in the Dutch pre-vocational context, the present study addresses the following research questions:

RQ1 What differences in English receptive vocabulary and grammar proficiency are there between pre-vocational CLIL pupils and non-CLIL pupils at the outset of pre-vocational CLIL?

RQ2 What is the effect of CLIL on the growth in English receptive vocabulary and contextual grammar proficiency for pre-vocational pupils?

RQ3 Is there a differential effect of CLIL for pre-vocational pupils in different school years and at higher or lower pre-vocational levels?

RQ4 Is there a moderating influence of potentially relevant background characteristics (gender, prior academic performance, home language, or prior target language instruction) on effects of CLIL on the development of English vocabulary and English grammar proficiency?

Based on the results of the previously-mentioned studies comparing vocabulary and grammar gains of secondary school CLIL pupils compared to their non-CLIL peers, our hypothesis was that the CLIL pupils' growth in English language proficiency would be significantly higher than that of the non-CLIL pupils for both measures of proficiency. We further hypothesize that the greatest growth would be shown in the first two years of secondary pre-vocational education, as the number of CLIL subjects and CLIL hours decrease after this period.

3.3B Context of Dutch pre-vocational CLIL

In the Netherlands, secondary education is highly streamed. Primary school pupils take a national standardized test (called 'Cito' or *doorstroomtoets*, further education assessment test) at age 11-12 which aims to measure verbal and mathematical ability and is considered a measure of general academic ability. The score on this test, together with the official recommendation of the primary school, determine the type of secondary school to which the pupil can be admitted. Approximately 50% of the pupils in the Netherlands are recommended for the six-year pre-university or five-year pre-professional secondary school tracks, while the other half proceed to the four-year pre-vocational secondary education (voorbereidend middelbaar beroepsonderwijs', or *vmbo*), which in turn usually has four sub-levels, hereafter called *pre-vocational levels*. In theory these four sub-levels are distinct, from highest to lowest in terms of academic challenge; in practice, however, they are more fluid and it is more accurate to speak of generally 'higher' and generally 'lower' (see section 3.6 for a description of the variable 'pre-vocational level' in this study). Approximately half of this pre-vocational population, both mother-tongue and migration-background pupils, have severe difficulties with reading comprehension in Dutch (Inspectie van het Onderwijs, 2008; OECD, 2023) and can be described as 'low achievers' (Trapman et al., 2014) based on the aforementioned Cito scores.

Analysis by the Dutch Education Inspection (Inspectie van het Onderwijs, 2016), has shown that pupils with lower-educated or lower-SES parents are significantly less likely to be recommended for the higher secondary tracks and more likely to attend one of the pre-vocational tracks, which also have a higher percentage of minority home language and migration background pupils. The national percentage of pre-vocational pupils with a home language other than the national language, Dutch, was 33.6% at the time of this research, compared to 22.3% for all secondary pupils (Centraal Bureau voor de Statistiek, 2016).

While the majority of the bilingual education streams in the Netherlands are in the more academic tracks, there is a small but enthusiastic CLIL presence at pre-vocational level, with 15 schools at the time of this research recruitment (2011) offering either a CLIL stream in parallel to a mainstream non-CLIL stream, or a schoolwide CLIL program. CLIL is most often found at the highest of the four pre-vocational levels, which is best characterized as 'average' in terms of academic ability, and—though rarer—there are a few CLIL streams and CLIL-only schools at the lower, more practice-oriented pre-vocational levels as well. English is by far the most common foreign language for CLIL. The Dutch national Standard for CLIL provision for pre-vocational schools at the time (Europees Platform, 2009) stipulated that 30% of the curriculum be given primarily in the target language, which amounts to approximately 275 hours per year including the English language lessons. The CLIL subjects must include at least one in the natural sciences and one in the social sciences. In the current study these ranged from mathematics, biology, and geography to

drama, arts, and cooking. Since the national final examinations at the end of the four-year pre-vocational secondary program are conducted in Dutch, the CLIL programs tend to be more concentrated in the first two years, tapering off in the final two years so that pupils become accustomed to using Dutch for the final examination subjects. This means less target language input time and fewer CLIL subjects in the last two years (years 3 and 4, age 14-17), and those CLIL subjects are then mainly non-academic, non-exam subjects such as crafts, drama or international projects (Wouters, 2012). While CLIL programs in the pre-university and pre-professional secondary schools have target proficiency levels for third-year pupils of respectively CEFR C1/B2 and B1+/B2, the aims are less ambitious for the fourth-year pre-vocational CLIL pupils: CEFR level A2 (lower pre-vocational levels) up to B2 (higher pre-vocational levels) for at least one receptive and one productive language skill. These proficiency levels are set out in the Dutch Standards for Bilingual Education (Nuffic, n.d.-b); the lower proficiency goals for pre-vocational education are specifically intended to make CLIL accessible and achievable for all.[9]

In contrast to the explicitly selective nature of most CLIL programs—whether through entry tests of general cognitive and academic ability, prior English level, attitude and motivation, or participation in a CLIL preparatory program (c.f. Rumlich, 2016)—pupils at pre-vocational level in the Netherlands are actually *negatively* selected by being excluded from the academically-higher secondary streams and the more intensive, more selective CLIL programs, and enter CLIL solely based on self-selection by the pupils and their parents.[10]

3.3C Design

The current study employed a two-year longitudinal quasi-experimental design with three repeated measurements, in a sample consisting of three school year cohorts, in which growth in English-language grammar ability and in receptive vocabulary was compared between pupils in CLIL streams and non-CLIL streams in pre-vocational education in the Netherlands.

3.3D Participants

In total, 603 pupils were included in this sample, (CLIL n=313 and non-CLIL N=290, comprising three cohorts (see Table 3.1). The cohorts consisted of pupils starting in year 1/

9 "The fundamental idea of pre-vocational CLIL is that all pre-vocational pupils, at all pre-vocational levels, should be able to participate. For that reason, the target language proficiency aims are differentiated. The minimum attainment level for the lower pre-vocational levels is not high, specifically in order to ensure that it is made accessible to all pupils, while still adhering to certain uniform demands of quality." *["Het uitgangspunt van t-vmbo is dat alle vmbo-leerlingen, in alle leerwegen, er aan deel moeten kunnen nemen. Daarom wordt er gedifferentieerd in het gewenste eindresultaat op het gebied van de taalvaardigheid. Het minimumniveau voor de BBL en KBL is niet hoog, juist om het tweetalig onderwijs, met behoud van uniforme kwaliteitseisen, voor alle leerlingen binnen bereik te brengen."]* - Toelichting bij de Standaard Europees Platform tweetalig vmbo (Europees Platform, 2009, p.1).

10 However, even conscious self-selection is not always present: focus group interviews with pre-vocational CLIL pupils in schoolwide-CLIL schools revealed that some pupils didn't even realize that they would be participating in a CLIL stream, having chosen the school for location, proximity, friends, or sibling reasons. See Chapter 5.

grade 7, age 12-13; year 2/grade 8, age 13-15; and year 3/grade 9, age 14-17), followed simultaneously for two years. Cohort year 1, therefore, started at the beginning of pre-vocational school until the end of year 2; cohort year 2 from the beginning of year 2 until the end of year 3; cohort year 3 from the start of year 3 until the end of the pre-vocational program in year 4. The sample included students of all four sub-levels or mixed levels of Dutch pre-vocational secondary school. There were 25 classes (15 CLIL classes and 10 non-CLIL classes) in seven schools; three of these schools offered only a CLIL stream, three offered a CLIL stream and a Dutch-language regular stream, and one offered only a Dutch regular stream. The schools were all located in the densely-populated western part of the Netherlands in urban or semi-urban areas. Over 40 different home languages were represented in the sample, and the percentage of pupils with a home language background other than Dutch, the national language and the default school language, was 38.4%. Of these, Turkish and Moroccan Arabic were by far the most common home languages, as these are the most common migration background languages in the Netherlands. English was used as a home language by less than 0.5% of the sample.

Table 3.1: Distribution of participants by CLIL/non-CLIL and cohort

CLIL or non-CLIL	cohort 1	cohort 2	cohort 3	TOTAL
CLIL	151	100	62	**313**
non-CLIL	163	70	57	**290**
TOTAL	**314**	**170**	**119**	**603**

With regard to the five pupil background characteristics, none showed significant relationships with participation in CLIL. Multilevel regression analyses, with participation in CLIL (yes/no) as the independent variable and each background characteristic as the dependent variable, indicated that none of the background variables were significantly related to participation or non-participation in a CLIL stream.

For the background variable *gender* (logistic regression), two variance levels were required (pupil and class; the rationale for including variance levels is explained in Section 3.6), and no significant regression effect was found (Wald test: $z = 0.012$, p = ns, $N = 603$). For *Cito* scores, three variance levels were required (pupil, class, and school), and the regression effect was again non-significant (Wald test: $z = 1.516$, p = ns, $N = 333$). For *home language*, three variance levels were required**,** and no significant regression effect was observed (Wald test: $z = 0.930$, p = ns, $N = 565$). Similarly, for *years of primary English instruction*, three variance levels were required, with no significant regression effect (Wald test: $z = 0.496$, p = ns, $N = 566$), and for *frequency of primary English lessons*, three variance levels were required**,** again yielding a non-significant regression effect (Wald test: $z = 1.510$, p = ns, $N = 565$).

3.3E Instruments

The instruments used in this study consisted of a biodata questionnaire, a receptive vocabulary test, and a lexico-grammar test.

Biodata questionnaire

Pupils filled in a questionnaire with personal characteristics including gender, general achievement test score (Cito), home language background, and number of years and frequency per week of primary school English lessons. The home language environment variable was created by adding the sum of four dichotomous variables representing the native language of the father and mother (1 for Dutch; 0 for other) and the language spoken by the pupils with his or her father and/or mother (Dutch or other), with a total score of 4 representing all Dutch, and a total score of 0 representing no Dutch. The years of primary school English was scored from 0 (no primary school English) to 4 (four or more years of primary school English), and the frequency of these lessons from 0 (no primary English) to 4 (four or more lessons per week). The relevant parts of the biodata questionnaire can be found in Appendix 2, Attachment 2.2.

Vocabulary test

The receptive vocabulary test was a 48-item paper-based multiple-choice four-picture test which used words and images from the Peabody Picture Vocabulary Test (PPVT, Dunn & Dunn, 1997), used in a similar format by de Kraay (2016) and van Batenburg et al. (2019). A picture format test was chosen in order to minimize the amount of reading required of these pre-vocational pupils. The target words were verbs, nouns, adjectives, adverbs, and short phrases selected from levels 1–12 of the Peabody test. Pupils had to choose the picture which best corresponded to each English word or short phrase. Dutch-English cognates were removed in an attempt to eliminate a possible disadvantage for pupils with a non-Dutch home language, as Goriot et al. (2021) found that the effects of the Peabody test were confounded by the interplay of similarities between Dutch and English words. The lexical items in the adapted test were distributed across frequency bands based on the BNC/COCA headword lists (Nation, 2017): 29% from the 1,000-word level, 31% from the 2,000-word level, 11% from the 3,000-word level, and 29% from the 4,000-word level and above, including a small number of vocationally relevant terms considered appropriate for this pre-vocational sample. Due to copyright restrictions, the full adapted version of the test is not included in the appendix. A sample page of the test is given in Appendix 3, Attachment 3.1.

Grammar test

The lexico-grammar test (Attachment 3.2) was a cloze test (gap-fill with 36 target words removed) in the form of a short story, developed specifically for this research. The test is referred to as a *lexico-grammar* test to reflect its focus on grammatical forms as they are realized in context, including a small number of lexicalized constructions and fixed ex-

pressions. After this initial description, the term *grammar* is used as a shorthand to refer to outcomes on this test. As with the vocabulary test, this format was chosen in order to limit the amount of text that needed to be read and to contextualize the target words and forms, while supporting sustained engagement for pupils who may find extended decontextualized reading tasks challenging. The format was also intended to remain minimally accessible to pupils with very limited English proficiency, while still being sufficiently engaging and informative for pupils with higher proficiency levels. To support global text comprehension without providing lexical or grammatical cues for the gap items, the task included a small number of illustrative images related to the narrative content. The words which pupils needed to fill in were primarily context-based grammatical items (definite and indefinite articles, personal and demonstrative pronouns, verb tenses, morphological inflections, dependent prepositions) corresponding to target grammar in the chapters of Dutch and English course books for these beginning and lower-intermediate learners.

3.3F Procedure

Both the adapted PPVT and the grammar test were piloted with a class of pre-vocational year 1 pupils from a school not included in the study. In addition, the instruments were discussed in a brief talk-aloud session with the head researcher to ensure that all items were unambiguous and that test levels were feasible for the target population.

All tests and questionnaires were subsequently administered in the research pupils' own classrooms by the head researcher, assisted by a teacher at the school. The vocabulary test and the grammar test were administered on paper three times over the course of two academic years, 82 weeks in total. The first measurement took place in the first few months of the school year (September-October); for the year 1 pupils this meant that the starting language proficiency level of the CLIL and the non-CLIL pupils could reasonably be compared. The second measurement occurred 30 weeks later at the end of the same school year (May-June). The final measurement was 52 weeks later, in May-June of the following school year.

3.3G Analyses

As stated in the preceding section, the sample comprised 603 participants. Statistical power was estimated for detecting a small effect ($f = .15$; difference in growth between CLIL and non-CLIL pupils) at $p < .05$, assuming a correlation of .50 between the three repeated measurements, a sphericity correction ($\varepsilon = .80$), and an intra-class correlation of zero. Under these conditions, power exceeded .95 with a sample size of 135 cases (Buchner et al., 2019).

When the intra-class correlation was assumed to be .10 and the mean class size equaled 10, the effective sample size was 135 when the actual sample size equaled $N = N_{e}ff \times [1$

+ ($N_{(}$clus$_{)}$ − 1) × ρ] = 135 × 1.9 = 257 (Hox, 2010). Because the sample size in all analyses exceeded 257 participants, statistical power was considered sufficient**.**

For a regression analysis with $p < .05$ and 10 predictors, a sample of 170 participants was required to achieve a power of .95 for detecting an R^2 deviation from zero (Buchner et al., 2019). Assuming an intra-class correlation of .10, this again implied that a sample of approximately 500 students was sufficient**.** With an estimated intra-class correlation of .10, this resulted in an effective sample size of 500 / [1 + 19 × .10] = 172. Testing a small R^2 increase ($f = .15$) after adding four predictors in a regression model with 20 predictors at $p < .05$ yielded a power of .95 with 130 participants; thus, the available sample provided sufficient power for the planned analyses.

For the summed scores of the instruments used to measure vocabulary and grammar proficiency, internal consistency was assessed using Cronbach's alpha. For research at the group level, alpha values of at least .60 are generally considered acceptable (Bryman, 2008), with values of .70 regarded as satisfactory and .80 as good (Berthoud, 2000). As shown in Table 3.2, all summed scores for the vocabulary and grammar tests demonstrated high reliability, with alpha values ranging from .854 to .948 across the three measurement occasions.

Table 3.2: Cronbach's alphas for measurement instruments (N=sample size; SE=standard error; SD=standard deviation)

Variable	Nr of items	Alphas	N	Means (SE in parentheses)	SD
1st measurement vocabulary	48	.854	486	24.475 (0.345)	7.599
2nd measurement vocabulary	48	.880	490	27.269 (0.369)	8.161
3rd measurement vocabulary	48	.887	401	29.424 (0.408)	8.175
1st measurement grammar	36	.945	491	15.075 (0.422)	9.349
2nd measurement grammar	36	.948	484	18.240 (0.442)	9.727
3rd measurement grammar	36	.946	406	21.138 (0.474)	9.541

Because the data were nested within naturally occurring groups (classes and schools), the research questions were addressed using multilevel growth models. A multilevel growth model is comparable to a repeated-measures general linear model with a multilevel approach, implying different variance levels in the random part of the model. It is a regression model in which repeated measures of the dependent variable are predicted by independent variables, with random intercepts specified for each level (class and/or school) that significantly improved model fit, indicating a significant intra-class correlation at that level.

In the present dataset, repeated measures were nested within pupils, pupils within classes, and classes within schools. Consequently, the maximum number of levels was four, and the minimum was two (as repeated measures and pupil levels constitute the minimum required levels). The measurement occasion was included as a variable in the regression equation of the growth models. This variable was labeled *time* in the tables. The time of the first measurement was set to zero; all subsequent time values represented the number of weeks elapsed since the first measurement. Accordingly, *time* was measured in weeks after the initial measurement occasion.

Thus, all multilevel growth models included a repeated-measures level (with three measurements for vocabulary or grammar, respectively) nested within a participant level. Levels for class and school were added only when they significantly improved model fit. The significance of model fit improvement after adding a variance level was tested by means of the difference in deviance between nested models. This difference follows a chi-square distribution with a number of degrees of freedom equal to the difference in the number of parameters estimated. Fit improvement after adding a variance level to the model is considered significant at the 5% level when the chi-square is significant at the 10% level (Hox, 2010), since variances cannot be negative.

The significance of regression coefficients was evaluated in two ways: by means of the Wald test (coefficient divided by the standard error) and by verification of the difference in deviance between nested models (with and without the predictor[s] tested). Effect sizes were estimated by comparing variance components of nested models and calculating proportions of explained variance at each level.

The independent variable *pre-vocational level* differentiated between the represented pre-vocational levels of pupils within a class. Pupils had been allocated to one of four pre-vocational levels, ranging from more academically demanding to more practical tracks. However, pupils' level could change during secondary school, and classes could include pupils from different levels, referred to as *brugklassen* or *dakpanklassen* ('bridge' or 'overlapping roof tile' classes), particularly in the first two years. In addition, schools often created their own class combinations due to logistical constraints.

In the present study, such combinations of levels within classes and changes in class composition over the two-year period were observed**,** making it difficult to distinguish clearly between levels. Consequently, at best a distinction between *higher* and *lower* pre-vocational levels could be made. This resulted in a class-level variable with five different values indicating the average degree of pupil levels within a class. The variable *pre-vocational level* was treated as a continuous variable.

To answer the first research question, the variable indicating whether pupils attended CLIL or not was added to the growth model containing only *time*, *pre-vocational level*, and *cohort* (school year) as predictors. If the addition of the CLIL variable significantly improved model fit, this indicated a significant difference between CLIL and non-CLIL pupils at the outset in cohort year 1, as main effects in growth models pertain to the intercept.

Research question two was addressed by comparing a baseline growth model including *time*, *pre-vocational level*, *cohort*, *CLIL*, and all relevant two-way interactions—excluding the *CLIL* × *time* interaction—with a subsequent model in which the *CLIL* × *time* interaction (and any associated higher-order interactions, if significant) was added. A significant improvement in model fit indicated an effect of CLIL on growth. This procedure was applied to both vocabulary and grammar outcomes (see Appendix 3, Tables 3.3 and 3.4, models 7b and 8; complete models in Tables 3.5 and 3.7).

Research question three was addressed by inspecting the magnitude and significance of all possible three-way interactions between *time*, *CLIL*, *pre-vocational level*, and *cohort*. Where significant, a three-way interaction between *CLIL*, *time*, and either *cohort* or *pre-vocational level* implied moderating effects of cohort and/or pre-vocational level on the effect of CLIL on growth in language proficiency.

To answer the fourth research question, interaction effects between pupils' background characteristics, growth in proficiency, and CLIL versus non-CLIL stream were tested**.** These individual characteristics included gender, general achievement test score (*Cito*), home language environment, years of primary school English instruction, and frequency (lessons per week) of primary school English. For each moderator variable and each measurement instrument, the model including the three-way interaction between *time*, *CLIL*, and the moderator variable was compared with the corresponding model without this interaction. Where the addition of the three-way interaction resulted in a significantly improved model fit, the proportions of explained variance attributable to this moderator effect were calculated**.** All continuous variables were grand-mean centered prior to inclusion in the regression analyses.

3.4 Results: vocabulary and grammar development

For clarity's sake, the results of the four research questions are given first for vocabulary, and then for grammar.

3.4A Vocabulary results

Before testing the effect of CLIL, the appropriate multilevel structure was established. In the regression analyses concerning English vocabulary, in addition to repeated measures and pupil level, a class level and a school level were required (see models 1–3, Table 3.5). The initial level (research question 1) of English vocabulary in cohort 1, at the start of the pre-vocational program and the start of CLIL, was significantly higher for the CLIL pupils than the non-CLIL pupils (model 6a, Table 3; χ^2=19.427; df=1, p<.001, a difference of 5.177 points). To estimate the effect of CLIL on growth in vocabulary after correcting for pre-vocational level and cohort, a model of reference was created with the predictors time, pre-vocational level and the two dummies for cohorts 2 and 3 as well as all two-way interactions between these variables. Subsequently a new model was created by adding the interaction term of time and CLIL. Comparing these last two models, growth in vocabulary is significantly larger (χ^2 = 5.571; df = 1; p < .01; 0.016 points per week) for CLIL pupils than for non-CLIL pupils (model 7 compared to model 6b; see Table 3.3). Participation in CLIL explains 0.8% of differences in vocabulary growth and 6.9% of differences in vocabulary between classes at the start (see Table 3.5).

Three- and four-way interactions between CLIL, time, and other variables were non-significant in the exploratory models (see Table 3.5). As a result, no moderator effects of pre-vocational level or cohort were retained, and these interactions are not included in the final models reported in Tables 3.3 and 3.5. Figures 3.1, 3.2, and 3.3 show the vocabulary growth per cohort of CLIL and non-CLIL pupils in each of the five pre-vocational levels, using the values from the multiple regression analyses for the model of best fit. Higher pre-vocational levels show greater growth for both CLIL and non-CLIL, and the positive effect of CLIL on vocabulary growth is approximately the same for all pre-vocational levels (research question 3).

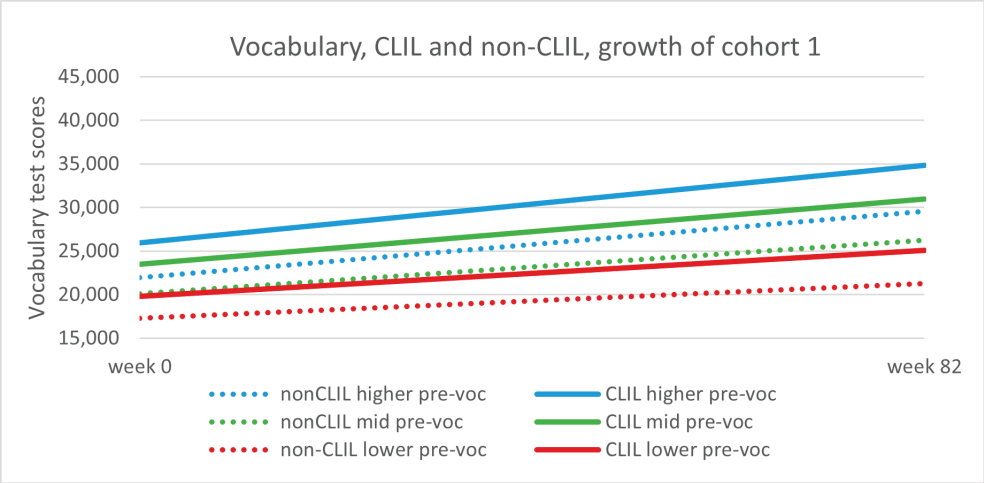


Figure 3.1: Vocabulary growth of cohort 1, different pre-vocational levels

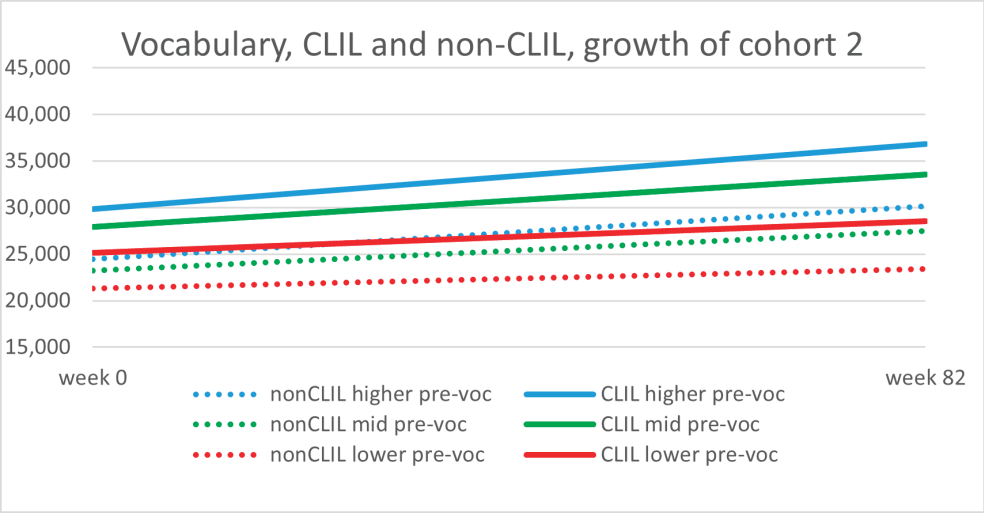


Figure 3.2: Vocabulary growth of cohort 2, different pre-vocational levels

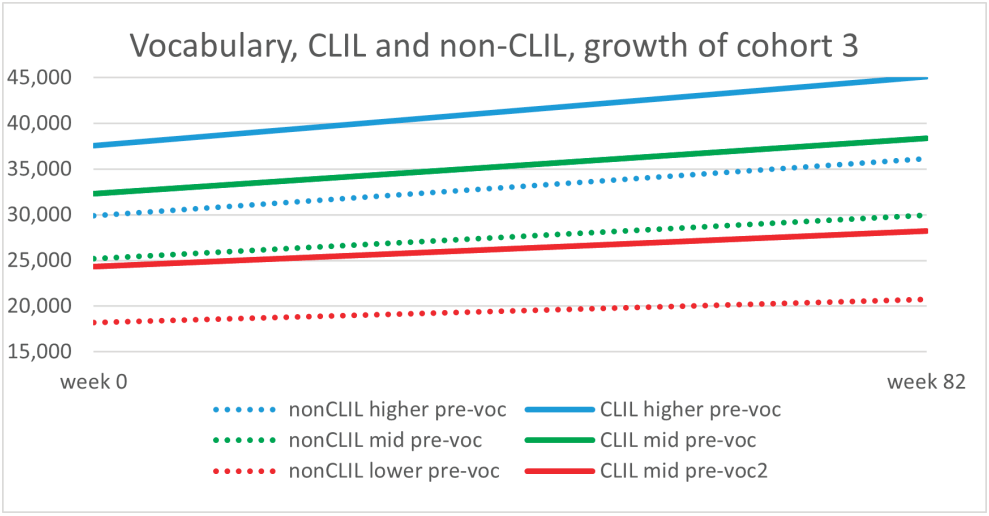


Figure 3.3: Vocabulary growth of cohort 3, different pre-vocational levels

As described in the Analysis section, we have, in addition to examining cohort and pre-vocational level, also examined whether the effects of CLIL on vocabulary are moderated by the pupil characteristics (research question 4). Table 3.6 (Appendix 3) shows that after adding the three-way interaction terms (even-numbered models), none of the pupil characteristics significantly moderates the effect of CLIL on vocabulary growth. In all cases, the relevant three-way interactions (time × moderator × non-CLIL) are non-significant.

3.4B Grammar results

The second set of results concerns grammar (Table 3.4). As with vocabulary, all four variance levels are required in the random part of the regression model. Regarding the first research question, the results for the grammar test are comparable: cohort 1 CLIL pupils start significantly higher than their non-CLIL peers (Table 3.4, model 6b, and Table 3.8; $\chi^2 = 18.040$, df = 1, $p < .001$, a difference of 6.821 points). Research question 2 is addressed by adding the CLIL × time interaction to the two-way interaction model (model 7 compared to model 6b; see Table 3.4 and Table 3.8). Across all years, pupils in the CLIL groups show significantly larger learning gains after correcting for pre-vocational level and cohort, as well as differences in growth due to these variables (0.011 points per week, one-tailed $p < .05$). The effect, however, is not significant when testing the deviance difference between both nested models and only 1-sided significant when we use the Wald test. Furthermore, the effect size is very small (0.5% explained variance of the growth).

The final set of three-way interactions added to model 7b (resulting in model 8), specifically the time × cohort2 × non-CLIL and time × cohort3 × non-CLIL interactions, addresses research question 3 by testing whether the effect of CLIL on grammar growth differs across cohorts. These results indicate that the positive effect of CLIL on grammar

development is strongest for the year 1 cohort, while no significant moderation by pre-vocational level was observed. The three-way interaction between cohort 2, time, and non-CLIL has a significant positive regression coefficient (.035, $p < .05$), as does the interaction between cohort 3, time, and non-CLIL (.072, $p < .001$). These findings indicate that the effect of CLIL on grammar growth differs across cohorts. In contrast, the three-way interaction between pre-vocational level, time, and non-CLIL has a negative but non-significant regression coefficient (–.012, $p > .05$), indicating that no reliable moderation effect of pre-vocational level on the CLIL–non-CLIL difference in grammar growth was found.

Comparing these two models shows that the addition of the three three-way interaction terms—two cohort-based interactions (time × cohort2 × non-CLIL and time × cohort3 × non-CLIL) and one interaction involving pre-vocational level—together with the CLIL × time effect, significantly improves model fit ($\chi^2 = 27.888$; df = 3; $p < .001$). Although only the cohort-based three-way interactions reached statistical significance, the inclusion of these interaction terms collectively explains an additional 3.68% of the variance in grammar growth (model 8 compared to model 7b; see Table 3.4 and Table 3.8). This indicates that CLIL accounts for 3.68% of the overall differences in growth in grammatical proficiency.

Figures 3.4, 3.5, and 3.6 show the results of the grammar test for the CLIL and non-CLIL pupils for different values of the pre-vocational level variable, using the values from the multiple regression analyses for the best-fitting model.

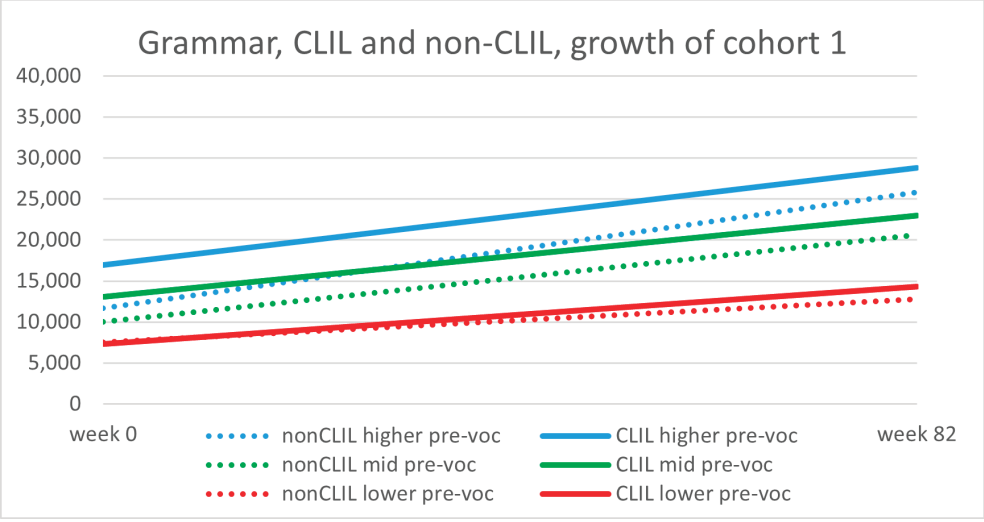


Figure 3.4: Grammar growth of cohort 1, different pre-vocational levels

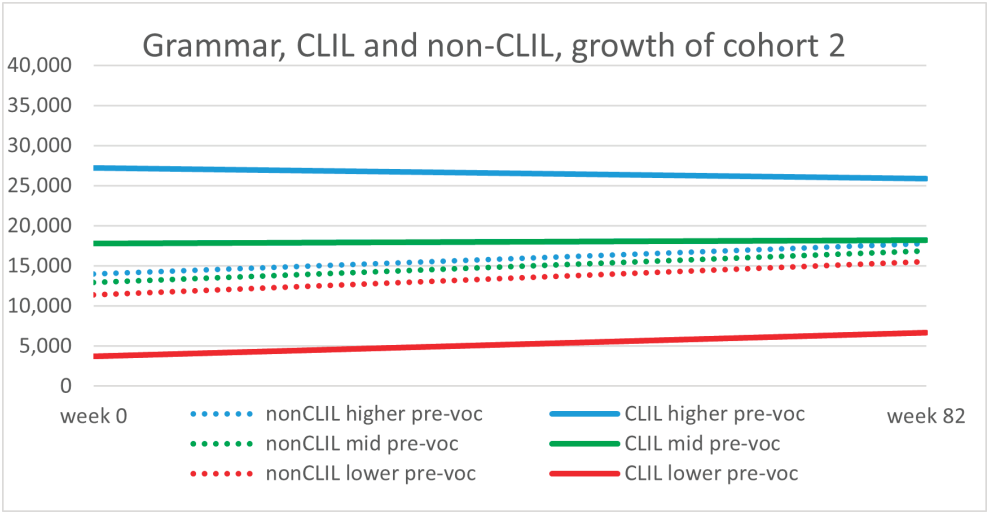


Figure 3.5: Grammar growth of cohort 2, different pre-vocational levels

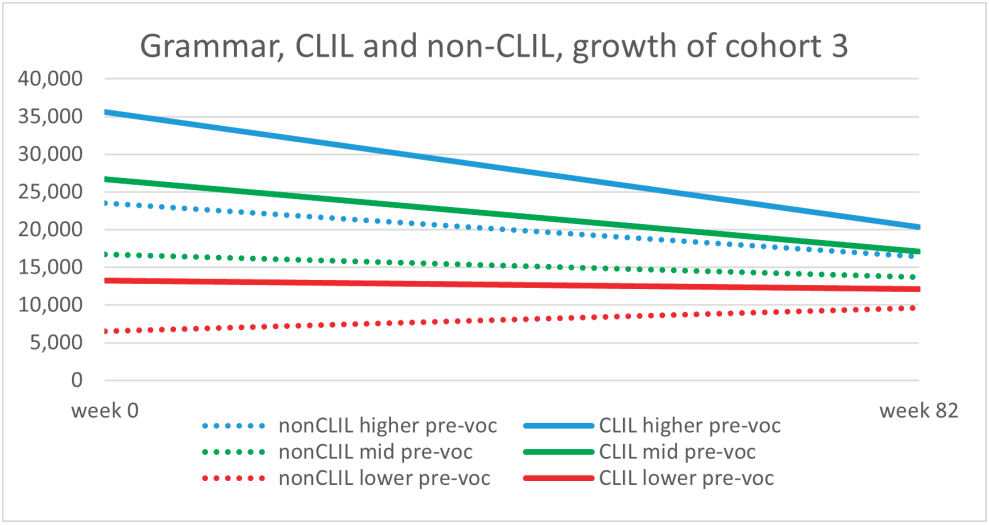


Figure 3.6: Grammar growth of cohort 3, different pre-vocational levels

Table 3.8, which presents the moderator effects of pupils' background characteristics, shows that after adding the three-way interaction between time, CLIL, and each moderator variable, no significant moderation effects are found. This indicates that the effect of CLIL on grammar growth is not moderated by individual pupil characteristics over and above the cohort effects discussed earlier**,** such as pre-vocational level, gender, prior achievement, home language, or exposure to English in primary school. In other words, these personal background variables do not significantly affect the mean difference in grammar growth between CLIL and non-CLIL pupils.

3.4C Summary of key findings

Across both vocabulary and grammar, CLIL pupils outperformed their non-CLIL peers on initial tests and demonstrated small but consistent growth advantages over time. For receptive vocabulary, the positive effect of CLIL on growth was broadly similar across all pre-vocational levels, whereas for grammar the CLIL advantage showed a non-significant tendency to be slightly greater among pupils in the lower-level groups. The largest gains for grammar were observed in the CLIL year 1 cohort, whose progress exceeded that of the older groups. While this pattern may indicate that the impact of CLIL is strongest in the first years of the program, it could also reflect contextual factors such as increased teacher experience or greater emphasis on target language exposure at that stage. None of the learner variables (gender, prior academic achievement, home language, or prior English instruction) significantly moderated the effects of CLIL for either vocabulary or grammar growth.

3.5 Discussion, limitations and recommendations

3.5A Interpretation of findings

Previous research has shown that pre-vocational CLIL pupils in the Netherlands develop a more positive attitude towards learning English than their non-CLIL peers (Denman et al., 2018; see also Chapter 2 of this dissertation). The results of the current study show a more positive growth in two aspects of target language ability as well.

The descriptive statistics show that the CLIL and non-CLIL groups were reasonably equally-matched, as none of the background variables is significantly related to participation in CLIL. However, there was a difference in the English-language proficiency between the two groups from the outset.

This is also the answer to the first research question, concerning the initial level of pupils in the CLIL and non-CLIL stream at the start of CLIL in cohort 1 (year 1). The CLIL pupils start with significantly higher scores on the pretests for both vocabulary and grammar (both $p<.001$). Considering the lack of any school-based selection procedure, this seems to indicate that most pupils who self-selectively choose for the CLIL stream at the end of primary school do so with confidence about their relative level of English.

In terms of longitudinal growth (research question two), the multilevel growth models show that the CLIL pupils show significantly more growth in both vocabulary and grammar than the non-CLIL pupils, albeit with small effect sizes, 0.08% for vocabulary and 3.68% for grammar, thus confirming previous claims that CLIL favorably affects target language receptive skills such as vocabulary (Dalton-Puffer, 2017; Ruiz de Zarobe, 2015), and aspects of grammar (Dalton-Puffer, 2008b; Pérez Vidal & Roquet, 2015). It is possible

that the CLIL pupils' higher grammar growth could be attributed to low initial proficiency, whereby more input in and exposure to the target language input contributed to a more rapid uptake at a beginning level. The small effect sizes here can also partly be explained by the fact that many unmeasured variables can also influence language proficiency growth, such as motivation, IQ, self-confidence, teacher input, or classroom dynamics.

Comparing the different cohorts (research question three), we can see a differential effect on growth. The growth in both vocabulary and grammar for cohort 1 (from the beginning of year 1 through the end of year 2) is significantly greater than for cohort 2 or cohort 3, with a significant advantage for the CLIL pupils in cohort 1. In fact, the cohort 1 CLIL had outscored the cohort 3 non-CLIL pupils after 82 weeks on the vocabulary and grammar tests, despite two years' additional English lessons for the older non-CLIL pupils. This corresponds with results of research which has attempted to compensate for the extra hours of target language exposure in CLIL programs by matching those pupils with older non-CLIL pupils (Jiménez Catalán & Pilar Augustín Llach, 2017; Pérez-Vidal & Roquet, 2015). It seems likely that this greater growth in the first two years is due to the greater number of CLIL subjects and larger number of hours of exposure in the first two years. The subsequent reduction in the number of CLIL subjects and exposure hours seems to correlate with a reduction in growth, as cohorts 2 and 3 continued to show significantly more vocabulary growth than the same non-CLIL cohorts, but the growth rate declined. For grammar, however, the growth of the CLIL pupils in cohort 2 stagnated, and the CLIL and the non-CLIL pupils in cohort 3 actually showed negative growth: lower scores at the end of the two-year study.

Although these results correspond to those found at higher school levels (Goris et al., 2013; Rumlich, 2017; Verspoor et al., 2013) this drop in both the CLIL and non-CLIL grammar scores of the year 3 cohort does not seem to have an obvious explanation: since the non-CLIL group shows a similar decline, the lack of growth cannot logically be attributed to the reduction in CLIL provision and exposure hours. Although the CLIL year 3 cohort was receiving less English-language input than their lower-year CLIL peers, there is a strong emphasis on grammar-based instruction in English lessons in the Netherlands (de Kraay, 2016; Jeffery & van Beuningen, 2020), even among the younger generation of teachers (Graus & Coppen, 2017), and one might assume that this would at least serve to maintain the level of grammatical knowledge. Possible explanations could include an increased focus on final English reading comprehension-based exams, pre-exam anxiety and concerns about exams in the non-CLIL language (c.f. Brevik & Moe, 2012), careless errors due to a perception that the test is too easy, or lack of motivation and interest (c.f. Doiz et al., 2014a).

There is no differential effect of the grand mean centered 'pre-vocational level' predictor on the effect of CLIL (also research question three) for vocabulary. Pupils in classes with

higher pre-vocational levels show significantly faster growth in vocabulary, but there is no significant growth difference between the CLIL and non-CLIL groups depending on their vocational level. Conversely, for grammar, classes with pupils in less academic pre-vocational levels show more effect of CLIL on grammar growth. In other words, CLIL appears to be more effective for lower pre-vocational pupils' grammar proficiency development than for pupils in higher pre-vocational classes. Interestingly, the aforementioned decline in grammar scores for cohort year 3 is most extreme for pupils in classes with pupils in higher pre-vocational levels, who also started with much higher scores, and almost negligible for pupils in classes with lower levels, who started with much lower scores but only decreased minimally.

3.5B Moderating variables and differential effects

The effect of the five pupil background characteristics as moderator variables—gender, general academic achievement score (Cito), home language background, and duration and frequency of primary English—proved non-significant. Other studies have indicated that CLIL may even the playing field so that boys overcome a traditional lag in foreign language learning; in this study this was not the case, as both CLIL and non-CLIL boys outperformed the girls, with no additional advantage for the CLIL boys. The advantage for boys may be related to their relatively higher participation in online and video gaming in English, as found by de Kraay (2016). As Cito score is directly related to the pre-vocational level to which pupils are assigned, any differential effect of Cito score may already have been subsumed in the pre-vocational level categorization. Regarding home language, although there was no significant effect of this on CLIL gains, neither were there losses: our results do not show that a minority-language background seems to have a facilitative or detrimental effect on vocabulary and grammar scores in CLIL participation, which is particularly relevant given the high percentage of minority-language and migration-background pupils in this study and in Dutch pre-vocational education in general. The duration and frequency of primary school English, while showing a significant effect on vocabulary and grammar gains for all pupils, had no significant effect on the growth of CLIL pupils.

3.5C Limitations and recommendations for future research

There are several limitations to this study. First, the quasi-experimental design meant that there was no random allocation into groups, implying that the internal validity is less than in a randomized controlled trial. Secondly, the two-year, three-cohort design of the study meant that different cohorts were measured simultaneously and predictions were made about growth over four years, although each cohort was only measured for two years. It is possible that each cohort might have had different outcomes if they had been followed for all four years; our comparisons over cohorts are based on different individuals. Thirdly, there was the largest sample size and thus the greatest power for cohort year 1 (both CLIL and non-CLIL) and smaller numbers and a weaker power for cohorts year

2 and (especially) 3. This is due partly to the novelty of CLIL at pre-vocational programs in the Netherlands: some programs had simply not existed long enough for there to be pupils in year 3. Lastly, because pre-vocational CLIL was still at a nascent stage at the time of the study, no pre-vocational school had yet achieved a national CLIL accreditation, and although some schools in the study benefitted from the experience of CLIL programs at their higher streams, CLIL implementation may not yet have been optimal at all participating schools. It is recommended that a similar research study be carried out at accredited pre-vocational schools in the near future to see if the results are similar or even more conclusive.

A few additional recommendations can be made. This study examined the development of pre-vocational pupils' vocabulary and grammar over time, comparing CLIL and non-CLIL groups. A similar study could compare the same groups' productive language development in the areas of writing and speaking. Additionally, since the specifics of curricular implementation of CLIL pedagogy were not a focus of this study, a further recommendation is for more research to examine the actual implementation of CLIL didactics and pedagogy on developing language skills, and compare language proficiency development of CLIL and non-CLIL groups regarding the effects of specific pedagogical approaches. The instruments used to measure vocabulary and grammar could also be employed in another sample, or a more standardized language proficiency test could be administered to a similar population to compare results. Finally, research on bilingual education and the CLIL approach for non-elite secondary pupil populations is sorely needed in other regional contexts as well. Kontio and Sylvén (2015) note the paucity of research on CLIL in vocational contexts, and Küppers and Trautmann state the case for "...a focus which takes into account the effects of CLIL on and the impact of CLIL on non-selected students, non-CLIL classes, exclusion processes, selection, social inequality..." (2013, p. 294). Only with a broader range of such research can informed conclusions about the suitability of CLIL for lower-achieving learners be drawn, and only with more CLIL opportunities for these learners can such research be conducted.

3.6. Conclusion

The main research question in this study was whether pre-vocational pupils, including those with lower scholastic aptitude, can benefit from a CLIL-based bilingual educational program in terms of enhanced English-language proficiency gains in vocabulary and grammar, compared to their peers in mainstream, non-CLIL programs with regular foreign language instruction. A total of 603 pre-vocational secondary pupils in the Netherlands were studied over the course of two school years, across three parallel cohorts spanning pupils in different years and sub-levels of the pre-vocational program. There were no admission criteria for participation, as enrollment in CLIL was based on self-selection. Over

one-third of the pupils had a migration background and reported a home language other than Dutch, the default language of instruction.

The measurement instruments included a receptive vocabulary test and a cloze grammar test. Multilevel regression analyses showed that, despite the absence of formal selection criteria, pupils who chose the CLIL stream already demonstrated significantly higher levels of English vocabulary and grammar at baseline. Over the two-year study period, CLIL pupils not only maintained this initial advantage but also showed greater growth in both vocabulary and grammar, albeit with small effect sizes (0.8% for vocabulary and 3.68% for grammar). The strongest growth in these skills occurred during the first two years of pre-vocational schooling (cohort year 1), which may be related to a reduction in the number of CLIL subjects and associated instructional time from year 3 onwards. The results for cohort year 3 showed a decline in grammar proficiency for both CLIL and non-CLIL pupils.

Although growth among CLIL pupils in lower pre-vocational level classes was generally slower than among those in higher-level classes, these pupils nonetheless tended to show greater language proficiency gains than their non-CLIL peers. Taken within the context of this study's early-stage implementation of pre-vocational CLIL, the findings suggest that CLIL can contribute to increased English language proficiency for lower-achieving pupils, with no evidence that its effects are moderated by pupils' home language background or prior exposure to English in primary education. To optimize the benefits of CLIL for this population, the findings support maintaining, rather than reducing, curriculum time for CLIL as pupils progress through the program.



Inclusive CLIL: Pre-vocational pupils' target language writing and speaking proficiency, fluency, and Willingness to Communicate

This chapter is based on a previously published article:
Denman, J., van Schooten, E., & de Graaff, R. (2022). Inclusive CLIL: Pre-vocational pupils' target language oral proficiency, fluency, and Willingness to Communicate. *AILA Review, 35*(2), 321–350. https://doi.org/10.1075/aila.21021.den

Note.
*Sections of this chapter reporting on pupils' speaking proficiency, speaking fluency, and Willingness to Communicate (WtC) were previously published in the article cited above. The present chapter extends that work by including the writing component of the study and incorporates additional analyses and revised framing. The terminology used in the dissertation refers to the educational approach as Content and Language Integrated Learning (CLIL)**;** the published journal article employs the same concept.*

Abstract

Bilingual education using a Content and Language Integrated Learning (CLIL) approach is widespread in secondary education throughout Europe and also found further afield. In many contexts CLIL seems to select or attract the more able and more academically-inclined pupils, or only be available to pupils in higher academic secondary streams. Positive effects of CLIL for target language proficiency development may be due in part to this cognitive or academic selection effect. Can the target language skills of pupils with lower scholastic attainment—a group which, in some contexts, has less access to CLIL programs—also benefit from the CLIL approach?

The current two-year longitudinal quasi-experimental research, part of a larger study, focused on the development of oral and written proficiency skills of three cohorts of 603 pre-vocational pupils in 25 classes in the Netherlands in both CLIL and non-CLIL programs. Pre-vocational secondary education in the Netherlands serves approximately fifty percent of the total pupil population, including a large percentage with a minority-language background, and consists of the least academic streams. Despite the lack of explicit school-based selection procedures for pre-vocational pupils' participation in CLIL, there were significant differences in favor of the CLIL groups in the initial levels of English oral proficiency and fluency, and willingness to communicate orally in a familiar situation, but not for English writing proficiency, writing fluency, or willingness to communicate orally in an unfamiliar situation. The CLIL pupils showed significantly more growth than the non-CLIL control group in writing proficiency and fluency as well as speaking proficiency, but not for speaking fluency or willingness to communicate. This positive result for the CLIL group did not appear to be moderated by pupil background variables, except for writing proficiency and fluency, which showed a small moderation effect of class year. Despite the small effect sizes found, these results are encouraging for the further development of CLIL provision for pre-vocational pupils in the Netherlands and elsewhere, and indicate that despite the cognitive challenges, the CLIL approach can have a positive effect on foreign language proficiency of pupils in less academic educational streams.

Keywords: bilingual education; CLIL; content and language integrated learning; inclusion; pre-vocational secondary education; oral proficiency; speaking; willingness to communicate

4.1 Introduction

The Content and Language Integrated Learning (CLIL) approach to bilingual education, in which some school subjects are taught through a second or foreign language and attention is paid to both subject content and the target language, was envisioned as "a pragmatic European solution to a European need" (Marsh, 2002, p. 11), with the goals of increasing foreign language competence so as to enable more mobility and cultural understanding across the European Union (Marsh, 2012). Supported by several Council of Europe initiatives, the CLIL approach spread rapidly throughout Europe, particularly in secondary education (Nikula, 2017a). Several decades later, various forms of CLIL provision are now part of educational systems in nearly all European countries, with a wide range of target languages but most commonly with English as the target language (EECEA, 2017; Dalton-Puffer, 2011). European CLIL programs share certain core characteristics: the target language is a foreign language rather than a second language; the CLIL teachers are themselves generally non-native speakers of the target language; CLIL lessons are school subject lessons, with additional foreign (target) language lessons; the CLIL lessons generally comprise less than 50% of the school curriculum (Dalton-Puffer, 2011). However, particularly in some European contexts, there is concern about a perception that CLIL only "'works' in 'elite' contexts, i.e. in private, urban schools with socio-economically and socio-culturally privileged children" (Pérez Cañado, 2020, p. 7) and about whether "rather than increasing the equality of opportunity, CLIL in certain contexts is subtly selecting students out" (Bruton, 2013, p. 593). In this light, increasingly more attention is being paid to inclusion and diversity in CLIL, such as in the six-country ADiBE research project (ADiBE Project, n.d.) which aims to make CLIL accessible to all learners, regardless of background or ability.

4.1A CLIL and productive language skills

One of the core questions is to what extent CLIL contributes to productive L2 development. As secondary school CLIL programs in Europe have often been offered only in the more academic school types (Feddermann et al., 2021) and positive CLIL L2 language development results have been attributed partly to the selection of the brightest pupils (Küppers & Trautmann, 2013), another question is whether CLIL can also benefit the development of those productive skills in learners with lower scholastic ability in a context with no explicit selection criteria.

Despite the lower amount of exposure to the L2 and the less ambitious language-learning goals in most CLIL contexts than in Canadian immersion contexts (for further comparison of immersion and CLIL, see Cenoz et al., 2014), CLIL still increases not only the contact time, the amount of exposure, and opportunities for L2 use compared to mainstream foreign language education, but also the quality of the interaction in the target language (Escobar Urmeneta, 2019).

In the case of the productive skill of L2 writing in CLIL, a survey by Dalton-Puffer (2008b) reports that writing proficiency is generally less positively affected than speaking. Studies comparing CLIL and non-CLIL pupils' results on writing tests show a mixed picture of results. CLIL pupils have been found to have a wider lexical range (Ackerl, 2007; Jexenflicker & Dalton-Puffer, 2010; Navés & Victori, 2010; Ruiz de Zarobe, 2010), display more grammatical accuracy (Roquet & Pérez Vidal, 2015), and show more written fluency as measured by number of words (Navés & Victori, 2010). Comparisons of general writing proficiency not focusing on specific sub-skills have shown that CLIL groups outperform non-CLIL groups (Alonso et al., 2008; Lasagabaster, 2008; Lorenzo et al., 2010; San Isidro, 2010; Verspoor et al., 2010; Verspoor et al., 2013; Zydatiß, 2012). On the other hand, other studies have found no significant advantage for CLIL pupils in the area of written vocabulary (Saladrigues & Llanes, 2010), accuracy (Nieto Moreno de Diezmas, 2016; Ruiz de Zarobe, 2010), fluency as measured by words per minute (Pérez Vidal & Roquet, 2015) or general writing proficiency (Saladrigues & Llanes, 2010).

The lack of conclusive results for writing may be due to the fact that learning at school is mainly an oral experience, and that CLIL provides fewer opportunities for writing than speaking (Surmont et al., 2014). It may also be the case that the productive skills, especially writing, need more time for CLIL to show a positive impact, as the receptive skills generally show larger gains (Dalton-Puffer, 2007).

Research results comparing the L2 speaking skills of CLIL and non-CLIL pupils generally report positive oral proficiency results for CLIL learners. In a survey of CLIL program L2 outcomes, Dalton-Puffer (2011, 2017) reports that the most noticeable advantage of CLIL pupils over their non-CLIL peers is in oral production, particularly regarding fluency, quantity, and risk-taking. Significantly higher results for CLIL pupils' general L2 speaking proficiency have been reported (Admiraal et al., 2006; Lasagabaster, 2008; Lorenzo et al., 2010; Martínez Agudo, 2019; Nieto Moreno de Diezmas, 2016), as well as for the oral proficiency sub-skills of grammar, lexical range, fluency, and pronunciation (Madrid & Barrios, 2018; Pérez Cañado, 2018; Ruiz de Zarobe, 2008). CLIL pupils also showed a significantly higher rate of speaking fluency as measured by number of words or words per minute (Dalton-Puffer et al., 2008; Juan, 2010; Zydatiß, 2007). On the other hand, some mixed results have been found after taking certain variables into consideration. Academic ability seems to show a differential effect: while CLIL pupils with average-to-high academic aptitude outscored their non-CLIL peers, those with lower scholastic attainment did not, struggling particularly with oral proficiency due perhaps to the cognitive challenges of CLIL (Mewald, 2007); we will return to the issue of academic aptitude in section 4.2. Time is also a factor; no significant oral fluency advantage for CLIL pupils was found after one year (Merino & Lasagabaster, 2018), or two years (Rallo Fabra & Jacob, 2015); these authors raise the question whether this is enough time for a significant advantage to emerge, as generally the development of L2 productive skills lag behind that of the receptive skills

(Pérez Cañado, 2018; Rallo Fabra & Jacob, 2015). Not only time, but also occasion and communicative need are considered crucial for creating a stimulating environment for developing productive L2 skills (Hüttner & Smit, 2014). Compared to mainstream foreign language classrooms, the CLIL approach is expected to offer more opportunities for L2 interaction and authentic communication (Dalton-Puffer et al., 2010b; Pérez-Vidal, 2009).

4.1B Willingness to Communicate (WtC)

For L2 interaction and communication to occur, however, learners must be *willing to use* the L2. Because more interaction influences the amount and frequency of communication, MacIntyre et al. (1998) propose that the development of 'willingness to communicate' (WtC) is the primary goal of foreign language instruction, as it is suggested to be "the most immediate determinant of L2 use" (Clément et al., 2003, p. 191). Not only are higher-WtC learners more likely to use the L2 more frequently; they are also more inclined to do so independently and thus help create a more active communicative classroom atmosphere, and they may extend their learning opportunities more readily to outside the classroom (Kang, 2005).

The WtC construct in L2 learning (MacIntyre et al., 1998) has been studied in both immersion and CLIL contexts. WtC has been found to be situational, and stronger inside than outside the classroom (MacDonald et al., 2003). Individual variables such as gender, age, and prior L2 experience can influence a learner's WtC (MacIntyre et al., 2002, 2003). It has been postulated that CLIL provision can increase WtC, and higher WtC helps raise L2 proficiency (Menezes & Juan-Garau, 2015). However, studies of WtC in CLIL do not show consistent results. In a Flemish context, interviews with teachers and parents reveal that they noticed CLIL pupils' increased willingness to communicate in the L2, including that of the less proficient pupils (Simons et al., 2019). A significant correlation between WtC and L2 proficiency was found by Menezes and Juan-Garau (2015), with the CLIL pupils scoring higher on both measures than their non-CLIL peers. In a longitudinal study in Germany, Italy and the Netherlands (Goris et al., 2013), CLIL pupils were found to have higher WtC scores than their non-CLIL peers already at the start of CLIL provision. There was no significant increase over time for the German or Italian groups and although the Dutch CLIL and non-CLIL pupils significantly increased their WtC, there was no significant growth advantage for the CLIL group over time (Goris et al., 2019). Lialikhova (2018) found an increase in WtC for mid- and high-achieving pupils, but no change for the lower achievers, who not only had the lowest level of WtC but also struggled with anxiety, low oral fluency, and the communicative demands of CLIL. From these studies we can conclude that it is essential to pay particular attention to the WtC of lower-attaining learners.

4.1C Individual differences and selectivity

Comparisons and evaluations of the various empirical results for writing and speaking are complicated because not all researchers have used an experimental design or stringently

controlled for initial levels (Bruton, 2011a; Verspoor et al., 2015) or appropriately controlled for selection effects (Goris et al., 2022; Paran, 2013; Pérez Cañado, 2020; Piesche et al., 2016). Additionally, individual differences between types of learners may also show differential effects, although these may be highly dependent on context. For instance, gender inequalities in language learning may skew results, as girls may outperform boys in mainstream foreign language learning but less so in CLIL (Lahuerta Martínez, 2015; Merisuo-Storm, 2007; San Isidro, 2010). Migration background, and particularly a minority home language should also be taken into consideration; although some CLIL studies in the German context reveal no significant differences in L2 proficiency growth between migration background pupils and their L1 German peers (Dallinger et al., 2016; Schwab et al., 2014), it is important to explore whether the CLIL approach might pose a risk to school success for less academically-inclined pupils with a minority home language, especially as teachers may subconsciously apply their own socio-cultural biases in the classroom (van den Bergh et al., 2010). Conversely, it is important to see if CLIL can benefit these pupils by reducing possible disparities in L2 attainment between majority- and minority-language pupils. In short, to isolate the effects of CLIL, it is important to control for a possible selection effect at the outset of CLIL provision, as well as for certain potentially intervening variables and learning prerequisites (Dallinger et al., 2018)—such as gender, home language, standard achievement test scores, and prior L2 instruction—which might moderate the results.

4.2 CLIL selectivity and learners with lower scholastic attainment

Research has shown CLIL pupils' advantage from the outset of CLIL over their non-CLIL peers in various target language skills (Admiraal et al., 2006; Alonso et al., 2008; Broca, 2016; Juan, 2010; Merino & Lasagabaster, 2018), implying that CLIL is more often followed by abler students and resulting in 'educational creaming' (Rumlich, 2017, p. 115). This selection effect may be at least partly responsible for the positive effects of CLIL, which could be caused by differences in motivation and predisposition as well as higher L2 ability from the start (Garcia López & Bruton, 2013; Küppers & Trautmann, 2013).

4.2A Evidence from immersion and CLIL contexts

Research on Canadian immersion has shown that pupils with low academic ability or low literacy development can benefit from bilingual immersion (Cummins, 1984a), outperforming control group peers in L2 proficiency, first-language development, and subject matter (Genesee, 2004; Genesee & Fortune, 2014). Considering Canadian immersion studies among pupils with lower IQ, learning disabilities, lower socio-economic status, and other at-risk factors, Genesee (1987, 2004) concludes that there is no evidence that these learners are disadvantaged in an immersion setting, and that they can profit from

bilingual education. However, a differential effect of pupils' academic ability has been found in immersion pupils' L2 proficiency development in all target language skills, attributed to the high cognitive demands of learning content through a foreign language (Genesee, 2004). In other words, low academic ability pupils benefitted from immersion, but not to the extent that the higher academic ability pupils did. There do not, however, seem to be many immersion studies addressing this issue.

4.2B Lower-attaining learners in CLIL research

Similarly, there is a distinct lack of empirical evidence in CLIL research regarding CLIL learners with lower scholastic attainment. Few CLIL programs (and, consequently, few studies) have expressly included learners with lower scholastic attainment, whether in heterogeneous or homogeneous groups. There are notable exceptions, such as in Andalucia, Spain (Lorenzo et al., 2021), Queensland and Victoria, Australia (Smala, 2022), the UK (Coyle & Meyer, 2021), and the aforementioned European ADiBE Project. Although pre-vocational CLIL pupils have shown more growth in positive attitudes towards learning English than their non-CLIL peers (Denman et al., 2018; see also Chapter 2 of this dissertation), very little is known about what effect CLIL might have on the L2 proficiency of pupils with average to below-average attainment. The positive results for CLIL pupils regarding target language proficiency, mentioned at the beginning of this section, may partly be attributed to the participation of selected, more academically gifted and more motivated learners in CLIL programs (Broca, 2016; Bruton, 2011b; Dallinger et al., 2016; de Bot & Maljers, 2009; Feddermann et al., 2021). The question is whether, after requisite controlling for initial proficiency level, lower-attaining CLIL pupils also develop higher L2 proficiency in writing and speaking skills than their non-CLIL peers.

The few studies that address this question yield mixed results. A study of modular CLIL in a small class in a German *Hauptschule* (the lowest of a streamed secondary education system) showed that over two years the greatest gains in oral proficiency were made in the first year, with no difference between the L2 development of the German L1 speakers and their migration-background peers (Schwab, 2013). In a one-year Belgian study, teachers report an increase in the speaking skills of weaker pupils and an enhanced willingness to participate (Simons et al., 2019). Other results, however, are less promising. In a longitudinal study of mixed-ability CLIL learners at CEFR A1–B1 level, Escobar Urmeneta (2004) found that the lower academic ability pupils' L2 speaking skills were of poorer quality than those of their higher-ability classroom peers, and in fact hardly improved, although these pupils did increase their self-confidence and attitude. Mewald (2007) found that while average and above-average learners benefitted from CLIL, the lowest-ability CLIL groups scored lower in L2 proficiency than their non-CLIL peers and struggled with aspects of the L2, especially oral proficiency. Gierlinger (2017) notes that teachers reported being unwilling to even attempt the CLIL approach with their lower-ability learners. It is no surprise that research results are sparse.

4.3 The present study: research context and questions

4.3A Research context and rationale

The highly-streamed Dutch educational system offers a structure within which to compare the L2 language development of lower-attaining CLIL and non-CLIL pupils' L2 language development. This streaming system sifts pupils at the end of primary education into various types of secondary schools based on standardized achievement tests and primary school teacher assessments. Over 50% of the total pupil population is directed into one of the four pre-vocational subtracks, with a disproportionate number of pupils with a first- or second-generation migration background: 62% of all secondary school pupils with a non-western migration background are allocated to pre-vocational education, and particularly to the two least academic pre-vocational subtracks (Nederlands Jeugdinstituut, 2021). Research, both international and in the Dutch context, has shown lower teacher expectations of pupils with lower parental education level and/or a migration background (Denessen, 2017); consequently, these pupils' potential may be hampered as well as structurally underestimated.

The Dutch secondary education system initially had an inherent selection process related to CLIL programs. Until 2009, only pupils with above-average score on scholastic aptitude tests and high teacher recommendations had access to the streamed school types with CLIL programs—which may have additional selection procedures—and there were no CLIL programs at all available to pupils in the pre-vocational tracks. By 2012, however, there were 15 secondary schools offering pre-vocational CLIL (and that number had doubled by 2022). There are no explicit selection criteria for pre-vocational CLIL, only self-selection: it is open to all, limited only by geographical proximity. The pre-vocational CLIL curriculum usually consists of three or four CLIL subjects plus English as a foreign language (EFL), amounting to about eight to twelve lesson hours per week, approximately 30% of the total hours or about 275 hours per year. Schools are free to decide which subjects will be part of their own CLIL curriculum, depending on teacher availability, teacher ability, and suitability for cross-curricular collaboration. In the case of the current study, there was no single CLIL subject common to all schools; the national CLIL Standard for pre-vocational CLIL (Nuffic, n.d.-b) only stipulates that there must be at least one subject each from the fields of social studies, STEM, and arts or physical education. In contrast, mainstream (non-CLIL) pupils have only EFL lessons, usually two or three lesson hours per week or about 120 lesson hours per year. Although English is an officially required subject in the final two years of primary school, and extracurricular exposure to English is ubiquitous in the Netherlands due to tourism, social media, music, and the lack of dubbing of films and series, there are huge differences among pre-vocational pupils' exposure to and level of English (de Kraay, 2016).

To the best of our knowledge the present study is the first quasi-experimental study focusing on the development of L2 oral proficiency, oral fluency, or the willingness to communicate of CLIL pupils at the pre-vocational level as compared to their non-CLIL peers.

4.3B Research questions and hypotheses

Our research questions, therefore, are:

RQ1 What differences are there between pre-vocational CLIL and non-CLIL pupils' English writing proficiency and fluency, speaking proficiency and fluency, and Willingness to Communicate at the start of CLIL in secondary education (grade 7)?

RQ2 What is the effect of CLIL on pre-vocational pupils' growth in English writing and speaking proficiency and fluency, and WtC?

RQ3 What are the differential effects of CLIL on growth in English writing, speaking and WtC, dependent on pupils' grade, pre-vocational level, gender, scholastic aptitude test score, home language, and prior English-language instruction at primary school?

Our working hypothesis, based on results from previous research mentioned above, is that the CLIL pupils' level of English writing proficiency and fluency, speaking proficiency and fluency, and WtC will be significantly higher than the non-CLIL pupils at the start of CLIL provision in grade 7, as a result of informal self-selection procedures. Considering prior research results on target language proficiency growth in CLIL, we also hypothesize that CLIL pupils' growth in writing, speaking and WtC will be significantly higher than that of the non-CLIL pupils. We further want to verify whether CLIL is differentially effective for pupils with different background characteristics such as gender, home language, scholastic aptitude test scores, and years of prior L2 instruction.

4.4 Method and instruments

4.4A Design

In a two-year (2012-2014) longitudinal quasi-experimental study in pre-vocational secondary education, the differences in growth in language proficiency between CLIL (experimental group) and non-CLIL pupils (control group) have been estimated. Three cohorts of pupils were followed for two years starting respectively in years 1, 2 and 3 of pre-vocational education and finishing respectively at the end of years 2, 3 and 4. In the first two years of CLIL in pre-vocational education there were more CLIL subjects in the curriculum and more exposure to the target language than in the third and fourth years,

when there were fewer CLIL subjects and less L2 exposure in order to prepare pupils for their final exams, which are in Dutch.

4.4B Participants

All fourteen pre-vocational schools with an English-language CLIL stream in the Netherlands (in 2012, at the time of planning) were invited to participate in the research; six schools signed on to the project and an additional non-CLIL school was included for balance and to increase the power of the sample. None of the CLIL programs had any kind of school-based selection criteria beyond self-selection by the pupils and their parents. This resulted in a convenience sample of 603 pupils (CLIL n=313 and non-CLIL n=290; CLIL: girls n=168, boys n=145; non-CLIL: girls n=137, boys n=153) which at the start included pupils of three different secondary-level grade cohorts (Table 4.1). All six CLIL schools had started with a CLIL program between 2009 and 2011, so the numbers per cohort were lower in grades 8 and 9 because only two of the participating schools had a CLIL program before 2011.

The sample includes all four sub-levels of Dutch pre-vocational secondary education[11], from the most practical, least academic to the more theoretical and academic; several class groups in the study combined two adjacent levels, particularly in the lower-level CLIL groups, or shifted some pupils to the next higher or next lower level during the course of the study. There were 25 classes (15 CLIL classes and 10 non-CLIL classes) in seven schools, and all classes consisted only of pre-vocational stream pupils; three of these schools were CLIL-only and catered primarily to the less academic/more practical pre-vocational streams. Three offered a CLIL stream and a parallel monolingual (Dutch-language) regular stream, offering pre-vocational education only in the most academic/more theoretical sub-level. The seventh school offered only a Dutch regular stream at a less academic pre-vocational sub-level and functioned as part of the non-CLIL control group. The participating schools all were located in the densely-populated western part of the Netherlands in urban or semi-urban areas. The percentage of pupils with a home language background other than Dutch, the national language and the default school language, was 38.4%, which is higher than the national average (33.6%) for pre-vocational enrollment (Centraal Bureau voor de Statistiek, 2016). In the study sample, there were 44 different non-Dutch languages spoken at home by either one or both parents. Of these, Turkish and Moroccan Arabic were by far the most common home languages in the current study. English was a home language for less than 0.5%, and in no case used by both parents.

11 For Dutch readers: *basisberoepsgerichte leerweg (bb), kaderberoepsgerichte leerweg (kb), gemengde leerweg (gl), theoretische leerweg (tl/t/mavo).*

Table 4.1: Distribution of participants by CLIL/non-CLIL and cohort at the start of the two-year study

CLIL or non-CLIL	cohort 1 (grade 7, age 12-13)	cohort 2 (grade 8, age 13-15)	cohort 3 (grade 9, age 14-17)	TOTAL
CLIL	151	100	62	313
non-CLIL	163	70	57	290
TOTAL	314	170	119	603

To verify the comparability of the experimental and control group at the start of the study, correlations between five pupil background variables and experimental or control group membership were calculated. Four of the personal characteristics do not show significant correlations with group membership (1=experimental; 2=control) (gender: $r(601) = .064$, $p = .115$; scholastic aptitude (Cito): $r(331) = .093$, $p = .091$; mean number of years English at primary school: $r(564) = .007$; $p = .872$; frequency of those lessons: $r(563) = -.057$, $p = .180$. A low but significant correlation was found for the home language environment: the CLIL group slightly more often had a non-Dutch home language environment: $r(563) = .117$, $p = .005$). In other words, the CLIL and non-CLIL groups did not differ in individual pupil characteristics, except for a slight difference ($r^2 = .014$) in language background.

4.4C Instruments

Biodata questionnaire

The biodata questionnaire contained questions about gender, primary school scholastic aptitude test score[12], home language background, and years and frequency of primary school English lessons. The home language background variable is a summation of four dichotomous variables; whether the native language of the pupil's father and mother is Dutch, and whether the language spoken at home is Dutch (range 0-4). The primary school English years variable ranges from 0 (no English lessons at primary school) to 4 (four or more years of primary school English lessons); the weekly frequency of these lessons was also scored from 0 (no primary English) to 4 (four or more lessons per week). The classes had already been separated by pre-vocational levels (sometimes combined) by the schools and because of middle-level combination classes, the four pre-vocational levels were recoded into five groups for this research (1 = most theoretical; 5 = least theoretical).

Writing test: writing proficiency and writing fluency

The writing test was piloted with a non-CLIL pre-vocational grade 7 class not participating in the research project, in order to evaluate the face validity of the instruments, check

12 https://www.cito.nl/onderwijs/primair-onderwijs/centrale-eindtoets; this score was at the time the most important determiner of allocation of pupils into a pre-vocational subtrack.

the time needed to administer them, and ensure that all instructions and questions were clear.

The test designed especially for this research project was based on a series of prompts in Dutch instructing the pupils to write a letter to an imaginary British pupil who would be visiting their school in the Netherlands. The prompts were structured similarly to simple writing tasks used in foreign language teaching in the Netherlands, with each prompt on its own line and addressing a slightly different topic (Appendix 4, Attachment 4.1). Designed to stimulate the use of a range of vocabulary, grammar, and language functions (University of Cambridge ESOL Examinations [Cambridge ESOL], 2012), the prompts gave the participants the chance to write about familiar subjects as well as more challenging ones. The pupils were instructed to try to write at least 100 words and to write in complete sentences if possible. Lexical variety, grammatical and syntactical accuracy, and general structure were included in a rubric with a scale of 0–6 (Attachment 4.3), based loosely on the holistic rubric used by Verspoor (personal communication, August 27, 2012) and Verspoor et al. (2012), and the Cambridge ESOL (2012) assessment scales for assessing writing proficiency, while the number of English words was used as a measure of writing fluency, defined as 'utterance rate and length' (Wolfe-Quintero et al., 1988, p. 14).

Speaking test: oral proficiency and oral fluency

The speaking test was designed especially for this research project and piloted with non-CLIL pre-vocational pupils not in the research sample. It was based loosely on the formats of the individual speaking sections of the Cambridge ESOL 'Key English Test for Schools' (2012). There were three short parts: some introductory general questions using the present tense, some questions using the past and future tenses, and a short discussion of two contrasting photographs designed to elicit a range of vocabulary and modal verbs (see Attachment 4.2). The pupils were instructed to use as many words as possible and try to speak in complete sentences. All speaking tests were conducted individually, recorded by the lead researcher, and lasted about 3:00 minutes, except in cases where the pupil indicated that he or she was completely unable to continue in English. Lexical variety, grammatical and syntactical accuracy, (lack of) Dutch interference, intelligibility, and general communicative competence were used as categories to measure oral proficiency in a rubric with a scale of 0 to 6 (Attachment 4.4), corresponding generally to the CEFR levels pre-A1 to C1 (Council of Europe, 2001). This holistic rubric was loosely based on the categories and descriptions of the Student Oral Proficiency Assessment Rating Scale (SOPA) (Thompson et al., 2002) and Cambridge ESOL (2012). The number of English words produced within the three-minute test was used as a measure of oral fluency defined as 'utterance rate and length' (Wolfe-Quintero et al., 1988, p. 14).

Willingness to Communicate

Willingness to Communicate (WtC) was measured by means of two self-rated Likert-scale questionnaires with statements related to oral communication in English: seven statements each for two different situational aspects of WtC, one inside and the other outside the classroom. WtC-'School' items related to L2 communication situations inside the classroom, and WtC-'London' were related to L2 communication in (imaginary) situations in London. The questionnaire items (see Attachment 4.5), based on the two sets of eight items for measuring WCT at school (here called WtC-'School') and in a hypothetical situation abroad (here called WtC-'London') (MacIntyre et al., 2001), described a variety of situations, some familiar (asking questions in class) and some unfamiliar (calling a venue in London to inquire about tickets).

4.4D Procedure

All tests and questionnaires were administered in the pupils' schools by the lead researcher, assisted by a teacher at the school. A biodata questionnaire (Appendix 2, Attachment 2.1) was filled out once, at the start of the study. The writing test (proficiency and fluency) and the WtC questionnaire were administered on paper three times over the course of two academic years, from the start of one school year until the end of the following school year, 82 weeks in total. The second measurement occurred 30 weeks later at the end of the first school year (May-June). The third measurement was 52 weeks after the second measurement, in May-June of the following year. The individual speaking test (proficiency and fluency) was administered twice, at the first and last measurements, 82 weeks apart. As all first measurements for both CLIL and non-CLIL pupils took place in the first weeks of the school year, the starting levels for writing proficiency and fluency, speaking proficiency and fluency, and WtC of the cohort 1/grade 7 CLIL and non-CLIL pupils could reasonably be compared. Due to scheduling conflicts at several schools, not all pupils were able to participate in all measurements. The writing test (measuring both proficiency and fluency) was completed at least once of the three times by 594 pupils, resulting in a data set with 1367 measures of writing. Only the 289 pupils who completed the speaking test at both measurement moments were included in the analyses (CLIL n=147; non-CLIL n=142). The WtC questionnaire was completed by 590 pupils for at least one of the three WtC measurements, resulting in a data set with 1330 measures of WtC. To obtain unbiased estimates, all cases with one or more valid measures of WtC were included in the repeated measures analysis using a full information estimation procedure (maximum likelihood) (Hox, 2010, p. 106).

4.4E Scoring and data preparation (Writing and Speaking)

Over the two-year research period, the three measurements for writing and the two measurements for speaking produced a total of 2,167 tests to be analyzed for proficiency and fluency. For proficiency, the writing and speaking tests were made anonymous, transcribed, and assessed holistically by teams of six or seven assessors consisting of the

lead researcher and English teacher trainees, including at least two native speakers. For logistical reasons it was not possible to use the same teams for all the assessments. Each team of assessors attended training sessions which used the assessment rubrics and a set of 'anchor' texts from the first measurement which served as marking models. Each assessor was given a random selection of the tasks for a particular measurement moment, allocated with overlap between all possible pairs in a team so as to measure the overall consistency of rater agreement. The tasks were marked on a 7-point scale, from a score of 0 (barely identifiable as English) to 6 (age-corresponding native speaker level). The scores for the tests were obtained using at least two raters per product. Examples of pupils' anchor texts for writing proficiency for scores of 0 through 4 are given in Attachment 4.6, and for speaking proficiency for scores 0 through 4 are given in Attachment 4.7 (no scores of 5 or 6 were awarded). For measuring writing and speaking fluency, the number of English words in each text was counted.

4.5 Analyses

4.5A Reliability

Before conducting analyses to answer the research questions, the reliability of several instruments was calculated: writing and speaking proficiency and fluency, and Willingness to Communicate ('school' and 'London'). For the first two instruments this was done by calculating the inter-rater reliability (IRR). For WtC, homogeneity was calculated (Cronbach's alpha). For research at group level, an alpha above .8 is considered excellent (Albers, 2017). The reliability of these WtC instruments is thus excellent (range .87-.89). Table 4.2 presents descriptive statistics for the two WtC instruments at the three measurement points. The labels 0-M, 1-M, and 2-M indicate the baseline measurement, the first follow-up (nine months later), and the final measurement (approximately 21 months after baseline), respectively.

Table 4.2: Cronbach's alphas, Means and Standard deviations for WtC measurement instruments

Variable	Nr of items	Alphas	N	Means (SE)	SD
0-M WtC London	7	.88	489	26.52 (6.38)	0.94
1-M WtC London	7	.89	490	27.32 (0.29)	6.42
2-M WtC London	7	.89	351	28.20 (0.33)	6.18
0-M WtC School	7	.87	489	16.45 (3.60)	1.09
1-M WtC School	7	.87	488	16.94 (0.16)	3.62
2-M WtC School	7	.88	351	17.15 (0.19)	3.51

n.b. WtC *London* items are scored 1–5 (max. 35); WtC *School* items are scored 1,2,3 (max. 21)

To determine the assessors' individual inter-rater reliability (IRR) for the writing and speaking proficiency tests, three parameters were calculated: two indicating whether raters agree on the ranking order of tests rated and one to verify differences in strictness (higher or lower ranking for the same speaking test). Since a portion of the rated texts for each pair of raters overlap, first the parametric and non-parametric correlations between these portions of scores were calculated. Both the parametric (Pearson) and non-parametric (Spearman's rho) correlations were calculated, as shown in Table 4.3, because the number of rating categories was rather low since only the lower five of seven categories were actually used. These correlations indicate whether raters ascribe the same rank order to the tasks rated. If they do not show high positive correlations, raters disagree about this rank order and thus have rated different aspects of the tasks. A high correlation between pairs of raters is a necessary, but not a sufficient condition for concluding that ratings are comparable, as raters showing high correlations might differ in strictness. The correlations found can be considered substantial to excellent (Albers, 2017): range Pearson's .701-.961; range Spearman's rho .707-.970.

Table 4.3: Correlations for paired assessors for speaking proficiency tests

Task and measurement	Number of tasks per pair of raters for Pearson's correlation and/or Spearman's rho	Pearson's correlation (range)	Spearman's rho (range)
Speaking task 0-measurement	min. 55, max. 88	.701–.876	.707–.898
Speaking task 2-measurement	min. 14, max. 24	.796–.961	.720–.970

The degree of consistency of agreement regarding strictness is determined by examining the degree of agreement between the mean assessment mark of each pair of assessors. Because each assessor was assigned a random sample of the total number of tasks, we may assume that, given equal strictness, the expected means for pairs of assessors will be equal for all tasks assessed by that pair. However, when the differences in means were tested with dependent t-tests per pair of assessors, it became apparent that the assessors were not all equally strict: nearly half of the holistic assessments of writing and speaking proficiency (52 out of 107) had significant differences in means in paired sample t-tests (see Appendix 4, Tables 4.4–4.6 for writing proficiency, and Tables 4.7–4.8 for speaking proficiency). Therefore, despite the high conformity of ranking order, the degree of strictness did not conform.

These discrepancies in strictness were then compensated by calculating a z-score for the tasks assessed by each assessor, and separately for each measurement moment; for each measurement moment, the z-scores have a mean of 0 and an SD of 1 so that scores may be considered equally strict. Again, this is permissible because of the random distribution of pupil tasks among the assessors. Because there were in all cases more than 100 tasks per assessor, we may assume that all means are estimations of the population mean (the

population of all pupils who did the respective task). In this way it becomes possible to examine *whether* pupils in CLIL classes make more progress over time than pupils in the non-CLIL control group by inspecting differences in mean ranks, but it is no longer possible to measure *how much* progress is made, as the means for all measurement moments have been forced to a value of 0.

4.5B Regression analyses

The research questions have been answered by means of regression analyses. Since there are naturally-occurring groups in the sample (classes and schools), it is necessary to verify whether these analyses should be multilevel. Dependent variables measured three times (writing proficiency, writing fluency, WtC-'London' and WtC-'School') are analyzed using repeated measurement models, in which the variable 'time' indicates the time that has passed since the first measurement, so time at first measurement is coded as zero. Dependent variables only measured twice (speaking proficiency, speaking fluency) can be analyzed in two different ways: using the pretest scores as covariate (Ancova) and using Change scores (posttest minus pretest). There is discussion about which of these methods is preferable, since working with pre-existing groups or non-random allocation to groups in a design with pre- and posttest and an experimental and a control group may result in spurious effects (Lord's paradox). Allison (1990) recommends only trusting results found in both types of analyses. Van Breukelen (2013) argues that Change scores are the better option when working with pre-existing groups. To see if there is a difference, we have analyzed the speaking proficiency and speaking fluency data using both methods.

For each dependent variable we first checked which variance components (random intercepts) should be incorporated in the random part of the regression model. For the repeated measurement models this is done starting with a model with a repeated measures level and a pupil level and time as the only independent variable, subsequently testing fit improvement after adding a random class intercept and a random school intercept. In the Ancova models the fit improvement of adding a class- or school level is tested starting with a model with only the pretest as predictor, and in the models using Change scores as dependent variable no predictors were added when verifying necessary variance components. Fit improvement is tested using the chi-square distributed difference in –2 log likelihood (deviance) between nested models, where one model includes and the other excludes the random intercept variance component. The probability of the chi-square is divided by 2 in these tests, since variances cannot be negative (Hox, 2010).

After establishing the random intercept variance levels needed in the random part of the regression model, we checked for each dependent variable analyzed in a repeated measures model whether pupils attending CLIL differ on the first measurement in grade 7 from pupils not attending CLIL. This was done by examining the main effects of group membership (CLIL and non-CLIL), which in a repeated measures model is an effect on

the intercept or the starting value, since time starts at zero, after adding all main and interaction effects between pre-vocational-level, grade, CLIL and time (research question 1). Where there are only two measurement moments (speaking proficiency and speaking fluency), the first measurement is used as an independent variable to see if CLIL membership evinces a difference after controlling for pre-vocational level and grade.

Testing the significance of adding the interaction between CLIL participation and time gives the results for answering to the second research question. This effect is established after correcting for the effects of time, grade and pre-vocational level.

To answer the third research question concerning moderator effects, interaction effects between background characteristics of pupils and growth in proficiency were tested. These seven moderator variables were grade and pre-vocational level, plus the five pupil background variables: gender, scholastic aptitude test score, home language, years of primary school English, and frequency (lessons per week) of primary school English. The variables 'years of primary English' and 'frequency of primary English' were also combined as an interaction term (years*frequency) to operationalize the 'intensity of primary English'. For the repeated measurement models (writing proficiency, writing fluency, and WtC) the interaction effect of interest is the three-way interaction between time, CLIL and each moderator variable; for speaking proficiency and speaking fluency (Ancova models and Change models), it is the interaction between CLIL and each moderator variable. Additionally, we checked whether effects of CLIL differ for different combinations of grade and pre-vocational level, which for the repeated measurement models implies four-way interactions (time with pre-vocational level with grade with CLIL).

Significance of regression coefficients was calculated in two ways: by means of the Wald test (coefficient divided by the SE) and by verifying the difference in deviance between nested models. If N is large enough, the Wald test generates a z-score. For variables at higher levels (class and school level), the critical ratio of the Wald test is interpreted as a t-value, with the number of the degrees of freedom equal to the number of observations at the higher level (the number of schools or classes) minus the number of predictors minus 1. The effect sizes were estimated by comparing variance components of nested models and calculating proportions of explained variance.

4.6 Results: writing, speaking, and Willingness to Communicate

4.6A Initial differences between CLIL and non-CLIL groups

The first research question was whether the starting level of the CLIL and non-CLIL pupils in cohort 1 (grade 7) was equivalent for writing proficiency, writing fluency, speaking proficiency, speaking fluency, and WtC ('London' and 'School').

For Writing proficiency (see Appendix 4, Table 4.9), after correction for cohort and pre-vocational level and the two-way interactions between these predictor variables (model 8), the non-CLIL pupils in grade 7 start approximately one-fourth point lower (-0.265, p = n.s.) on a seven-point scale than their CLIL peers in grade 7 who are just beginning with the CLIL program. After adding the three-way interactions between CLIL, grade, and pre-vocational level (Table 4.9, model 9), this difference remains non-significant, implying no initial advantage for the starting CLIL pupils after correcting for the effects of cohort and pre-vocational level.

For writing fluency (Table 4.10, model 8), after correction for the main effects of grade and pre-vocational level, and the two-way interactions between these variables, the CLIL pupils in grade 7 show a positive advantage at the outset, writing approximately 22 words more on average at the beginning of grade 7 ($p < .05$). However, after the inclusion of three-way interactions between CLIL, pre-vocational level, and grade (Table 4.10, model 9) the difference is no longer significant and no initial advantage for the starting CLIL pupils was found.

For speaking proficiency and speaking fluency (Table 4.11), the multilevel models with the pretest scores as dependent variables show that at the start of pre-vocational secondary education in grade 7, the CLIL pupils had significantly higher speaking proficiency scores ($p<.01$, model 4) and used significantly more words in the speaking test (37.6 words more, $p<.01$, model 8) than their non-CLIL peers. After controlling for grade and pre-vocational level, the total variance explained by belonging to a CLIL class at the start of CLIL in grade 7 was 12.5% for speaking proficiency and 10.1% for speaking fluency (Table 4.11).

The measurement of WtC-'London' shows that after correction for the main effects of grade and pre-vocational level, the CLIL pupils start significantly higher at the beginning of secondary school than their non-CLIL peers (Table 4.12, model 7), but after inclusion of the two- and three-way interactions between these variables and CLIL, the difference between the two groups of pupils at the outset is no longer significant. In Table 4.13, however, the results for WtC-'School' indicate that the difference between the non-CLIL

pupils in grade 7 and their peers just starting CLIL remains significant even after the two- and three-way interactions are added (Table 4.13, model 8: -1.003; model 9: -.963; $p<.05$).

In summary, regarding the answer to the first research question, it appears that at the beginning of pre-vocational CLIL, the starting grade 7 CLIL pupils had similar initial levels of writing proficiency, writing fluency, and WtC-'London'. However, these pre-vocational pupils just starting CLIL were already significantly more proficient and more fluent in speaking, and also reported significantly more L2 WtC in a familiar school situation than their non-CLIL peers.

The second research question concerned the effect of CLIL on the growth in L2 writing proficiency, writing fluency, oral proficiency, fluency, and WtC, controlling for the initial starting differences in cohort 1. These results were obtained using the same instruments as those used for the third research question (*Is there a differential effect of CLIL on growth in English writing proficiency, writing fluency, speaking proficiency, speaking fluency, and Willingness to Communicate, dependent on individual variables?);* therefore, the results for these two research questions are reported together for each instrument, rather than separately.

4.6B Writing proficiency

The regression analyses (Table 4.9) show that in addition to the repeated measures level and pupil level, a random intercept at class and school level needs to be included in models predicting writing proficiency. Since we use z-scores of pre- and posttest (see Analysis), real growth is not measured, only relative growth or differences in rankings between pre and posttest. The analyses are built up by adding one or more predictors or interactions to the next model.

After controlling for the main effects of grade and prevocational level and their 2-way interaction terms (model 7), adding the interaction between CLIL and time (model 8) shows that there is a significant effect of CLIL ($p<.001$) but with a small effect size. CLIL explains 1.9% of the *relative* growth (differences in rank from pre- to posttest). After adding the 3- and 4-way interactions the effect of CLIL is no longer significant. Adding these interactions seems to result in a fragmentation of the effect over all interactions which contain the CLIL variable. In short, CLIL shows a significantly positive but small effect on writing proficiency but only when we do not control for three- and four-way interactions between pre-vocational level, CLIL, time and grade.

In the tables with pupil characteristics variables (Tables 4.14–4.15), none of the personal characteristics variables (gender, scholastic aptitude test score, home language, years or frequency of primary school English instruction) moderated the effect of CLIL on the growth in writing proficiency. Pre-vocational level also shows no moderator effect. Grade,

however, does moderate the effect of CLIL on growth in writing proficiency. In Table 4.9, model 8, CLIL shows a small overall positive effect (regression weight of non-CLIL= -0.003; p<.001). After adding the three-way interactions however, this effect is no longer significant while the moderator effect of cohort 2/grade 8 (the three-way interaction between Time, non-CLIL and grade 8=0.005; p<.05) shows a negative effect of CLIL for students in grade 8. This moderator effect is also apparent in Table 4.14, models 1, 2, 5 and 6, where the small positive effect of CLIL in cohort 1/grade 7 and cohort 3/grade 9 similarly is absent in cohort 2/grade 8 (e.g. model 1 where the overall effect of non-CLIL is shown in the regression coefficient of -0.004 (p<.001), but for grade 8 this is neutralized through a moderator regression coefficient of 0.006 (p<.01) of 'time*non-CLIL*grade 8').

4.6C Writing fluency

The preliminary models (Table 4.10) show that three random intercept levels need to be included in the random part of the regression model: class, pupil and repeated measures. The mean number of words shows significant growth for all pupils (main effect of time, all models; p<.001). Model 7 is introduced with all two-way interactions except non-CLIL*Time, in order to be able to calculate the effect size of participation in CLIL on growth in writing fluency after correcting for differences in growth caused by grade or pre-vocational level. In model 8 we see that after these corrections, CLIL has a small but significant positive effect on growth in writing fluency (p<.05), explaining 1% of growth in fluency. This effect becomes even larger when we correct for the three- and four-way interactions (models 9 and 10).

The written fluency of the lower pre-vocational levels grows more slowly in both CLIL and non-CLIL, but overall, there is a significant higher growth for the CLIL pupils compared to the non-CLIL pupils, regardless of their pre-vocational level (see Table 4.10; higher values indicate lower academic tracks).

Just as with the results for writing proficiency, none of the personal characteristics variables results in a significantly better model fit, meaning that no significant moderating effect of these variables on the effect of CLIL on writing fluency was found (see Tables 4.16–4.17). Pre-vocational level does not moderate the effect of CLIL on growth in writing fluency, but grade does (Table 4.10). The effect of CLIL for the two years of cohort 1 is shown in the regression coefficients of time*non-CLIL (in Table 4.10, model 9: -0.423* and model 10: -0.442*). However, this positive effect of CLIL for cohort 1 does not apply to cohort 2 and 3, since the three-way interactions between cohort 2/grade 8, time and non-CLIL and cohort 3/grade 9, time and non-CLIL are significantly positive (except for time*grade8*non-CLIL in model 10) and as large or even larger (0.414* to 0.511*) than the positive effect of CLIL in cohort 1/grade 7 (-0.423 to -0.442). This implies that CLIL only shows a significant positive effect on growth in writing fluency in the grade 7 cohort.

4.6D Speaking proficiency

The multilevel models for speaking proficiency (models 1–4) and speaking fluency (models 5–8) are presented in Table 4.11. The preliminary ANCOVA analysis shows that three levels of variance are needed: pupil, class, and school. After controlling for pre-vocational level and grade (Table 4.18, model 8), there is a small significant effect of CLIL ($p<.05$, total percentage of explained variance 2.4%). Using Change scores (Allison, 1990), the results are the same: CLIL has a significant positive effect ($p<.05$%, total explained variance 3.0%) on the growth (shift in ranking) in oral proficiency (Table 4.19).

Tables 4.20 and 4.21 show that none of the moderator effects, including the moderator variables grade and pre-vocational level, proved statistically significant. The same results are found using the Change score approach (Tables 4.22–4.23).

4.6E Speaking fluency

The analyses for speaking fluency followed the same pattern as for general speaking proficiency. The Ancova approach (Table 4.24) has three random intercept levels: pupil, class, and school level. In model 8 we see that adding the CLIL variable results in a significant model fit improvement ($p<.05$). CLIL pupils gain 20.63 words more ($t=8.550$; $df=19$; $p<.001$) than non-CLIL pupils from pretest to posttest. The percentage of total variance in learning gain explained by CLIL is 5.9%. In the Change score approach (Table 4.25), we need the same three random intercept levels, but the CLIL variable (model 8) is non-significant ($b=-13.969$; $t=1.573$; $df=20$; p=n.s.). The Ancova and the Change score analyses thus indicate the same trend towards more fluency development for the CLIL pupils, but according to van Breukelen (2013) the Change score analysis is preferable. Strictly speaking, we cannot conclude that there is a significant effect of CLIL on speaking fluency.

The Ancova (Tables 4.26–4.27) and the Change score (Tables 4.28–4.29) analyses for oral fluency both show that there are no significant moderator effects of any of the five background variables, nor of pre-vocational level, on the effects of CLIL on speaking fluency. However, both types of analyses reveal a marginally significant ($p<.10$) moderator effect of the scholastic aptitude test on the effect of CLIL (Table 4.26, model 7 and 8 and Table 4.28, model 6 and 7), indicating that the higher the CLIL pupils scored on this test at the end of primary school, the smaller the effect of CLIL on speaking fluency. In other words, CLIL pupils with *lower* scholastic aptitude scores showed *more* speaking fluency growth than the higher aptitude CLIL pupils, as compared to their same-aptitude non-CLIL peers. The effect of CLIL on growth in fluency is thus larger for pupils scoring lower on the scholastic aptitude test.

4.6F Willingness to Communicate (WtC)

For WtC-'London' (Table 4.12), two random intercept levels are used in the analysis: repeated measures and pupil level. WtC-'School' (Table 4.13) also includes a class level

intercept variance. Both CLIL and non-CLIL pupils indicate significantly more WtC over time, but CLIL does not show any additional effect on growth of WtC, whether in London or at school, after adding the combinations of three- and four-way interactions of grade, pre-vocational level, CLIL with the variable 'time' (Table 4.12, model 10; Table 4.13, model 10). Since there are no effects of CLIL on these measures, no moderation analyses with background variables were conducted.

4.7 Discussion and conclusion

4.7A Summary of findings

The aim of the present study was threefold. First, we wanted to ascertain whether at the start of CLIL the level of English writing proficiency, writing fluency, speaking proficiency, speaking fluency, and Willingness to Communicate of pre-vocational CLIL pupils was equivalent to that of their non-CLIL peers, in the absence of any explicit school-based selection procedure. Second, we wanted to analyze whether the CLIL pupils showed more positive development than the non-CLIL pupils in these productive skills and WtC over a research period of two school years. Thirdly, we wanted to see if there were differential effects of five background variables which have been shown to moderate the effects of CLIL in some previous studies.

First, there was no initial significant difference between the two groups for writing proficiency after controlling for pre-vocational level, grade, and the interaction between these two variables. For writing fluency, we do find a significant difference of 28 words in favor of the CIL pupils after correcting for pre-vocational level, grade, and the two-way interactions between these two variables. However, after adding the three-way interactions between pre-vocational level, grade, time, and CLIL, the difference is reduced to 17 words and is no longer significant. This lack of a significant effect may be due to a lack of writing experience, as the exposure to English at most Dutch primary schools is mainly oral (de Kraay, 2016), and the skill 'writing' is absent from the primary school English curriculum guidelines, apart from a minimal amount of exposure to spelling (Thijs et al., 2011). In spite of this prior lack of experience with English writing, the fact that the CLIL cohort 1 pupils wrote more words on average—even though their writing proficiency was not significantly more advanced than that of their non-CLIL cohort 1 peers—suggests a greater readiness to express themselves in writing and less hesitation to attempt written communication in English, even at a modest proficiency level.

Although no formal selection criteria or procedures were in place, pupils who chose for the CLIL stream at the start of secondary school (cohort 1/grade 7) showed significantly higher initial levels of speaking proficiency and speaking fluency. Generally, these CLIL pupils also rated themselves as significantly more willing to communicate in English in

their school context. They seemed to be more positively orientated towards English and use it more willingly from the outset, at least in the classroom, even if their language level was low.

Like the results for writing, there was no significant difference found between the two groups at the outset for Willingness to Communicate in an unfamiliar English-speaking environment (London). The divergence in WtC at the start of cohort 1 could be explained by situational context: it appears that CLIL pupils are more self-confident and less anxious than non-CLIL pupils about communicating in familiar situations, but not in unfamiliar ones.

4.7B Interpretation and implications

The second question concerned the longitudinal development of these skills for three cohorts of CLIL and non-CLIL pupils. The writing and speaking proficiency scores showed a significantly higher growth for the CLIL pupils, albeit with small effect sizes. These small effect sizes may partly be explained by the assumption posited by Verspoor et al. (2013) in research with CLIL and non-CLIL pupils in the first three years of a more academic secondary stream: progress at a lower starting proficiency level is generally faster than at a higher level, where relatively more linguistics subsystems must become more advanced and complex to show progress. The cognitively challenging, context-reduced language of school subjects requires a relatively longer time to begin to master than basic L2 skills (Collier, 1989). The same trend can be seen in writing fluency, also with a small effect size: the CLIL pupils' scores showed higher growth over time. For speaking fluency, the two types of analyses did not yield the same results: the Ancova analysis showed a significant greater gain for the CLIL group, but although the Change score analysis followed the same positive trend, that result was only significant at 10%. As described above, the Change scores are preferable in this non-random sample with pre-existing groups (van Breukelen, 2013). Thus, compared to some other studies (e.g. Dalton-Puffer et al., 2008), we cannot confirm the CLIL advantage for speaking fluency as measured by number of words. This invites speculation: as CLIL is theoretically based on the communicative approach to language learning, it should contribute substantially to the development of oral fluency. Perhaps, in the current study, schools and teachers were not yet able to optimally foster communicative classroom practice in the target language. Most were new to CLIL and to the challenges of adapting a fusion of a foreign language and their subject content to a scholastically-challenged pupil population. It is also possible that the pre-vocational CLIL subject teachers' English proficiency level was not yet optimal for them to serve as ideal language models and stimulate target language communication, or that they were not sufficiently able to simplify their language to enable comprehension, or that they felt that as subject teachers, their primary aim was to impart subject content rather than encourage L2 communication.

Regarding the results of the effects of CLIL on growth of Willingness to Communicate, neither WtC-'London' nor WtC-'School' showed any significantly larger gains for the CLIL group, similar to the results in Goris et al. (2019). It is possible that the WtC of the CLIL pupils did not significantly increase because they experienced anxiety or a lack of self-confidence about meeting the more communicative demands of CLIL (Lialikhova, 2018). We did not gather data on perceived motivation. However, with respect to attitude, Denman et al. (2018; see Chapter 2 of this dissertation) found that pre-vocational CLIL pupils scored significantly higher than non-CLIL pupils on four out of five attitudinal constructs, although this does not appear to translate to increased WtC in the current study. As not only affective constructs, but also a learner's perceptions of competence affect their WtC (Baker & MacIntyre, 2000), it may also be that CLIL learners develop a more critical and nuanced view of their own competence through more exposure to a wider variety of input and productive situations, and that their WtC does not increase more rapidly than that of their non-CLIL peers. Although a symbiotic relationship between CLIL and WtC has been suggested whereby CLIL helps to develop WtC and WtC in turn has a positive effect on language skills in CLIL (Kang, 2005; Menezes & Juan-Garau, 2015), the current results do not confirm this. Increasing opportunities for productive use of the target language and more attention to communicative competence might help to make learners both more willing and more competent.

4.7C Moderator effects and learner variables

We were also interested to know whether CLIL was more or less successful with pupils in different pre-vocational subtrack levels, different years, or with certain background characteristics (research question 3). Regarding pre-vocational level, the analyses showed no moderating effect of this variable on CLIL provision, indicating that the L2 gains of the CLIL pupils in the less academic pre-vocational levels were not less than those of their peers in the higher pre-vocational levels. There was, however, a significant effect of cohort/grade: the CLIL pupils in cohort grade 7 showed the greatest growth over two years in writing proficiency, writing fluency, and speaking fluency compared to their non-CLIL peers. In the higher cohorts/grades this positive advantage diminished; this deceleration might have been partly because the CLIL hours of exposure and number of CLIL subjects decrease in pre-vocational years 3 and 4 as preparation begins for Dutch-language final exams; this same dip in L2 development was also noted in a higher-level CLIL track (Verspoor et al., 2013). Another possible explanation of this advantage for this youngest cohort could be their enthusiasm for school and also for CLIL at the start of secondary education and correspondingly more amotivation and school alienation as pupils grow older, especially for lower achievers (Morinaj et al., 2020). It seems also reasonable to assume that since pre-vocational CLIL was a novelty when the older cohorts started, the schools and teachers became more skilled in CLIL teaching and understanding the particular needs of this group of learners, and subsequent CLIL cohorts likely benefitted from this acquired experience.

Regarding the five pupil characteristics, four of the five background variables—gender, home language, and years or frequency of primary school English lessons—had no significant moderating effect on the effect of CLIL provision. This result contrasts with some previous research (Lahuerta Martínez, 2020; Merisuo-Storm, 2007) which has shown that CLIL can help equalize gender-based differences in foreign language learning; in our case, there was no starting advantage for CLIL girls or 'catching up' for CLIL boys.

Similarly, there was no differential effect of the variable indicating the amount of the majority language (Dutch) spoken in the home environment on the effect of CLIL, which indicates that a native Dutch language background, or a migration background and different home language, offers no significant advantage or liability for target (English) language proficiency in CLIL. This confirms other studies (Schwab, 2013; Somers, 2017) which maintain that there is no inherent disadvantage for minority-language pupils in CLIL. This may be partly dependent on CLIL lessons following the 'target language = classroom language' principle (Westhoff, 2005) rather than making extensive use of the majority language, where native Dutch pupils would likely have an advantage.

We also found no moderating effect of 'prior years of primary school English' or the frequency of those lessons; our results indicate that CLIL is no more or less effective for pre-vocational pupils who have had more or fewer years and/or frequent primary school English lessons. This reinforces research by de Kraay (2016), who maintains that in effect the English-language instruction at primary school in the Netherlands is generally playful and easygoing, beginning effectively anew at absolute beginner level at secondary school. Perhaps CLIL builds on a foundation of extramural exposure to English, particularly through English-language television and gaming (Naber & Lowie, 2012). The lack of a moderating effect of prior English lessons underscores the inclusive, egalitarian aspirations of pre-vocational CLIL in the Netherlands, as it means that pupils who have had only minimal exposure to English prior to secondary school are not at a disadvantage in CLIL.

Finally, there was also a marginally significant moderating effect ($p<.10$) of the primary school scholastic aptitude test on the effect of CLIL for speaking fluency, both the Ancova and the Change score analyses. Pupils with a lower scholastic aptitude test score benefitted more from CLIL than pupils with a higher aptitude score; that is, they showed more growth in number of words spoken. Similar to the writing and speaking proficiency results, this might be attributed to faster progress at a lower level of fluency.

4.7D Limitations and recommendations

There are a number of limitations to the present study. Although our biodata questionnaire included items designed to measure the amount of out-of-school contact with English, too many pupils did not complete these consistently, so we could not reliably include extracurricular English as a possible moderating variable. It was also not possible

to test subject-related speaking proficiency or vocabulary, since there was no common CLIL subject which was offered at all schools. Because subject-specific language most likely varied considerably between schools, it was decided not to take this into account in testing. In any case, no validated tests are available for subject-specific language in lower vocational levels. The writing and speaking tests therefore targeted generic, familiar language and may not have given all pupils the opportunity to optimally display subject-specific language learned in CLIL classes. We also used a holistic scores rubric to measure writing and speaking proficiency; although this included the domains of grammatical, lexical, and syntactical range and accuracy, these domains were part of a composite score rather than being scored separately. The use of z-scores for the writing and speaking proficiency measurements implied that it was no longer possible to measure the specific amount of progress over time—which would have been desirable—but it was only possible to determine that the gains were significantly larger in the CLIL group. Regarding the WtC questionnaire, this was limited to the pupils' own self-assessment rather than a more objective measurement, and they may have either over- or underestimated themselves in some cases. The questionnaire, and hence the current study, did not explore the broader issues of motivation, attitude or self-confidence, all of which are aspects of WtC (MacIntyre et al., 1998).

Further studies, therefore, could include a measure of extracurricular English to explore a correlation between this and CLIL/non-CLIL gains, and to what extent the gains might be attributable to extracurricular English rather than to CLIL, as has been done in a more academic Dutch secondary education context (Verspoor et al., 2011). Further studies might also focus on disciplinary literacies or content subject attainment levels by comparing results from different schools with one or more common CLIL subjects, or use more detailed linguistic analyses to explore the L2 writing and speaking gains made over time in various linguistic domains, such as the use of chunks (Smiskova et al., 2012) or other analyses of complexity, accuracy, and fluency (Housen et al., 2012). Regarding the rating of the writing and speaking tests, a closer alignment of the raters' assessments could be achieved by providing additional training, or by having the raters confer until a consensus is reached (Verspoor et al., 2012). Finally, Mearns et al. (2020) have emphasized the importance taking not only academic factors but also affective factors into consideration when comparing differences between different groups of learners. Future studies could explore correlations between pupils' WtC self-assessments and their actual performance, perhaps including variables related to motivation, self-confidence, enjoyment, or anxiety. There is clearly a need for further research in pre-vocational and other less privileged contexts in order to build up an empirical foundation for a more complete understanding of inclusive bilingual education with CLIL.

4.7E Conclusion

This two-year longitudinal study compared three cohorts of CLIL and non-CLIL pre-vocational pupils in 25 classes of the least academic secondary levels in seven schools in the Netherlands. It measured written and oral proficiency and written and oral fluency development in English as well as L2 willingness to communicate (WtC) inside and outside the classroom. The results generally show a difference in favor of the pre-vocational CLIL pupils at the outset in grade 7, despite the lack of any school-based selectivity. Over time there was a significant advantage for the pre-vocational CLIL pupils over their non-CLIL peers for writing proficiency, writing fluency, and speaking proficiency. A significant advantage for the youngest CLIL cohort was also found for speaking fluency, although this advantage was not present in the older cohorts, possibly partly due to the decrease in CLIL subjects and hours. There was, however, no significant WtC growth advantage for the CLIL pupils. Interestingly, pupils with a minority home language or who had little to no primary school English-language instruction did not appear to be at a disadvantage in the CLIL program. Despite the small effect sizes found, these nascent pre-vocational CLIL programs have shown that non-selected, non-elite pupils—also when they are academically-challenged, have a migration background, a learning or behavioral disability, or otherwise can be considered 'at-risk'—can clearly benefit from bilingual education with a CLIL approach. As issues related to educational equality and inclusivity are gaining more attention in diverse CLIL contexts, the current study offers results that bode well for the future of more inclusive CLIL programs.



5

“They underestimate us”: Pupil and teacher perceptions of CLIL in pre-vocational secondary education

Abstract

This mixed-method, multi-perspective study explores how teachers and pupils in Dutch pre-vocational secondary education perceive and experience Content and Language Integrated Learning (CLIL). The research focuses on pupils with lower scholastic aptitude across all four levels of Dutch pre-vocational secondary education (*voorbereidend middelbaar beroepsonderwijs*, or *vmbo*[13]*)* and their teachers, in the early stages of CLIL implementation (2012-2014) in this educational track. Data were collected from 53 CLIL teachers—both English and subject specialists—and 267 pupils from 15 CLIL classes at the six *vmbo* schools included in the broader research project, three of which offered CLIL throughout the school and three which provided a CLIL stream alongside a regular Dutch-medium stream.

Combining quantitative questionnaire data with qualitative focus group interviews, the study investigates how teachers and pupils evaluate the affordances and challenges of CLIL for this learner group. Unlike the more academically oriented CLIL contexts often described in the literature, this setting foregrounds issues of affect and inclusion: pupils frequently reported feelings of being underestimated, and valued the empathy, patience, and encouragement shown by their teachers. Using thematic analysis (Braun & Clarke, 2006, 2022), the chapter identifies areas of convergence and divergence in stakeholder perceptions of classroom language use, pedagogical strategies, and the balance between language and content learning. In line with calls to foreground stakeholder voices in CLIL evaluation and educational innovation (Coyle, 2013; Coyle et al., 2010; Pérez Cañado, 2020), the study highlights the importance for both pedagogy and policy of listening to those most directly impacted by its implementation. Although CLIL teachers experienced many of the same pressures related to workload, time, and materials as their colleagues in more academic tracks, and faced additional challenges arising from their pupils' relatively limited target language proficiency, they were nonetheless broadly positive about their own linguistic and pedagogical development within the CLIL environment. Pupils, too, were positive about CLIL and their improvement in English, but the findings also reveal striking variation among pupils, classes, and programs in terms of perceived challenge, teacher language use, and overall engagement. Conducted during the formative years of pre-vocational CLIL in the Netherlands, the study offers rare empirical insight into how teachers and pupils jointly make sense of bilingual education in an inclusive, non-selective setting.

13 Because this chapter deals extensively with the authentic voices of teachers and pupils, the Dutch term for pre-vocational or junior vocational education (*vmbo*) is used more frequently in this chapter.

5.1 Introduction and rationale

In the history of CLIL research, early studies were mainly focused on the learner outcomes of target foreign language learning and linguistic competence, and only later began to describe actual CLIL practice and explore the perceptions of CLIL participants (Coyle, 2007; Dalton-Puffer et al., 2022; San Isidro & Pérez Cañado, 2023). There is still a lack of research focusing on the perceptions of the key participants, and further exploration is called for (Barrios, 2022; Barrios & Milla Lara, 2020; Cenoz et al., 2014; Hüttner et al., 2013). An investigation into the perceptions of primary stakeholders in Content and Language Integrated Learning (CLIL) can provide valuable insights into the practical realities of this educational innovation. The perceptions of these stakeholders on the front line—CLIL pupils and their teachers—can reveal the strengths and weaknesses of the approach, inform future pedagogical strategies, and contribute to the broader discourse on bilingual education.

More specifically, research into the perceptions of pupils and teachers in CLIL programs can help identify disjunctures between the perceptions of pupils and their teachers regarding their respective expectations, satisfaction, and needs, as their different opinions on classroom teaching and learning are not always clearly communicated to each other (Wegner, 2012). They can serve to fine-tune and improve CLIL provision and coordination between language teachers and subject teachers (Bonnet, 2012) and can offer first-hand information about teacher training needs. They can cast light on pupil motivation and engagement (Dörnyei, 2019) and facilitate the identification of learner emotions, which is important because positive and negative learner emotions have been shown to respectively have a significant positive or negative effect on their foreign language performance and progress (Botes et al., 2022; Dewaele et al., 2025; Lasagabaster & Sierra, 2009). They can provide an assessment of CLIL teaching from a different angle than that of researchers or outside experts (Mahan & Norheim, 2021), and help both researchers and practitioners better understand whether CLIL is being implemented effectively (Pladevall-Ballester, 2015).

Of particular relevance for the current study are a number of other purposes/benefits: they can help others to make a decision regarding the possible implementation or improvement of CLIL at pre-vocational level or for less academic learners in general (van Dongen, 2017), and help bridge the gap between research claims and actual classroom experiences and emotions, which can be especially important with lower-proficiency learners (Pladevall-Ballester, 2015). They can enable more detailed comparisons between research in other CLIL contexts and settings (San Isidro & Pérez Cañado, 2023), or comparisons in the same CLIL context but with different participants or at a different time. And finally, they can provide insights into how quantitatively-measured learning outcomes—especially with the same participants—might be interpreted, enabling more

nuanced speculation through the integration of quantitative and qualitative approaches (Bonnet, 2012).

5.2 Theoretical and empirical background on CLIL perceptions

It may be useful to review the principles guiding the CLIL approach before examining prior research on CLIL participants' perceptions of the success or failure of some of the aspects of CLIL practice. According to CLIL scholars, CLIL lessons should have certain foundational characteristics, some of which will be more relevant for either content or language teachers at any given time. They should:

- be at a target language level that offers comprehensible input (Krashen, 1981), but is still (minimally) challenging (de Graaff et al., 2007; van Kampen, 2019) and is at the right cognitive level (Dale et al., 2011)
- be primarily content-driven, at least in the CLIL subject lesson (Coyle et al., 2010)
- have interrelated learning goals for both language and subject content (Coyle et al., 2010, p. 36; Coyle & Meyer, 2021)
- encourage output, stimulating pupils' language production to communicate with the teacher and each other in the target language (Dale et al., 2011; de Graaff et al., 2007; van Kampen, 2019)
- be taught by teachers with a range of learning skills and strategies for both language and content acquisition, and with a high enough L2 level to be a language model (Ball et al., 2015)
- include a variety of didactic activities and tasks (Coyle et al., 2010) and include multimodal input (Dale & Tanner, 2012)
- avoid decontextualized grammar and vocabulary, but still include some 'focus on form' in a counterbalanced approach (Lyster, 2007; Swain, 1993) and implicit or explicit corrective feedback (de Graaff et al., 2007)
- include a focus on genre-specific language in subject lessons, where appropriate (Llinares, 2015)
- help develop cultural awareness and intercultural understanding, whether at the micro- or macro-level (Coyle et al., 2010).

From early on, charges of elitism and educational creaming have been leveled at CLIL programs. These were not unfounded, as most early CLIL programs were aimed specifically at, and through various selection procedures limited to, pupils who excelled academically (Apsel, 2012; McClintic, 2022; Otinowska & Forys, 2017; Paulsrud, 2009; Siepmann et al., 2023, Verspoor & Edelenbosch, 2011). To some extent, CLIL programs are still associated with prestige (Madrid & Pérez Cañado, 2018), but the landscape is gradually changing, partly due to calls for more attention to diversity, equity, and inclusion,

and partly because the generally successful selective, high-status CLIL programs are being expanded to include average, non-selected, or less academically privileged learners. However, research on lower academic achievers in CLIL in this area remains scant. This is especially true for the perceptions, emotions, and experiences of these pupils and their teachers, and their recommendations for improvement. There are gaps in the literature regarding stakeholder perceptions in the less privileged learner context, including in the pre-vocational (*vmbo*) context in the Netherlands (van Dongen, 2017), which is the context of the current study.

CLIL research is characterized by a broad spectrum of contexts, including location, mother tongue, intensity of the CLIL program, teacher and pupil L2 proficiency level, extramural exposure to the target language, national policy, school type, educational level, and academic stream. The literature germane to this qualitative chapter has been selected from secondary and late primary education, and is mainly—but not exclusively—from various European contexts. The following section presents an overview of research results on the perceptions of CLIL pupils and their teachers regarding various aspects of CLIL. The overview is divided into positive perceptions, negative perceptions, and perceptions related to academically-challenged learners: first from teachers and then from pupils, followed by their suggestions for improvement. A summary of a preliminary study of pre-vocational CLIL teachers and pupils in the Dutch context (Denman et al., 2013) is given at the end of this overview.

5.2A Teacher perceptions in CLIL research

One of the most frequently reported CLIL teachers' positive perceptions is the noticeable target language gains made by their pupils (Iwaniec & Halbach, 2024; Lancaster, 2016; McDougald, 2015), particularly in the area of oral comprehension and willingness to communicate (Milla Lara & Casas Pedrosa, 2018; Pladevall-Ballester, 2015), as well as content knowledge (McDougald, 2015).

Most CLIL teachers are non-native speakers of the target language (Dalton-Puffer et al., 2010a) and they observe that not only does their own target language proficiency improve, but as language learners, they empathize with their pupils regarding the shared linguistic challenge of CLIL (Iwaniec & Halbach, 2024). They also notice aspects of CLIL pupils' positive affect: more self-confidence, interest, enthusiasm, and enjoyment (Hunt, 2011; Milla Lara & Casas Pedrosa, 2018), and that pupils participate actively and are more engaged (Coonan, 2007), with the result that they are quieter and more concentrated, and classroom behavior improves (Hunt, 2011).

Teachers also note that CLIL teaching involves making some changes to their methodology, pedagogy, and/or language use. They appreciate the active, hands-on aspect of CLIL pedagogy (Bauer-Marshallinger et al., 2023) and the change from a grammar-based

methodology to one with more communicative goals (Iwaniec & Halbach, 2024). They report that CLIL pupils are offered more appealing content in the target language than in traditional language lessons (Hunt, 2011) and that—at least in the case of English-language CLIL—the larger variety and availability of resources in English reduces the dependency on textbooks and leads to more interactive tasks and 'real' activities (Iwaniec & Halbach, 2024). When linguistic comprehension cannot be taken for granted, teachers take greater care in ensuring that the content is understood (Coonan, 2007) and change their teaching methods to be more learner-centered and learner-responsive (Iwaniec & Halbach, 2024; Kashiwagi & Tomecsek, 2015). They appreciate the flexibility of using the L1 to help overcome learning difficulties (Coonan, 2007). And finally, they report that despite the extra effort, work and time involved, the investment is worthwhile (Banegas & Cerrato Will, 2022; Iwaniec & Halbach, 2024; Milla Lara & Casas Pedrosa, 2018).

Other results highlight CLIL challenges and frustrations. The top challenge according to teachers is a lack of appropriate materials and resources, which seems to be directly related to teachers' perception of a heavy workload and not enough time. Materials may be unsuitable, unreliable, unmotivating, or must be adapted to the teachers' and learners' own context (McClintic, 2022; Merisuo-Storm, 2007; Pérez Cañado, 2023; Roiha, 2014; Siepmann et. al; 2023; van Kampen et al., 2020), and the process of finding the right level can be "very, very difficult" (Gierlinger, 2017, p. 106). Other teachers mention the lack of time to prepare CLIL lessons and incorporate CLIL methodology (McDougald, 2015; Pladevall-Balester, 2015), saying that the workload is generally heavier for CLIL lessons (Gálvez Gómez, 2022) and that while it is possible to go into a non-CLIL lesson unprepared, this is impossible for a CLIL lesson due to the extra L2 preparation necessary (Iwaniec & Halbach, 2024). A lack of time also creates challenges for collaboration with CLIL colleagues and for peer support (Pérez Cañado, 2023; Pladevall-Balester, 2015). In contrast to the positive perception of pupil behavior noted above, some teachers found CLIL pupils' behavior difficult and undisciplined (Gierlinger, 2017), or noted that pupils paid much less attention when the L2 was used (McClintic, 2022).

CLIL teachers also voice concerns about the expected acquisition, use, and level of the target language. They question if there is sufficient learning, especially for the productive skills (Milla Lara & Casas Pedrosa, 2018), and state that it is difficult to get learners to use only the target language (van Kampen et al., 2020). Pupils' attention wanes if their L2 level is too low to enable comprehension (Pladevall-Ballester, 2015), and teachers of CLIL subjects are sometimes insecure about their own L2 level (Iwaniec & Halbach, 2024; Oxbrow, 2018; van Kampen et al., 2020).

Finally, teachers state the need for more pre-service and/or in-service CLIL methodology training (Gálvez Gómez, 2021; McClintic, 2022; San Isidro & Lasagabaster, 2019), especially

student-centered methodology in order to cater to diversity (Pérez Cañado, 2023), and more institutional support in general (Pladevall-Balester, 2015).

The perceptions of the CLIL language teacher and the CLIL subject teacher sometimes diverge. While subject teachers are concerned that the content learning will suffer if the classroom language is oversimplified (McClintic, 2022; Seikkula-Leino, 2007), and that the content goals and standards will be lowered due to the difficulty of learning through the target language (Wegner, 2012), English teachers worry that content learning will come at the expense of linguistic accuracy (Banegas & Cerrato Will, 2022). Subject teachers may struggle with the integration of language and content, particularly the role of language in their subject lessons (Skinnari & Bovellan, 2016), and may consider it unnecessary to make methodological changes to their 'non-CLIL' way of teaching (Iwaniec & Halbach, 2024). As content experts, they may have a deep-set belief that language teaching is not part of their role (Bauer-Marschallinger et al., 2023), or only incidentally (Gierlinger, 2017). This has ramifications for coordination between the two groups of teachers: as "tribal animals," (van Kampen et al., 2020, p. 863) the borders which exist between language and subject teachers impede collaboration between these "territorial beings" (van Kampen et al., 2020, p. 869).

Teacher perceptions of the lower academic achievers in CLIL have been reported in some mixed-ability groups, and also in streamed lower academic ability groups (Schwab, 2015; Schwab et al., 2014; van Dongen, 2017). They echo some of the same teacher perceptions but also introduce some observations particularly relevant for this group of learners. On the positive side, teachers note pupil L2 gains, particularly for oral proficiency, and pupil engagement exceeds expectations (Schwab, 2014, 2015). This engagement is due to the target language being used as a tool to achieve task completion and the 'real' learning of content, rather than as an end in itself (Hunt, 2011). Lower academic achievers benefit particularly from the learner-centeredness of CLIL (Iwaniec & Halbach, 2024) and the extra efforts teachers make to ensure that the challenges of learning through a foreign language are overcome (Schwab, 2015). Weaker learners are especially supported by the judicious use of the L1 to make content accessible and cater to diversity, though this also aids their more proficient peers (Bauer-Marschallinger et al., 2023; McClintic, 2022). They value using code-switching and the L1 to facilitate comprehension, particularly at the start of CLIL, but warn of the risk that it becomes habitual and gradually used more and more instead of less and less with pre-vocational groups (van Dongen, 2017). They note the necessity of mastering a wide range of techniques to stimulate oral communication with and among weaker learners (Lo & Macaro, 2015), and that an investment in resources and teachers' pedagogical knowledge is required for CLIL to achieve some success with underachieving pupils (Seikkula-Leino, 2007). If this success is not achieved, teachers feel that their professionalism is called into question (van Dongen, 2017).

CLIL teachers of these learners note similar challenges as their peers with more academically-inclined pupils, albeit slightly differently. They struggle with the increased workload and time pressure, finding it especially difficult to locate material at the right level especially since CLIL material used with more academic learners is not suitable (van Dongen, 2017). They report that lower achievers need additional scaffolding to complete tasks, which requires more teacher time and attention (Lialikhova, 2018). These pupils are more self-critical and their self-esteem can suffer when they have to face the challenges of CLIL (Seikkula-Leino, 2007). Finally, for a number of reasons, there may even be a lack of teachers willing to implement CLIL with less proficient learners (Gierlinger, 2017; van Dongen, 2017).

5.2B Pupil perceptions in CLIL research

Research into the perceptions of CLIL pupils has shown that they chose the CLIL stream (when there was a choice to be made) for different reasons: the advantages of improved English proficiency for future studies and employment (Bakken & Brevik, 2023; Fernández Agüero & Hidalgo-McCabe, 2022; Mearns et al., 2020; Paulsrud, 2018; Sieben & van Ginderen, 2014), an extra challenge (Mearns et al., 2020; Weenink, 2005), "something new and different" (Mahan & Norheim, 2021, p. 81), but also purely pragmatic reasons such as the location or siblings at the same school (Weenink, 2005). Pupils from a low SES background and immigrant minority pupils also felt that English proficiency would be useful for future job opportunities (Bakkens & Brevik, 2023; Mady & Arnett, 2009; Sieben & van Ginderen, 2014; Sierra & Lasagabaster, 2008).

In general, CLIL pupils were generally quite positive about their improvement in English (Doiz et al., 2014b; Gálvez Gómez, 2021; Lasagabaster, 2017; Zanoni, 2021) or another target language (Coyle, 2013). They indicated intrinisic motivation for learning: they liked the novelty of CLIL learning (Pladevall-Ballester, 2015) and thought it was 'cool' to be able to use English as a lingua franca for travel abroad and for entertainment such as gaming and watching films (Buckingham et al., 2023). They also indicated extrinsic motivation for improving their English through CLIL: for future work and studies, travel, and communication with other people and cultures (Doiz et al., 2014b; Fernández Agüero & Hidalgo-McCabe, 2022; Lasagabaster, 2017; Zanoni, 2021).

Regarding the perceptions of CLIL methodology, the pupils see an 'authenticity of purpose' (Pinner, 2021) in the learning and use of English (Pladevall-Ballester, 2015), and value using authentic materials (Pinner, 2021) and talking about 'real' topics (Coyle, 2013). They sense a high level of CLIL teacher collaboration (Oxbrow, 2018). They appreciate the variety of activities in CLIL lessons and often describe these lessons as more fun and more interesting (Coyle, 2013; Doiz et al., 2014b; Hunt, 2011; Otwinowska & Forys, 2017). They identify certain CLIL teacher skills that help them learn content through language: clear explanations, lots of repetition, patience, and a high L2 proficiency (Padevall-Ballester,

2015) which makes them good language models (Buckingham et al., 2023) and note that the CLIL teachers put more effort into ensuring pupil comprehension of the subject matter than non-CLIL teachers (Iwaniec & Hallbach, 2024). They value being given personalized feedback and detailed assessment guidelines (Pérez Cañado, 2023) and perceive no negative effect of learning content through the L2 compared to the L1 (Lasagabaster, 2017). Pupils also report a classroom environment that is more disciplined (Buckingham et al., 2023) and more conducive to engagement, learning, and communication, resulting in a feeling of pride (Coyle, 2013).

There are also less positive pupil perceptions of CLIL in a variety of contexts. Perhaps unsurprisingly, the most common of these is a lack of comprehension. Some pupils report that they frequently do not understand the target language (Coyle, 2013; Doiz et al., 2014b; Otwinowska & Forys, 2017; Pladevall-Ballester, 2015; Zanoni, 2021) or the content—especially STEM subjects—when delivered in the L2 (Barrios & Acosta-Manzano, 2022; Pihko, 2010, cited in Roiha, 2019; Seikkula-Leino, 2007). They struggle with oral communication, which impedes learning when they have more to say than they can actually verbalize and have more questions than they know how to formulate (Wegner, 2012). Writing is also seen as a particularly challenging skill (Coyle, 2013). Problems with comprehension and communication can lead to language anxiety and insecurity (Lasagabaster, 2017; Somers & Llinares, 2021), resulting in reduced participation (Pihko, 2007) and intellectual helplessness (Otwinowska & Fortys, 2017). Pupils are concerned about their academic performance, perceiving that it is harder to get good grades in CLIL subjects (Lasagabaster, 2017; Paulsrud, 2018), and they fear being sanctioned for language errors (Siepmann et al., 2023).

There are also negative perceptions of CLIL methodology and classroom practice. Pupils in some contexts describe CLIL lessons as 'boring' due to teachers' reiteration of the material to ensure understanding, or a dislike of the topics and grammar-focused activities (Doiz et al., 2014b, p. 137). They perceive an overabundance of teacher talking time, translation, or cognitively undemanding work such as copying (Coyle, 2013), or overly simplified content (Zanoni, 2021). They note a lack of methodological variety and the predominance of teacher-frontedness (Pérez Cañado, 2023), especially in the transition from primary to secondary school CLIL (Oxbrow, 2018). They feel they do not get enough personal feedback or assistance from peers (Pérez Cañado, 2023), although some say they are afraid of negative judgement from peers (Siepmann et al., 2023) or, conversely, that their own learning is hampered by the low level of the other pupils (Lasagabaster, 2017). There are divergent perceptions regarding the teachers' use of the target language, as some pupils perceive their teachers' L2 proficiency as being insufficient (Mahan & Norheim, 2021; Otwinowska & Forys, 2017; Zanoni, 2021), or are not positive about their CLIL teachers who insist on only using the L2 (Siepmann et al., 2023). Pupils are dissatisfied with the books and materials (Siepmann et al., 2023), calling them boring (Doiz et al., 2014b) or

unsuitable for learning science (Mahan & Norheim, 2021), or insufficient for catering to diverse student needs (Pérez Cañado, 2023).

5.2C Comparative perspectives and gaps in the literature

Within certain CLIL studies which examined both teacher and pupil perceptions, there are some disjunctures. Pupils estimated the level of student-centeredness and the amount of didactic variety to be less than their teachers did (Bauer-Marschallinger et al., 2023, and they estimated that their teachers used the L1 more than their teachers indicated (Siepmann et al., 2023). The pupils also had a more negative perception of their teachers' target language proficiency and communicative skills than the teachers themselves did (Pérez Cañado, 2016, 2023; Siepmann et al., 2023). Pupils also reported serious comprehension and communication problems, while the teachers claimed their pupils were able to follow the lessons and participate (Pladevall-Ballester, 2015).

There were also contradictory expectations: teachers want to simplify concepts and not explain complicated vocabulary, but pupils felt they needed subject-specific vocabulary to understand the reading material or communicate about the topic; these contrary expectations can lead to misunderstanding and frustration (Wegner, 2012).

Within the teacher and pupil groups, perceptions did not always align. English teachers had a more positive perception than subject teachers of the L2 proficiency of the same pupils, possibly because the subject teachers focus more on communication about the content (Milla Lara & Casas Pedrosa, 2018). Pupils also disagreed about the use of the L1 in the classroom: while some felt it was essential, others wanted less (Lasagabaster, 2017).

Just as in the overview of teacher perceptions above, there is a very limited amount of research on how weaker learners in mixed-ability or lower-streamed groupings perceive learning through CLIL. However, most results show similarities with the perceptions of the stronger learners, especially from a study of German *Hauptschule* pupils in short CLIL modules (Schwab, 2015; Schwab et al., 2014). These less academic pupils note English proficiency improvement and growth in their communicative competence, and feel that learning English has broadened their cultural horizons. They like the use of authentic material in CLIL and appreciate the variety of activities, especially the fun ones. They maintain that they learn more in the CLIL subject lessons than they would have if taught in German; the researchers speculate that this may be because of the extra effort and scaffolding provided by the CLIL teachers.

Like their more academic peers, the less academic learners also struggle with comprehension and oral proficiency (Pladevall-Ballester, 2015) and related difficulties learning content (Lasagabaster, 2017), Some felt it was too challenging (Schwab, 2014, 2015) and simply gave up (Coyle, 2013). Otwinowska and Fortys (2017, pp. 473-474) summed up the

problem of cognitive overload for weaker learners in CLIL: "…their cognitive capacities are very likely to be impaired due to prolonged exposure to incomprehensible output… The results are a clear indicator that CLIL may not work for all learners because some cannot cope with its cognitive demands". Similarly, Zydatiß (2012, p. 27, *emphasis added*) concluded that involving *all* learners in CLIL may "adversely affect their academic success at school".

Of particular relevance for the present research is the preliminary study by Denman et al. (2013), conducted in five Dutch pre-vocational schools, which combined interviews and surveys to explore teacher and pupil perceptions of effective CLIL practice. Although generally optimistic and enthusiastic, the participating teachers acknowledged their pupils' limited English proficiency and short attention spans, noting the consequent need for brief, varied, and creative activities that allow for personal expression and learner choice. They emphasized that pre-vocational pupils must perceive a clear purpose for language use, ideally linked to real-life situations or vocational relevance. The European and International Orientation (EIO) component—especially projects involving authentic international contact—was viewed as a powerful means of enhancing this sense of authenticity, though teachers also recognized that developing a well-structured EIO program requires considerable time and coordination. Teachers highlighted the importance of a strong, collaborative CLIL team and ongoing professional development, and stressed that lessons should prioritize communication in English, with a focus on meaning and fluency rather than on linguistic accuracy alone.

The pupils in that study largely echoed these views: they valued teachers who spoke English consistently, encouraged them to do likewise, and provided clear corrective feedback. Like their teachers, they preferred active, varied, and creative lessons over predictable, textbook-based work, describing 'learning by doing' as both enjoyable and effective. More than 70% indicated that they would recommend the CLIL stream to others. Overall, the study suggests that successful pre-vocational CLIL can be both motivating and linguistically effective, provided it is supported by skilled, flexible teachers and sustained institutional commitment.

In sum, while this overview highlights some of the positive and negative perceptions of pupils and their teachers, these can show considerable variety depending on context (Doiz et al., 2014b). Activities which motivate or discourage pupils in different contexts and at different levels may be very different, and the dearth of research into pupil and teacher perceptions of CLIL for academically-challenged learners in different contexts makes it difficult to draw overarching conclusions concerning their emotions, attitudes, and experiences in CLIL. The current study aims to contribute to the CLIL knowledge base about this group of learners and their teachers.

5.3 The Dutch pre-vocational CLIL context

As outlined in Chapters 1–4, the Dutch secondary education system is highly streamed, with pupils placed at the end of primary school into one of several tracks according to teacher assessment and scholastic aptitude. The pre-vocational track (*voorbereidend middelbaar beroepsonderwijs*, *vmbo*) serves pupils in the least academic, four-year track and prepares them for upper-secondary vocational education (*mbo*). It comprises four levels—from *basisberoepsgerichte leerweg* (basic vocational) to *theoretische leerweg* (theoretical)—and typically serves pupils who benefit from more practical, hands-on approaches to learning.

Bilingual education based on the Content and Language Integrated Learning (CLIL) approach (*tweetalig onderwijs*, *TTO*) was originally confined to higher academic tracks, but since 2009 a growing number of *vmbo* schools have introduced CLIL programs using English as the medium of instruction. These pre-vocational CLIL programs differ from the academic ones described earlier in several respects: they are non-selective[14], open to all pupils regardless of English level or prior achievement; they have reduced language-exposure targets and adapted proficiency outcomes; and they emphasize accessibility and inclusion rather than elite enrichment. At the time of the data collection (2012–2014), pre-vocational CLIL was still in its early developmental phase: only about fifteen schools nationwide offered a *vmbo* CLIL stream or whole-school model, and most were still experimenting with CLIL pedagogy, materials, team-teaching arrangements, and the development of teachers' own English proficiency.

5.4 Research questions and methodology

With the previous literature review in mind and considering the research gap regarding less academically-inclined CLIL learners in general, and specifically pre-vocational CLIL pupils in the Netherlands, the current study aims to answer the following research questions:

RQ1 How do pre-vocational CLIL teachers in the Netherlands (RQ1a) and their CLIL pupils (RQ1b) perceive CLIL classroom practice?

RQ2 What recommendations do these teachers and their pupils make for the improvement of CLIL?

14 The situation regarding selectivity has changed slightly since 2014: a few pre-vocational CLIL schools now impose minimal selection criteria (van Dongen, 2017) and some higher educational tracks have reduced selection criteria (Onno van Wilgenburg, personal communication, May 10, 2024).

5.4A Research design

The current study is part of a larger longitudinal research project which collected data from 2012–2014 which also includes purely quantitative studies. The present study uses a mixed-methods design with a multiperspective approach and single-moment data collection for four instruments: a CLIL teacher survey, a CLIL pupil survey, several face-to-face interviews with non-CLIL pupils, and extended interviews with 15 focus groups, one from each CLIL class in the study. It is primarily an across-cohort comparison, as two groups of frontline stakeholders (teachers and pupils from the same schools) are compared, but also a within-cohort comparison to a certain degree, as pupils in six different schools, four different pre-vocational educational streams, and three different class years are compared.

5.4B Participants

The present study involved the same schools, teachers, and pupil cohorts as those described in the preceding empirical chapters, ensuring continuity across the mixed-methods design. Six *vmbo* schools participated: three all-CLIL schools (where all classes had some subjects taught through CLIL) and three dual-stream schools offering a CLIL stream alongside a regular Dutch-medium stream. Together these represented the full *vmbo* spectrum and a range of CLIL models across the participating schools.

Teachers. Fifty-three teachers participated, including subject-matter specialists and English teachers working within the CLIL framework. Teaching experience ranged from 1 to 41 years ($M \approx 12$ years), yet CLIL-specific experience was limited to the first three years of pre-vocational implementation. Most teachers had joined the initiative with little prior exposure to bilingual methodology. Their self-reported English proficiency ranged from intermediate (CEFR B1) to (near-)native (C2), and many expressed both enthusiasm and uncertainty as they adapted lessons, materials, and classroom interaction for this new learner group.

Pupils. The 267 participating pupils (aged 12–16) represented all four *vmbo* levels and roughly balanced gender groups. About one-third reported a home language other than Dutch—most commonly Turkish or Moroccan Arabic—reflecting the linguistic diversity typical of (semi)urban *vmbo* settings. As noted in earlier chapters, entry to these CLIL programs did not depend on test scores or recommendations: pupils self-selected into CLIL, often out of curiosity, parental encouragement, or a desire to improve their English, but also for reasons unrelated to CLIL or English. Consequently, classes were heterogeneous in English proficiency but relatively homogeneous in general academic ability, reflecting placement within specific *vmbo* subtracks. At the time of the data collection, all schools had a maximum of three years of CLIL experience at this level, making this study among the first to capture stakeholder perceptions from the formative phase of pre-vocational bilingual education in the Netherlands.

- **School A** (3 groups participating, years 1, 2 & 3), with both a CLIL and a non-CLIL stream, only offered the most academic pre-vocational level (*vmbo-t/mavo)*, in addition to also housing the higher academic tracks where a CLIL stream had already existed for a decade. Teachers in the pre-vocational stream could benefit from the in-house CLIL knowledge and experience of their more experienced colleagues; the pre-vocational CLIL teachers were expected to use English almost exclusively. The pre-vocational CLIL stream pupils were in a combination class with a higher academic level (*havo)* for the first two years, allowing for shifting between levels.
- **School B** (1 group, year 2), an urban school with a highly multicultural and multilingual pupil population, also offered a CLIL and a non-CLIL stream at only the most academic pre-vocational level.
- **School C** (2 groups, year 2), a suburban school, also offered CLIL and a non-CLIL stream at the most academic pre-vocational level. The CLIL stream had an on-off modular structure throughout the school year, with 4 to 6 weeks of all subjects (with the exception of the modern languages Dutch, German, and French) being taught with the CLIL approach, followed by 4–6 weeks of the same subjects being taught in Dutch, but not repeating the content. This pattern was repeated throughout the school years.
- **School D** (4 groups, years 1, 2 & 3), a suburban school, was CLIL-only and offered pre-vocational education at all of the pre-vocational levels, sometimes with mixed-level groups and including pupils with diagnosed learning difficulties. It was the first pre-vocational school to offer a CLIL stream, modelled after the school's CLIL stream in the higher tracks (at a different location). School D's distinguishing CLIL program feature was a focus on cross-curricular projects.
- **School E** (3 groups, years 1 & 2), a suburban school, was also CLIL-only and served the less academic pre-vocational levels. School E had extra school hours for self-directed or teacher-supported completion of assignments, promoting itself as a 'no homework' school.
- **School F** (2 groups, year 1), an urban CLIL-only school, served only the least academic pre-vocational levels in both single-level (*basis* and *kader*) and mixed-level (*basis-kader*) classes. The CLIL subject teachers used both Dutch and English, with particular emphasis on basic English vocabulary and word lists, calling their approach 'CLIL lite'.

With the exception of School C, which used the CLIL approach for nearly all subjects but only half of the time, the schools offered three or four school subjects in the CLIL stream for the first two years of the four-year program, in addition to English as a Foreign Language (EFL) classes. In the third and fourth years this was reduced to one or two non-exam subjects.

In short, considering various characteristics of the six schools—CLIL/regular streams versus all-CLIL, pre-vocational level, school years, CLIL subjects, variation in the CLIL ap-

proach, and whether or not there was a pre-existing CLIL program in the school's more academic tracks—it is clear that each school had a different combination of characteristics, and no two schools were entirely comparable.

5.4C Instruments and data collection

Three theory-informed and purpose-built instruments—questionnaires and focus group interviews—were used to measure stakeholders' perceptions of various aspects of the CLIL programs. Some contained questions intended to elicit essential background information relevant to the research questions.

CLIL Teacher questionnaire (RQ1a and RQ2)

A questionnaire was used to elicit the perceptions of the CLIL teachers (n=53), as it was not possible to conduct individual interviews. The content of the questionnaire was informed by prior research on key dimensions of CLIL implementation, including teachers' use of the target language, pedagogical balance between content and language, affective stance, and perceived affordances and constraints of CLIL teaching (e.g. Coyle et al., 2010; Dalton-Puffer, 2007; de Graaff et al., 2007; Pérez Cañado, 2016). A pilot version was used with teacher trainee students and subsequently refined (see Appendix 5, Attachment 5.1). The final version consisted of biodata items and three open questions plus a self-rating scale: what are the most positive aspects of CLIL teaching, what are the most negative aspects, and what two things would most improve the CLIL stream teaching at your school? The self-rating scale collected background data on the teachers' own use of the target language for five different language use areas. These were chosen based on informal interviews with student teachers in a pre-vocational teacher training program: classroom management, instructions, subject content, language feedback, and personal interaction. Data were collected from the teachers in the second year of the longitudinal study so that all teachers would have had at least one year of CLIL teaching experience.

CLIL pupil questionnaire (RQ1b)

Similar to the teacher questionnaire, there were several open questions (on the longer biodata questionnaire; see Appendix 2, Attachment 2.1) asking pupils to name the most positive and the most negative aspects of CLIL learning, plus a 4-point Likert scale question about whether, in retrospect, they would choose the CLIL stream again. The open-ended questions were designed to elicit pupils' own evaluations of CLIL, in line with previous research emphasizing the importance of learner perceptions, motivation, and affect in CLIL contexts (Doiz et al., 2014b; Lasagabaster, 2011). The retrospective choice question has been widely used in CLIL research as a proxy indicator of overall satisfaction and perceived value of the program (e.g., Denman et al., 2013; Schwab, 2014). Data were collected from the CLIL pupils in the first few months of the first year of the longitudinal study, so the first-year pupils had only had a few months of CLIL.

CLIL pupil focus group interviews (RQ1b and RQ2)

Fifteen focus group interviews of about 40 minutes were held with a group of four to seven pupils from each of 15 of the CLIL classes in the research sample (n=70), about 25% of the 267 CLIL pupils, who were chosen by their teachers to take part in the focus groups. The teachers were asked to choose pupils with a range of abilities and school motivation, language background, and gender, who were generally representative of the class and who would be honest and forthcoming in their responses. These semi-structured interviews consisted of 13 questions (Attachment 5.2), formulated with the help of student teacher trainees who were teaching at CLIL *vmbo* schools, and were conducted in the focus group pupils' own schools. The questions targeted dimensions frequently examined in qualitative CLIL research, such as learner enjoyment and frustration, perceptions of language use, classroom interaction, and suggestions for improvement (e.g., Dalton-Puffer, 2007; Lasagabaster, 2011; Schwab, 2014). The questions were the same for all groups; follow-up questions were limited to requests for clarification or exemplification. They were designed to elicit the pupils' perceptions of some of the main characteristics of CLIL teaching and learning, as well some background information about their reasons for choosing the CLIL stream. These interviews were held at the end of the first year of the larger two-year research study, meaning that the year 1 pupils had had a full year of CLIL, the year 2 pupils two years, and the year 3 pupils three years. The participating focus group pupils were assured that their comments would be made anonymous, not shared with their teachers during their time at school, and would only appear in the final research publications. The focus group pupils were given the choice of speaking in Dutch or English; all 15 groups chose to speak in Dutch "*omdat ik dan precies kan zeggen wat ik bedoel*" ["because then I can say precisely what I mean"]. Approximately 600 minutes of focus group interviews were recorded and transcribed verbatim, yielding a final document of about 20,600 words which was then translated into English.

5.4D Data collection procedure

Permission/compliance was obtained in the following ways: for the pupils, the participating schools notified parents in advance of the study and gave them the opportunity to withdraw their child from participation in the study; none did. For the teachers, participation was voluntary but the CLIL coordinators at each school ensured a high rate of return by promising the teachers that the questionnaire would be short and easy to complete in either Dutch or English.

The qualitative methodology used for the other three instruments was thematic analysis (Braun & Clarke, 2022), primarily in an inductive, experiential framework using *in vivo* coding, but also deductive in that some coding was led by prior research and theory. This was chosen so as to best identify and group the main themes and categories of the results of the three CLIL instruments—teacher and pupil questionnaires and the focus group interviews—and to compare those themes, categories, and responses from the different

perspectives. Data analysis and coding focused on identifying patterns and themes in the participants' verbatim oral and written responses. It followed Braun and Clarke's (2006) six prescribed thematic analysis steps: familiarization (including transcribing), coding, including both initial coding and focused coding (Charmaz, 2006), generating themes and categories, reviewing (and revisiting) themes, defining themes, and writing up. In all instruments, data saturation was achieved by ensuring that all words, phrases, and/or sentences were coded at least once.

The teacher questionnaires were distributed by the school's CLIL coordinator and collected about a week later in a sealed envelope to ensure privacy. The teachers thus had some time to think about their answers, unlike the pupils, who were under some time pressure to complete their questionnaire. There was some pre-coding based on the general results from the literature review, and the two overarching themes were 'positive' and 'negative', corresponding to the first two questionnaire items. Teachers were given the option to answer in either Dutch or English, so answers in Dutch were translated into English. The results from the teacher questionnaire were then coded and grouped into categories and in some cases into sub-categories. The teachers' answers within categories were also examined for patterns based on the teachers' role (English or subject teacher) and the pre-vocational level at which they teach (more academic or more practical).

The data from the CLIL pupil questionnaires were entered into SPSS and coded inductively, as there were no premeditated codes or categories based on prior research or theory. Pupils generally wrote only a few words, which simplified the coding process. Initially (and coincidentally) there were 31 codes each for 'most positive' and 'most negative'. Codes and data were subsequently re-examined and combined into categories, keeping the main themes of positive and negative.

The 15 focus group interviews were much more extensive than the teacher or pupil questionnaires and resulted in much more data. All focus groups were given the choice of interview language (English or Dutch) and all chose Dutch. The interviews were recorded, transcribed, re-ordered in two ways (by pre-vocational class and by interview question) and entered into the software program Atlas.ti (ATLAS.ti Scientific Software Development GmbH, 2022) for initial coding. This process of coding and the subsequent process of categorization and theme development was driven by the data content itself, rather than being guided by prior theoretical constructs or prior hypotheses (Braun & Clarke, 2022). Data saturation was achieved through assigning one or more codes to all sections of the data, which captured the pupils' own perspectives, emotions, and experiences.

5.4E Data analysis procedures

The initial coding resulted in over 100 codes, which were consolidated into 23 codes in the iterative process of examining both latent and semantic meaning in the coded data.

As patterns became more recognizable, these were aggregated into categories not only to make the data analysis more manageable, but also to enable easier comparisons with the data from the pupil and teacher questionnaire data. Ultimately, four major themes were generated, one of which addressed the first research question because it was related to a situation (choice for CLIL) prior to the start of CLIL, whereas the other data were related to experiences, perceptions, and emotions after the start of CLIL.

During the processes of study design, data collection, analysis, and report writing, the checklist *Consolidated criteria for reporting qualitative research* (COREQ) was used where applicable to ensure adherence to good practices in qualitative research reporting (Tong et al., 2007).

The quotations from teachers and pupils presented here were selected in most cases because they were representative of a wider group. An attempt was also made to highlight diversity by giving voice to a range of experience, CLIL subjects, schools, class years, pre-vocational levels, and pupil home languages.

5.5 Results: teacher and pupil perspectives on CLIL

5.5A Teacher perspectives (questionnaire data)

In total, 53 teachers in the six pre-vocational schools in the research project completed all or most of the teacher questionnaire. These teachers were either CLIL subject teachers or English teachers of the same classes of CLIL pupils in the study, constituting about 80% of the total number of these pupils' CLIL teachers. The background question about L2 use was answered by 48 teachers; it asked them to estimate how often they used the target language (as opposed to using Dutch) for five different language functions in the educational context: classroom management, instructions, subject content delivery, language feedback, and personal interaction. The results are given in Figure 5.1, with a distinction made between English teachers (dark bars) and subject teachers (light bars); not all teachers filled in an answer for each language function, and half of the teachers from School E failed to return their questionnaire. The number in each language function category indicates the number of English teachers and/or subject teachers who estimated their use of the target language at that percentage.

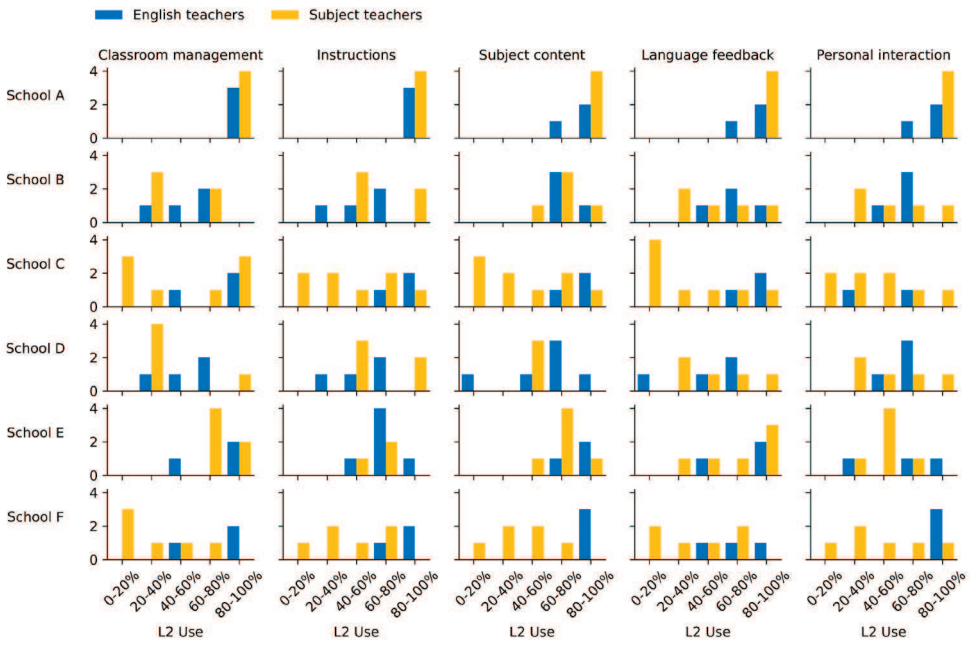


Figure 5.1: Teacher self-report results: estimate of their percentage of English use for five classroom language functions

As shown in Figure 5.1, there are some discrepancies among the schools: not only the number of English and subject teachers in each of the six schools and different self-estimations, but also different target language use across language functions. Although the self-estimations are necessarily subjective and there is no way to determine whether the categories were temporally equal, a few patterns emerge. First, the number of teachers reporting their use of English as 60% or less was nearly identical (17 to 19 teachers) across all language functions, although there is a considerable difference between schools. The subject and English teachers from School A—which had a more theoretical pre-vocational program and years of experience with CLIL in the higher streams—reported the highest target language use across all language functions. The subject teachers, some of whom also taught in the school's more academic *havo* or *vwo* CLIL stream, reported an even higher use of English than the English teachers. Conversely, subject teachers at schools D and F (mostly or entirely lower pre-vocational classes) reported the lowest target language use, with two subject teachers at School F reporting respectively 20% or less and 40% or less target language use across all five language function categories. With the possible exception of School A, the English teachers generally used the target language more often than the subject teachers across all function categories.

Theme: CLIL is enjoyable

Regarding RQ1, of the 53 teachers, 47 answered the first open question, *What are the most positive aspects of teaching pre-vocational CLIL?,* with 11 teachers giving two or three answers, for a total of 56 answers. These were coded and categorized in the iterative process described above, resulting in the following number of answers in categories and sub-categories (Table 5.1). It should be noted that, contrary to the recommendation of Braun and Clarke (2022), frequency counts and some quantitative language is used in the reporting and analysis of the results. This is to indicate the relative frequency of some general teacher perceptions as opposed to others, though all respondents' perceptions are equally valid and relevant.

Table 5.1: Categories of teacher responses, Theme 'CLIL is enjoyable'

Category (Theme: positive)	**# responses (total 56)**	**sub-category**
(Pupil L2 growth) Development of pupils' proficiency in English	25	
Teacher affect (positive)	13	General teacher enjoyment (5) Positive challenge/motivation for the teacher (4) More satisfying teacher-pupil interaction (4)
Positive aspects of CLIL didactics	7	
Pupil affect (positive)	4	
Teacher L2 language improvement	4	
Other pupil-related aspect	3	Pupils develop more cognitive skills (1) Good for their future (1) Pupils become more internationally-oriented (1)

Positive points directly concerning the pupils comprised over half of the aspects (32) mentioned by teachers, while the remainder concern the teachers themselves (24). The positive features noted by these teachers thus pointed to benefits for both the pre-vocational CLIL pupils as well as for their teachers.

The development of pupils' English language skills was mentioned most as a positive aspect of pre-vocational CLIL teaching by teachers in all six schools. The most frequently-used words were *improve, progress, speak/oral proficiency, vocabulary*.

> Seeing pupils not only learn to excel at the language but also acquire the culture—they become more international, they break out of the borders of having only one culture. (English teacher, School A)

> The pupils learn more English and become more used to speaking English. (Economics/ICT, School D)

Progress made by pupils in English is noticed by external subject experts. (Biology, School C)

The second-most mentioned category was that of a positive teacher affect (enjoyment, positive challenge, motivation, or interaction with pupils), sometimes in combination with a positive pupil affect. This was noted by teachers in almost all schools, with the words *motivation, positive challenge,* and *interesting* frequently used:

Committed teachers, high teacher motivation, slightly more motivated pupils (Art, School A)

More positive challenges and more possibilities for interesting lessons with CLIL, this increases teacher motivation (History and European/International Orientation, School A)

It's an extra positive challenge for pupils and teachers (Social work studies, School F)

A wider range of pedagogical activities, and specifically CLIL activities, were mentioned by teachers in all schools. Here the words *authenticity, hands-on,* and *didactic activities* were cited often:

Authenticity of language learning in a variety of activities (English, School E)

Pupils are well-engaged with CLIL, especially with 'hands-on' activities (Geography, Economics, Religion, School F)

The CLIL activities, more didactic activities (History/Social Studies, School C)

They also cited the positive challenge and the importance of the teacher-pupil bond with this type of pupil:

Enabling students to use and develop their language skills more, even though their level of learning is not the same as the higher educational streams (English, School C)

The pupil-teacher relationship is even more important, especially with 'special needs' pupils. (English, School B)

In CLIL, there is a closer relationship between teacher and pupil. (English, School F)

The category 'Teacher language' was mentioned only by subject teachers, whose own target language level is officially required to minimally B2 but is most often C1. English teachers in the Netherlands must demonstrate CEFR level C2 (Cambridge CPE) to gain teaching certification, so it is perhaps unsurprising that they did not mention language improvement as a positive aspect of teaching in the CLIL stream.

Theme: CLIL is frustrating

The second question, *What are the most negative aspects of pre-vocational CLIL teaching?* was answered by 44 of the 53 teachers, yielding 53 answers in seven categories (Table 5.2). There was some overlap with the categories identified in the answers to the 'positive aspects' question above, and a few answers fell into two categories.

Table 5.2: Categories of teacher responses, Theme 'CLIL is frustrating'

Category (Theme: negative)	**# responses / codes (total 53)**	**(sub-categories)**
Pupil English level	25	Level too low for learning through English (14) Subject issues: subject learned less well, subject too difficult for *vmbo* CLIL (9) Huge range of proficiency levels in the class (2)
Organizational limitations	10	Not enough preparation time (6) Extra work: trips, excursions (2) Hard to find materials (1)
Pupil behavior	8	Pupils refuse to speak English (4) Lack of discipline, concentration; bad behavior (4)
Dutch language issues	4	Less time for Dutch, bad for pupils (3) Final exams are in Dutch (1)
Teacher English level	3	Teachers don't use enough English (1) Teachers' low English level (1) Stress factor using English (1)
Teacher affect	2	Lack of motivation (1) Not as humorous in English (1)
Pupil affect	2	Low self-image (1) Unmotivated (1)

As with the positive points, the negative aspects mentioned by the teachers relate to both the pupil experience and the teacher experience, but here there are more pupil issues mentioned.

Also related to RQ1, the most frequently mentioned negative aspect concerned pupils' English proficiency. Teachers across all six schools reported that limited language proficiency constrained subject understanding, particularly in CLIL subject lessons. Recurring phrases in this category included *low language level*, *don't understand*, and *too difficult*. As one teacher noted,

> The pupils' language level is too low, so they don't grasp the subject. (Math, School B)

Another teacher captured this as a mismatch between linguistic and cognitive demands:

> Big difference between language level (low) and content level (higher). (Biology, School C)

A comparison of the positive and negative answers from individual teachers reveals that some comments about the pupils' language level fell under both themes: while being encouraged by the English language improvement, they were also skeptical about some pupils' ability to follow the subject lessons or learn the content.

> The difference between the language levels of the pupils is very high. Some pupils can follow everything, but I notice that not all the pupils understand English. (Physical education, School A)

And some teachers felt that their (subject) colleagues' English level was insufficient:

> The level of English language skills of the subject teachers is not always sufficient when it comes to teaching the language aspects (English, School C)

Organizational limitations were reported by teachers in four schools and were primarily related to time pressure and workload. Recurring words in this category included *time*, *workload*, *preparation*, and *materials*. Teachers indicated that limited preparation time constrained the extent to which CLIL-specific activities could be integrated into lessons. As one teacher explained,

> Lack of preparation time means we have less CLIL focus in the lessons. (Math, School C)

Several teachers also mentioned difficulties in sourcing suitable CLIL materials:

> It's very difficult to find the right materials and books. (Geography/History, School F)

The responses in the category 'Dutch language issues' reflect the ambivalence of some teachers regarding bilingual education for these pupils, focusing either on the need for pupils' Dutch-language proficiency, or the switch back to Dutch for most CLIL subjects in years 3 and 4:

> Many pupils aren't even able to speak Dutch properly, so why are we teaching in English? (Dutch, Math, Geography, History, School D)

> Doubts about whether it is going to work out for the pupils if it's not in Dutch—isn't it too difficult/confusing? (Physical education, School C)

> The pupils have a hard time adjusting back to Dutch for exams in their final year. (English, School E)

Finally, several teachers mentioned aspects related to pupils' affect which could also be considered typical characteristics of some pre-vocational pupils:

> The pupils often feel that they are not smart enough and they say that up front, without even trying first. (English, School B)

> Pupils' lack of motivation, apathy, and lack of understanding the importance of education (English, School F)

> Behavioral issues are more common, maybe due to pupils' frustration with not understanding. (Social studies, School F)

Theme: CLIL can be improved

The final open question for the teachers was: *What two things would make pre-vocational CLIL even more effective for you/your pupils/your school?* (RQ3). Teachers were free to address issues related to teaching, learning, organization, policy, etc. A total of 47 teachers made 76 suggestions, with several spanning more than one category.

As with the positive and negative CLIL aspects, the results were coded and examined to discern categories. Four main categories were identified, each with sub-categories (Table 5.3):

Table 5.3: Categories for teacher responses, Theme 'CLIL can be improved'

category	**# responses** (76 total)	**Sub-categories**
CLIL	28	Teacher collaboration & communication (16) CLIL training (12)
Rules and policy	24	Various rules/policy changes (17) Selection of pupils for CLIL (9)
Support	19	More time (11) Better materials (5) Infrastructure (2)
Language	12	Improve teacher language (7) Improve pupil language (4)

The largest category of suggested improvements concerned CLIL pedagogy, particularly the need for stronger teacher collaboration and communication, as well as additional CLIL training. Teachers in all six schools made suggestions in this category. Notably, relatively few suggestions came from School E, an all-CLIL school with daily mandatory teacher meetings, whereas teachers at School F ('CLIL lite') repeatedly emphasized the need for improved teamwork.

Suggestions referred both to collaboration among CLIL subject teachers and to closer cooperation between subject teachers and English teachers. For example, one teacher called for

> More teamwork, i.e. integrated projects with two or more subjects and teachers. (Geography, Economics, Religion, School F)

Others stressed the need for clearer, shared CLIL frameworks in joint practice:

> Clear CLIL guidelines that are actually used in practice by the teachers working together—English teachers and subject teachers. (English, School F)

Many teachers in five of the six schools stated a desire for more CLIL training. The two schools which had offered a CLIL stream the longest (Schools A and D) had the most teachers expressing this, while not a single teacher from School F mentioned CLIL training.

> More training during teaching, and not just at the start (Art, School A; Biology, School D)

> Give teachers help in adjusting their own English language level to fit the pupils' needs and capabilities (Biology, School C)

Suggestions related to Rules and Policy fell into two sub-categories: general policy, and the question of whether pupils should be selected. There were conflicting policy suggestions related to the use of English in the classroom:

> All teachers have to be on the same page about teaching CLIL lessons in English. (Biology, School D, School F)

> Pupils don't have to use English all the time in the classroom if they are working together to understand the subject. English-only is not essential. (Math, School A)

Other policy-related suggestions were broad in scope and addressed multiple levels, including classroom practices, school organization, and national assessment structures. Teachers referred to issues such as curricular time allocation, institutional sustainability, and the communication of program expectations. Several comments emphasized the importance of clearly preparing pupils for the demands of CLIL from the outset, as one teacher noted:

> We have to be really clear to the pupils from the very beginning that we are a bilingual school with CLIL. (Biology, School D)

Other suggestions reflected organizational and financial considerations, for example:

> More pupils in CLIL: this would make the CLIL stream more cost-effective and increase acceptance by non-CLIL teachers. (English, School A)

A smaller number of teachers proposed more far-reaching policy changes, such as aligning assessment structures more closely with CLIL provision.

> Final exams in English instead of Dutch for CLIL subjects (English, School E)

Regarding pupil selection, the national Standard for pre-vocational CLIL at the time stated that it was intended to be inclusive and accessible to all, yet teachers at four schools felt that certain selection procedures should be introduced. This was not expressed by any teachers from School C, with its modular 'CLIL weeks—Dutch weeks' system, or from School F, the self-proclaimed 'CLIL lite' school with its focus on basic vocabulary acquisition.

> Make a condition that pupils already have some knowledge of English / Entrance placement test: minimum starting English level (Art, School A; Social studies, School D; English, School E)

> Special CLIL classes for the highest pre-vocational level. No CLIL stream for the lowest pre-vocational level. (English, School D)

> Select only pupils who are motivated and chose for CLIL themselves. (Art, School B)

Support in terms of more time for teachers was mentioned by many teachers in four schools, especially in School A where the teachers had the example of higher-level CLIL stream to follow:

> More preparation time (Geography, Music, History, Math, English, School A; Geography, English, School B; English, Math, School C; Physical Education, School F)

Finally, suggested improvements related to language fell into two categories: teacher language and pupil language. With regard to teacher language, teachers referred not only to subject-specific language use in CLIL lessons but also to their own general English proficiency. As one teacher noted:

> Teachers need more general English training. (Biology, School D)

Some teachers also located language-related challenges earlier in the educational trajectory, pointing to the role of primary education in preparing pupils for CLIL.

> Primary schools should incorporate more English into their curriculum so pupils are better prepared. (English, School C)

5.5B Pupil perspectives (questionnaire data)

The 266 CLIL pupils filled out a biodata questionnaire which contained three items used in this qualitative study (RQ1b). The questions were in Dutch and nearly all pupils wrote their answers in Dutch.

1. What do you like best about the bilingual education/CLIL program? (open question)
2. What do you like least about the bilingual education/CLIL program? (open question)
3. If you could choose again, would you choose for the bilingual education/CLIL program?

The responses were coded in SPSS because this was also used for the other biodata entry. After coding under either 'positive' or 'negative', the codes were re-examined, combined, recoded, and translated into English. The related themes 'CLIL is enjoyable' and 'CLIL is frustrating' were assigned to allow the most direct comparison with the teacher responses to generally the same first two questions. Coding was relatively simple because most pupils wrote only one sentence, phrase, or single word. A third theme was only for the pupils' results, as it related to their responses to the question whether, in retrospect, they would choose CLIL again.

Theme: CLIL is enjoyable

Question 1: What's the best thing about bilingual education/CLIL? (266 responses)

In the first round of coding there were 31 codes identified for positive pupil perceptions of CLIL (Table 1), which were then recoded into 6 categories: English language learning, Nothing/Everything/Don't know, CLIL/Bilingual education characteristics, School-specific characteristics, Specific school subject, and Challenging yes/no (Table 5.3). Again, al-

though Braun and Clarke (2022) discourage reporting absolute frequencies for category mentions in qualitative analysis, in this case there is added value in giving numbers for the categories because of the huge discrepancy in frequencies. A high frequency count is likely to be indicative of the strength or consistency of a perception category, while the less frequent categories highlight different but still salient facets of the theme which might simply not have come to mind at the moment of data collection.

Table 5.4: Categories for pupil questionnaire, Theme 'CLIL is enjoyable', with counts per year and pre-voc level

Category	# year 1 (n=125)	# year 2 (n=90)	# year 3 (n=51)	TOTAL Year 1,2,3 (n=266)	Of which higher pre-voc (n=168)	Of which lower pre-voc (n=98)
English language learning (positive)	62	49	17	**128**	88	40
Nothing is good	10	5	10	**25**	10	15
Everything is good	12	10	4	**26**	13	13
Don't know	12	6	5	**23**	7	16
BE/CLIL characteristics	16	16	12	**44**	38	6
School-specific	4	2	2	**8**	4	4
Specific BE/CLIL subject	6	-	-	**6**	2	4
Challenge yes/no	3	2	1	**6**	6	-
Total # responses	125	90	51	**266**	168	98

Responses related to the positive aspects of learning and using English were widespread across schools and year groups and were also expressed by pupils with non-Dutch home languages. Pupils frequently referred to the ease and speed of learning English, opportunities for communication, and the international usefulness of English as a lingua franca. The most frequently-used words were *speak/talk, other people/other countries, contact, easy, learn more*. As one pupil explained,

> That you learn English really easily and so you can talk with people from other countries. (Year 1, School E)

Several multilingual pupils explicitly framed English as a global means of communication:

> That it's in English and almost the whole world understands English. (Year 2, Arabic home language, School C)

Pupils also compared their progress favorably to that of non-CLIL peers:

That the CLIL classes know more English than the non-CLIL classes. (Year 2, School C)

However, some of the positive comments seemed to indirectly indicate a lower amount of target language use than might be expected in a CLIL classroom (*emphasis added*):

Sometimes we get to speak English (year 1, School F)

That the teachers *sometimes* also speak English (year 3, School D)

That there is *some* English in every lesson (year 1, School F)

Pupils were also positive about features commonly associated with CLIL, particularly international orientation and experiential learning. Cultural content, school trips and exchange opportunities, and varied forms of classroom work were frequently mentioned and were often framed as both reasons for choosing the CLIL stream and as factors sustaining motivation and enthusiasm over time. Recurring words in this category included culture, trips, groups, variety, and projects.

That we also learn things about England, because I really like England. (Year 3, School A)

The trips abroad and excursions (year 2, School B)

There's a lot of variety. (Year 2, School E; Year 1, School D)

Learning Maths and Biology in English. (Year 1, School A)

Some pupils referred to positive aspects of CLIL that were specific to their own school or micro-context rather than to the CLIL program more generally. These included local organizational arrangements such as smaller class sizes, project-based learning structures, reduced homework, or the presence of international teaching staff. As one pupil noted:

We have teachers from other countries. (Year 1, School E)

Finally, some pupils explicitly framed the level of challenge in CLIL—whether experienced as demanding or as relatively easy—as a positive aspect of the program.

> It's an extra challenge to improve my English (year 1, School A)

> It's fun but it's very hard (year 2, School D)

> It's really easy (year 3, School D)

Theme: CLIL is frustrating

In the initial coding stage there were 31 codes identified as negative pupil perceptions (Table 5). After focused coding these were compacted and recoded into 5 categories: Nothing/everything/don't know, English language, Teacher or subject, Organization/rules, and Not enough English. The first and largest category included the most frequent sub-category, 'nothing'—which is ambiguous because it could mean that the pupil has nothing to say, or that there is nothing negative about CLIL (a reverse positive response).

Categories and counts

Question 2: What do you like least about bilingual education/CLIL? (270 responses; 4 pupils with 2 answers)

Table 5.5: Categories for pupil questionnaire, Theme 'CLIL is frustrating', with counts per year and pre-voc level

Category	# year 1 (n=125)	# year 2 (n=90)	# year 3 (n=51)	TOTAL Year 1,2,3 (n=267)	Of which higher pre-voc (n=169)	Of which lower pre-voc (n=98)
Nothing to dislike	40	30	25	**95**	55	40
Dislike everything	1	2	2	**5**	2	3
Don't know	13	5	5	**23**	9	14
English language (negative)	50	32	11	**93**	61	32
Teacher/School subject	13	9	3	**25**	20	5
Organization/rules or school policies	4	10	5	**19**	16	3
Not enough English/ other pupils	1	2	4	**7**	6	1
Total # responses	122	90	55	**267**	169	98

Aside from the first, more ambiguous category, the most frequently mentioned negative aspect concerned difficulties in understanding and using English, often accompanied by worries about receiving low grades in subjects taught through the target language. This

concern was most pronounced among pupils in Year 1, but was also reported in Years 2 and 3. Across schools, cohorts, and pre-vocational levels, pupils frequently used terms such as *moeilijk* (difficult/hard), *don't understand*, *confused*, *mixed up*, and *no idea*, indicating both linguistic difficulty and broader cognitive overload.

> Speaking English, because I can't do that very well yet. (years 1 & 2, School C)

> That you always get low grades because it is really hard (year 1, School A)

> I don't understand anything and sometimes I don't know what I'm supposed to do/I have no idea what's going on. (year 1, School D; year 1, School E; year 1, School F)

Some pupils also expressed frustration with specific school subjects or with teacher behavior related to comprehension and explanation. Recurring references included a perceived lack of support when pupils did not understand, teacher-dominated talk, and, in one case, the presence of a non-Dutch-speaking teacher. As one pupil explained:

> Sometimes I don't understand but the teacher thinks it's easy and doesn't help you and always makes fun of you. (Year 1, School A)

Another pupil linked comprehension difficulties explicitly to the teacher's language background:

> I don't like Maths in English because then you don't understand it and the teacher doesn't know Dutch. (Year 1, School E)

A smaller number of pupils described lessons as difficult to follow due to extended teacher talk without sufficient opportunities for clarification.

> The teacher just keeps talking and I don't understand anything (year 3, School D)

Pupils in the higher years were sometimes critical of the organization of CLIL, particularly in relation to modular arrangements and the reduction of CLIL subjects and hours in preparation for Dutch-language final examinations. In schools using a modular CLIL approach, pupils contrasted CLIL weeks unfavorably with periods taught entirely in Dutch:

> I like the weeks when everything is in Dutch, but not the CLIL weeks. (Year 2, School C)

Others expressed frustration with the abrupt shift back to Dutch for exam subjects in Year 3:

> I don't like it that starting in year 3 all exam subjects are suddenly in Dutch. (Year 3, School A)

Some pupils described this shift as effectively erasing CLIL from their school experience:

> I don't really notice that there is CLIL, we don't really have it anymore. (Year 3, School D)

Other pupils' limited participation or low proficiency in English was also a source of frustration, particularly when it reduced the use of English in lessons intended to be primarily in the target language. Pupils referred to classmates who did not take CLIL seriously or avoided speaking English, with the result that lessons reverted to Dutch:

> Not everyone is good enough in English and then they just don't participate. (year 2, School B)

> Because then it ends up mostly being in Dutch after all. (Year 3, School D)

Others emphasized that not all pupils adhered to the expectation to use English in class:

Some pupils don't take it seriously. Not all pupils speak English in class when we are supposed to. (Year 3, School D)

Would you choose the CLIL stream again?

The third question on the pupil questionnaire asked pupils to answer the question, *If you could choose again, would you choose the CLIL stream?* using a four-point Likert scale. The results are given in Table 5.6.

Table 5.6: Would the pupils choose the CLIL stream again?

RESPONSE	# Year 1 (n=125)	# year 2 (n=90)	# year 3 (n=52)	TOTAL years 1,2,3 (n=267)	Of which higher pre-voc (n=169)	Of which lower pre-voc (n=98)
Yes, definitely	63	43	31	**137**	95	42
Yes, but it's harder than I thought	33	22	6	**61**	43	18
Probably not	24	19	8	**51**	20	31
Definitely not	5	6	7	**18**	11	7
TOTAL # responses	125	90	52	**267**	169	98

Question 3: Choose CLIL again? (n=267)
In total, 74% of the CLIL pupils answered either 'yes, definitely', or 'Yes, but'. The younger pupils were slightly more positive about choosing CLIL again, but the length of their experience was quite limited. Pupils in the more academic pre-vocational classes were more positive than those in the less academic ones; the percentage of lower pre-vocational pupils saying they would probably not or definitely not choose CLIL again was twice as high as in the higher levels.

5.5C Pupil perspectives (focus group interviews)

Initial choice for the CLIL stream
It was speculated that information about pupils' reasons for choosing the CLIL stream in the first place might add nuances to their perceptions of the CLIL program one or more years later. This background information was gathered during the pupil focus group interviews. These pupils were asked why they chose for CLIL, and whether it was their own choice. The 70 focus group pupils gave 98 answers; some answers contained two reasons. The distribution of the answers in 12 categories is given in Figure 5.2, where the dark bars indicate a deliberate choice for the CLIL stream; the lighter bars indicate that the decisive factor was something other than CLIL. For the deliberate CLIL choices, the most-cited reason was the pupils' perception of English as 'fun' (*leuk)*. Future studies/ employment and the anticipation of a positive challenge or adventure were both named equally frequently, but less than half as often as 'fun'. Nine pupils said it was not their own choice, but that their parents chose for them, thinking more English would be useful or practical. A few pupils mentioned trips and excursions, some pupils thought it would be easy because they thought their English was already good, and one pupil was guided by a primary school's recommendation.

Factors other than the CLIL stream led to the choice for several pupils. Fifteen pupils—nearly all of them pupils from the CLIL-only schools—said that they had no idea they (or their parents) were choosing a school with bilingual education. Their primary reasons for choosing a particular school were a convenient location, an advertised 'no homework' school, siblings or friends attending the same school, or not registering in time for the pupil's first choice of school. In all, about one-third of the pupils said that the CLIL stream or extra English was not a factor in their school choice.

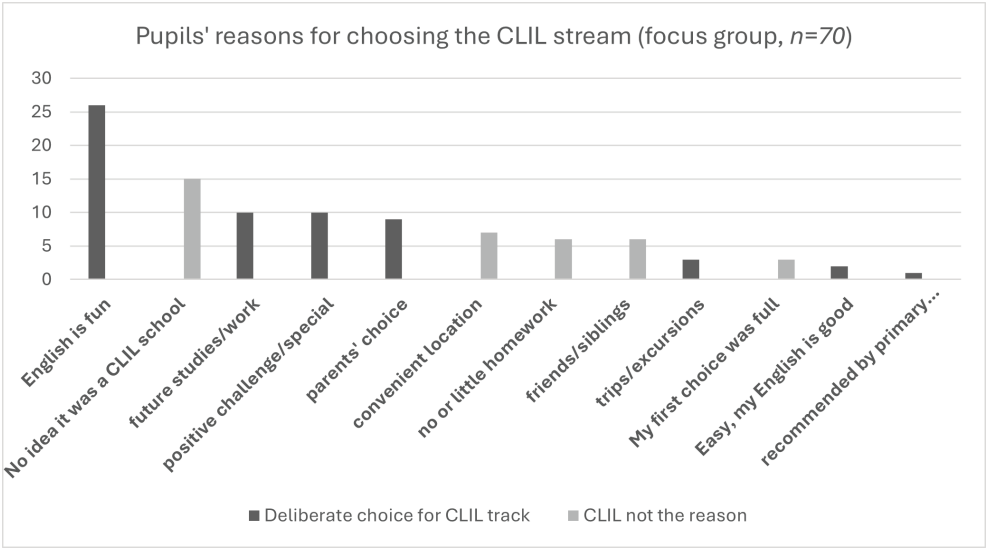


Figure 5.2: Focus group pupils' reason for choosing the CLIL stream

After limiting two-pronged responses to the first reason given per pupil, the answers indicate that 45 of the 70 pupils (or their parents) had consciously chosen for bilingual education. Of these, answers related to the English language were given most often, with the words *fun* and *like English* occurring most frequently:

> I like English and I already use it a bit as a second or third language (year 1, School F)

> It's fun to learn English. (year 3, School D)

> I had always thought that English was a fun language and it would be fun to do it. (year 3, School A)

Some pupils described having chosen CLIL deliberately for the perceived benefits of English for their future studies or employment. This form of future-oriented motivation was expressed only by pupils in the more academic CLIL streams and often involved a combination of parental influence and personal interest. One pupil explained:

> I chose CLIL on purpose, because English is a handy language and if you can speak it well then you get further in life and it looks good on your CV. (School B, Year 2)

Others described how the decision was shaped jointly by parents and pupils:

> My parents said that this would give me more opportunities later, a wider range to choose from, so I went along with this. But I actually chose myself, because it sounded fun and I already liked English at primary school. (School A, Year 2)

A smaller number of pupils framed their choice in terms of longer-term personal aspirations:

> I really want to be an English teacher and I thought CLIL would help. (School A, Year 3)

Other reasons for choosing CLIL related to its cultural and experiential character, including opportunities for international contact, trips abroad, and the perception of CLIL as offering something different from regular schooling.

> If you can speak English well, you can come into contact with people from other countries. (year 2, School A)

> It sounded like this school had a really active CLIL program compared to other schools that had CLIL but no trips abroad like to England, where we went this year. At this school they pay more attention to that. (year 2, School B)

> It's like an adventure, because normally at school you do nothing…it's boring, so I thought it would be fun to do something different from what the others were doing. (year 2, School A)

Of the remaining pupils, 25 reported having chosen the CLIL school for reasons unrelated to CLIL or bilingual education, a pattern that was particularly common in the less academic streams. Some pupils explicitly indicated that CLIL did not factor into their choice at all:

> I thought it was a nice school, but not because there was CLIL. There are extra cooking lessons, so you're doing something and not just writing a lot. More English, I didn't think about that. (year 1, School E)

Others described their school choice as shaped primarily by parental decisions or practical considerations:

> My parents said I had to go to this school because we live nearby. (year 1, School F)

A smaller number of pupils referred to the availability of additional support as a decisive factor:

> Because I am dyslexic and there are more English hours and because you can do all your homework at school. (year 2, School B)

Fifteen pupils reported being unaware at the time of enrollment that they had chosen a CLIL or bilingual education program. This lack of awareness was reported in all three less academic CLIL-only schools and in one of the more academic CLIL-stream schools. Some pupils only realized the extent of English use after starting the program:

> I chose this school because lots of my friends chose it. I didn't know there would be more English. When you're 12 years old you're clueless anyway. After a few weeks I thought wow, there's a lot of English! (year 1, School E)

Others described having chosen the school for practical reasons without realizing that it offered CLIL:

> The school is close to my house. I didn't know it was CLIL. (year 1, School F)

Lack of awareness was also reported across cohorts within the same school:

> I didn't know there would be subjects given in English. That was a surprise. (year 1 & year 3, School D; five pupils)

In some cases, this occurred despite parental involvement in the selection process:

> I looked at the website, also with my parents, but we didn't realize it was CLIL until after I had registered. (year 2, School E)

As with the pupil questionnaire results, the three general themes identified following the iterative cycles of coding and code aggregation of the 15 extensive focus group interviews were *CLIL is enjoyable, CLIL is frustrating,* and *CLIL can be improved.* Because the interviews were much more in-depth than the teacher and pupil questionnaires, the thematic analysis resulted in more categories. Attachment 5.3 provides the full codebook.

Theme: CLIL is enjoyable

The subsections below reflect the main categories identified in the thematic analysis (see Appendix 5, Attachment 5.3).

Improvement in English and perceived progress

Improvement in English was mentioned most frequently by focus group pupils in the more academic CLIL stream. Pupils often evaluated their progress in relation to non-CLIL peers and emphasized that the increased linguistic demands were challenging but manageable. One pupil explained:

> I was helping a [non-CLIL] friend with her English but it was so much easier than what we get; they are just not as far as we are. (year 1, School A)

Others highlighted that CLIL required them to work with material that initially exceeded their perceived language level:

> We have to do assignments about the books we read, which are often above our level of English but we can do it. (year 1, School A)

Several pupils described a rapid developmental trajectory, particularly when English input was sustained across subjects:

> I only started learning English in the CLIL program… in the beginning it was hard… but after a few months you get better and better and now it's easy to speak English. (year 2, School A)

Pupils also referred to increased use of English beyond the classroom, reinforcing their sense of progress:

> We speak English a lot… also outside of the lessons… The more you speak English, the easier it gets. (year 3, School A)

International orientation, exchanges, and cultural experiences

Pupils were particularly enthusiastic about trips abroad, the cultural and socio-cultural dimension of CLIL, and international orientation (EIO) activities. This was especially evident among first-year pupils, for whom international exchanges were still a future prospect. In focus groups, pupils jointly expressed anticipation about traveling abroad and meeting pupils from other countries:

- I think it would be really fun to go abroad.
- But we are going to!
- Yeah, but not until the end of year 2 or year 3.
- In a year or two maybe some foreign pupils will visit us. (year 1, School E)

Some pupils had already experienced forms of international contact, which they described in social and playful terms:

- We sent cards to Finnish pupils and they want us to do an exchange with them in the future and then speak more English. We still have contact with some of them on social media.
- Yeah, everybody says you have a chickie in Finland. [laughter] (year 1, School F)

For pupils in higher years at one school, international trips were described as a recurring and well-structured part of the CLIL program:

> We go to England every year. The first year to Canterbury, the second year to Harwich near Dover, and this year we visit a school in Ash. (year 3, School A)

Visible teacher collaboration and curricular integration

Pupils also referred positively to instances of teacher collaboration and curricular integration that were visible to them. In some cases, this involved coordination between English teachers and subject teachers around language demands in subject lessons:

> Some teachers go to Mr. X (the English teacher) and ask him to pay special attention to some aspect of English that we need for a presentation or something… then the English teacher works with the other teachers or makes word lists for us. (year 2, School B)

Pupils also described coordinated assignments across subjects:

> Some teachers organize things so that we get a combined assignment, and sometimes the lesson content is coordinated for several subjects. (year 2, School B)

Finally, collaboration was experienced through joint activities organized by CLIL teams:

> We had a games day organized by all the CLIL teachers; it was all outside and they made sure we spoke English. (year 3, School A)

Teacher English proficiency and communicative adequacy

In a small number of focus groups, pupils evaluated their teachers' English proficiency positively and emphasized that communicative adequacy mattered more than linguistic perfection. Pupils noted that their teachers' English was sufficient for instructional purposes, including providing feedback:

> Our teachers' English is definitely good enough to give [corrective feedback], also to correct our grammar. (year 2, School A)

Pupils also expressed tolerance and empathy toward teachers' language errors, framing these as normal and even educational:

> Everybody makes mistakes and we also learn from that, so if a teacher makes a mistake it's just normal. (year 1, School A)

In one case, pupils described a reciprocal dynamic in which they supported a teacher with language difficulties:

> Our biology teacher is dyslexic and sometimes makes spelling mistakes on the board, and then we help her. Actually we like that. (year 1, School A)

Consistent target language use and peer norms

Finally, some focus group pupils described positively the consistent use of the target language in CLIL lessons and the explicit encouragement to use English not only with teachers but also with peers. Pupils emphasized that speaking English was actively expected and, over time, became a shared norm:

> All CLIL teachers require us to speak English with each other. And if someone doesn't, then we [the pupils] help each other to keep speaking English. (year 2, School A)

Pupils also described how teachers reinforced this expectation and how pupils themselves took responsibility for maintaining it:

> We speak English with the teachers and also with each other. If we speak Dutch, the teachers correct us… Then they say, 'This is CLIL.' … and if someone still keeps speaking Dutch, then I remind them. (year 2, School A)

Didactic structure and instructional clarity

In some cases, pupils were able to articulate in considerable detail what they perceived as didactically distinctive about CLIL lessons compared to non-CLIL lessons. These descriptions focused on lesson structure, transparency of learning goals, and opportunities for reflection and evaluation. One pupil explained:

> There is a clear structure in the CLIL lessons: they always begin with a 'warmer' and end with a concluding activity and then we evaluate it. The teachers tell us what the goal of the lesson is and at the end we see whether we have reached that goal. In the other lessons we don't do this at all, but in the CLIL lessons we do it a lot. (year 2, School C)

Pupils also emphasized that this structured approach was accompanied by systematic checking of understanding and individualized support, including the strategic use of Dutch when needed:

> The teacher gives a little test to see if we understand the lesson, and if we don't understand then you raise your hand and say, 'I don't understand', and she explains what you don't understand and if you still don't understand then she explains it again in Dutch. (year 1, School F)

In addition, some pupils explicitly valued attention to language form and corrective feedback within subject lessons, indicating that linguistic accuracy was treated as an integral part of CLIL instruction rather than as the sole responsibility of English teachers:

> The other CLIL teachers also pay attention to English grammar and they let us know when we make a mistake. That's what a good teacher should do… (year 1, School A)

Multimodal delivery and engaging teacher behavior

Beyond lesson structure and instructional clarity, pupils also highlighted teachers' delivery style and use of multimodal support as contributing to both enjoyment and comprehensibility in CLIL lessons. They referred to explanations that combined visual, auditory, and demonstrative elements, as well as repetition and personalization:

> He explains it on the board and makes drawings on the board, and then he writes it on the board so you hear it and also see it. (year 1, School E; year 1, School A)

> They repeat the explanations many times and demonstrate it with the smartboard, underline, act it out, draw it. (year 3, School D)

Pupils also associated effective CLIL teaching with distinctive teacher behavior and energy, which they felt helped sustain attention and engagement:

> He's really lively when he's explaining something, he jumps around, he is always hyper. It's not dry or boring. (year 1, School D)

Finally, some pupils emphasized the importance of personalized explanation and responsiveness to questions:

> He explains really well and demonstrates how he would do the problem… and if you have questions, he answers them in a really personal way. (year 3, School A)

While many pupils described CLIL as enjoyable when lessons were well structured, supportive, and engaging, focus group data also revealed a contrasting set of experiences. In the following section, pupils articulated sources of frustration related to difficulty, comprehension, and the organization and implementation of CLIL, particularly when these supportive conditions were perceived as absent or inconsistently applied.

Theme: CLIL is frustrating

Perceived underestimation and constrained challenge

A recurring source of frustration concerned pupils' perceptions of being underestimated, both in terms of the level of challenge offered in CLIL lessons and the assumptions underlying pedagogical decisions. Pupils described instances in which instructional choices appeared to be guided by anticipated limitations rather than by pupils' expressed willingness or perceived capacity to engage with English-medium materials. This was particularly salient in relation to reading materials and the balance between Dutch and English use in the classroom.

Several pupils reported that teachers avoided more demanding English-language materials on the grounds that they would be too difficult for *vmbo* learners. As one year-1 pupil explained:

> We never have to read stories, they say that's too hard for us. In the whole year we have only read one story, Cinderella. (year 1, School E)

Such decisions were not always experienced as supportive. Instead, pupils emphasized that the absence of challenge led to frustration, particularly when they believed they were capable of more than teachers appeared to assume:

> But they underestimate us, we can do it. They just have to give us the chance. (year 1, School E)

A similar pattern emerged in pupils' accounts of language choice during CLIL lessons. In some cases, accommodations intended to reduce linguistic difficulty were experienced as unnecessary or even counterproductive. One pupil described a shift from English- to Dutch-medium materials that conflicted with their own assessment of task difficulty:

> At one point they said, "Well, we're going to use the Dutch book because you all think English is too difficult." Well, I thought it was a shame, because I thought the English book was actually easier. (year 2, School B)

Others expressed comparable frustration when teachers reduced exposure to English through immediate translation, interpreting this as a lack of confidence in pupils' comprehension rather than as targeted support:

> If the teachers consistently speak English, yes, it's harder, but sometimes they just translate everything right away because they don't think we can understand it. (year 2, School D)

Together, these accounts suggest that pupils' frustration was not primarily linked to the presence of linguistic challenge, but to situations in which challenge was constrained on the basis of assumed limitations, resulting in a mismatch between pupils' perception of teachers' expectations and their own sense of capability.

Difficulties with specific linguistic skills and CLIL subjects

Some pupils expressed frustration with particular linguistic skills or with specific CLIL subjects, especially when task demands exceeded their sense of control or when the perceived learning payoff was unclear. These frustrations were not uniformly directed at English use itself, but were often linked to specific skill areas or subject-related language demands.

With regard to language skills, pupils most frequently mentioned difficulties with productive skills and with grammatical metalanguage. Speaking English in class was described as particularly challenging, sometimes leading to avoidance rather than engagement:

> I think speaking English is the hardest. So I just don't do it. (year 1, School E)

Grammar was experienced as frustrating when formal terminology obscured procedural knowledge, resulting in uncertainty rather than support:

> I think grammar is the hardest because I don't know all the terms… I just want to DO it. (year 2, School B)

Similarly, writing was often associated with negative affect, particularly when tasks were perceived as repetitive or unengaging:

> I find writing difficult. I hate it. We only have boring writing assignments, no creative writing, ugh, not fun. (year 1, School E)

In addition to language skills, pupils reported frustration with certain CLIL subjects, particularly STEM subjects, where subject-specific terminology and language use intensified

cognitive demands. One pupil described how English-only instruction in mathematics limited comprehension and led to concerns about content learning:

> One teacher [Math] only spoke English… I have the feeling that I learned less math because of that. (year 2, School E)

Biology was similarly experienced as linguistically demanding, with pupils questioning the relevance of specialized vocabulary for their broader English development:

> Biology has lots of hard words; sometimes you are supposed to figure out the meaning of the word from the text, but you don't even understand the whole text. You don't really have the feeling that you are learning much English. You learn a lot of new words, but they're not words you would normally use, like names of plants, so it feels useless. (year 1, School C)

These accounts indicate that frustration arose when specific skills or subjects were experienced as linguistically demanding without sufficient support, clarity, or perceived relevance, rather than from the use of English as such.

Frustration with international orientation (EIO) activities

Some pupils expressed frustration with elements of the international orientation (EIO) component of CLIL, particularly when activities were experienced as unengaging, fragmented, or insufficiently realized in practice. Rather than functioning as a meaningful extension of English use or international engagement, EIO activities were often described as lacking relevance, continuity, or genuine interaction.

At the classroom level, pupils questioned the value of certain EIO-related tasks, especially when topics were perceived as disconnected from their interests:

> We have to do presentations [for EIO] about topics that nobody thinks are interesting. Nobody wants to do them. (year 2, School A)

Attempts to establish international communication were also described as inconsistent or short-lived, which undermined pupils' motivation to engage with EIO activities over time:

> We tried to have an e-mail exchange with pupils in Canada, but it didn't work… then we never had any more contact. (year 3, School A)

For pupils in higher year groups, frustration was further intensified by the perception that international experiences were limited to a small subgroup rather than being accessible

to the cohort as a whole. One pupil described how both outgoing exchanges and contact with visiting pupils were restricted to a select group:

> A few pupils went to Germany, France, Turkey and Spain [7 pupils total]. The rest did nothing… There were also exchange pupils from abroad who came to our school, but only the pupils from the 'exchange group' had any contact with them. (year 3, School D)

In addition, some pupils reported disappointment when international exchanges that had been presented as part of the CLIL program were not realized as anticipated:

> In year 1 they said that some pupils would go to Germany, and in year 4 to England. But last year they didn't go to Germany after all. (year 2, School D)

Taken together, these reports suggest that dissatisfaction with EIO was less related to the idea of international orientation itself than to its perceived lack of substance, continuity, and inclusive reach, which reduced its perceived value for pupils over time.

Limited didactic activation in CLIL lessons

Some pupils reported that CLIL lessons were not experienced as more didactically activating than their non-CLIL lessons. Rather than perceiving CLIL as involving greater variation or more interactive forms of instruction, these pupils described lessons as largely teacher-centered and similar in structure across subjects and languages.

As one pupil summarized, CLIL lessons were not associated with increased activity or engagement:

> We are not more active or anything like that. It's all the same: just sit and listen. (year 1, School E)

Others emphasized the lack of variety in lesson structure, particularly the predominance of workbook-based tasks and extended periods of passive listening:

> Nothing different—maybe once, otherwise just work in your workbooks. Not that there's a different activity after ten minutes, absolutely not. No variety. (year 3, School D)

Interestingly, some pupils explicitly contrasted CLIL subject lessons with English as a Foreign Language (EFL) lesson, noting that interactive activities were more common in the latter:

> In the English [EFL] lessons there are more activities in the lesson, but not in other [subject] lessons. (year 1, School D)

Taken together, these descriptions suggest that frustration did not stem from CLIL content or language use per se, but from the perception that CLIL lessons did not differ meaningfully in instructional approach or level of activation from non-CLIL lessons.

Peer-related challenges in linguistically diverse classrooms

Some pupils also expressed frustration related to the behavior, language levels, or learning attitudes of their peers during CLIL lessons. These accounts suggest that peer dynamics sometimes complicated classroom interaction and contributed to pupils' negative experiences, particularly in contexts with wide variation in English proficiency.

Several pupils noted that differences in language level affected the pace and flow of lessons. In some cases, classmates were perceived as giving up too quickly, even when contextual clues were available:

> Sometimes you can see from the context of the sentence what it means, but some pupils immediately say, "I don't understand!", even though you can see what it means. (year 2, School B)

Others emphasized that large disparities in English proficiency made it difficult for lessons to meet everyone's needs simultaneously:

> I think there is a big difference in the English levels. Some children have Elementary level and others have a very high level, but they all get the same lesson. (year 2, School B)

These differences also affected peer support within the classroom. Pupils described situations in which those with higher proficiency felt unable or unwilling to support classmates who struggled, particularly when this interfered with their own work:

> Most of the others didn't know much English, and the ones who did didn't feel like explaining everything to them because they already had their own work to do. (year 2, School B)

Finally, some pupils pointed to social and affective factors that limited open communication about difficulty. Fear of negative judgment sometimes prevented classmates from admitting that they did not understand the material:

> There are children who lie… when they have to do the assignment, they don't understand it at all but they don't dare say so, they think they are losers or something. (year 2, School B)

Across these accounts, frustration related to peer dynamics appears to arise not only from differences in proficiency, but also from the social pressures surrounding help-seeking and participation in CLIL classrooms.

Interpersonal teacher–pupil relationships and affective alignment

Some pupils described frustration arising from their interactions with certain CLIL teachers, particularly in situations where they struggled to understand content or made linguistic mistakes. These accounts reveal not only irritation or impatience, but also moments of fear, embarrassment, and emotional withdrawal, which shaped pupils' willingness to participate and seek help during CLIL lessons.

Several pupils reported that teachers reacted strongly when pupils did not immediately understand explanations or made errors, interpreting such responses as anger or loss of patience:

> Some teachers get angry really quickly if you don't understand or make a mistake. (year 1, School C)

In some cases, these reactions were experienced as intense and emotionally unsettling. Pupils described situations in which teachers raised their voices or responded sharply to questions, which discouraged further interaction:

> Some teachers get really angry and start screaming. That's really unpleasant for us… Sometimes you want to ask a question because you don't understand something, so you go to her and she says, really angrily, "Sit down!" or she yells, "What?" and then you get scared. (year 1, School A)

Pupils also emphasized that such responses were not merely momentary discomforts, but affected how safe they felt in the classroom, particularly when they were trying to engage seriously with the lesson. One pupil described how repeated negative feedback and labeling were experienced as especially discouraging:

> Screaming, constantly telling us that we're not doing well, saying that we are acting out typical pre-vocational behavior; that's really hard to hear when you're trying your best. (year 2, School A)

In addition, some pupils reported that attempts to use English were met with ridicule rather than encouragement, further heightening their sense of vulnerability:

> They make fun of us sometimes, they laugh at us when we try to speak English. And then you try to say something and she says, "That's wrong! Ha ha ha!" (year 1, School C)

Taken together, these interactional experiences suggest that pupils' frustration with CLIL was shaped not only by linguistic or cognitive demands, but also by their perceptions of teachers' interpersonal responses to difficulty. For some pupils, these interactions created feelings of fear or discouragement that limited their willingness to ask questions, take risks, or participate actively in English.

Teachers' English proficiency and consistency of language use

Some pupils also reported frustration related to the way English was used by their CLIL teachers, particularly with regard to teachers' perceived level of proficiency and the consistency of target language use. These accounts indicate that pupils were not indifferent to how English was modeled in the classroom; on the contrary, they were acutely aware of both linguistic accuracy and language-switching practices, and evaluated these in relation to their own learning and motivation.

Several pupils expressed concern about teachers' English proficiency, particularly when teachers' spoken English was perceived as inaccurate or unnatural. In some cases, this was described as demotivating rather than supportive:

> Mr. X can't speak English very well. He thinks he's speaking English, but it's not good. When I hear him talk, I think that I don't want to speak English because he sounds so weird that I don't want to do it anymore. (year 1, School D)

Concerns about linguistic accuracy were also raised in relation to errors made by teachers or in instructional materials. Pupils described such errors as confusing, particularly when they were dismissed rather than addressed:

> Sometimes they make mistakes in English. Ms. X made a mistake on the board once and we pointed it out, but she said, "Just shut your mouth and keep writing." (year 1, School F)

> The tests often have mistakes in English and that's really confusing for us. (year 3, School A)

In addition to proficiency and accuracy, pupils frequently commented on inconsistent language use during CLIL lessons. Unsignaled switching between English and Dutch was experienced as confusing and, in some cases, cognitively overwhelming:

> The teacher starts a sentence in English but then switches: "Good morning, *wij gaan aan het werk*" … That doesn't set a good example. (School F, year 1; School D, year 3)

> Sometimes it's the same word in English as in Dutch but if you don't know which language the teacher is speaking, then you don't know what's what… In Biology the teacher sometimes speaks English and sometimes Dutch… all mixed up. It makes you dizzy. (year 1, School F)

These classroom perceptions suggest that pupils' frustration was shaped not only by how much English was used in CLIL lessons, but by how consistently and confidently it was modeled. For some pupils, perceived inaccuracies and unstructured codeswitching reduced clarity and undermined their willingness to engage with English themselves.

Limited attention to language and corrective feedback in CLIL lessons

Some pupils expressed frustration with the limited attention paid to English language use in CLIL subject lessons, particularly with regard to corrective feedback. Rather than experiencing CLIL as an integrated focus on both content and language, these pupils described situations in which language-related feedback was vague, absent, or confined to English as a Foreign Language (EFL) lessons.

Several pupils reported that when correction did occur, it was not always specific enough to be helpful. Being told that something was wrong without clarification led to uncertainty and self-doubt:

> Sometimes you hear that you're doing something wrong, but you don't know what and they don't tell you, and you think, "Everything I do is wrong." (year 1, School A)

Others emphasized that the absence of corrective feedback limited their opportunities to learn from mistakes, particularly when teachers responded positively without addressing language use:

> You don't learn anything if the teachers can't explain the language themselves… You don't get feedback. And you do want it. (year 1, School F)

From the pupils' perspective, attention to English was largely restricted to EFL lessons, with little explicit focus on language during CLIL subject lessons:

> [Do the subject teachers spend time on the English language?] No, not really, only in the English [EFL] lessons. (year 1, School E)

This separation was articulated explicitly by some pupils, who described CLIL subjects as focused exclusively on content rather than on language learning:

> English is about English, Religious Studies is about Religious Studies, not about English. It's not like you suddenly talk about the English language during a Biology lesson or something like that. (year 1, School F)

The pupils' descriptions suggest that frustration arose when pupils perceived limited opportunities to receive meaningful feedback on their English use in CLIL lessons, which reduced both clarity about their performance and the perceived language-learning value of subject lessons.

Limited amount of English in CLIL lessons

Many pupils expressed frustration with the limited amount of English used during CLIL lessons. Notably, such concerns were raised primarily by pupils in the lower *vmbo* tracks, who described CLIL subject lessons as involving little sustained use of English. Rather than experiencing CLIL subjects as predominantly English-medium, these pupils reported classrooms in which English was used infrequently or not at all, particularly outside of English as a foreign language (EFL) lessons.

We think it's a pity that there's not more English. They don't speak English very often, really little. (year 1, School E)

Others reported that English was largely absent from CLIL subject lessons altogether, reinforcing the perception that English use was confined to EFL classes:

No, the teachers don't usually speak English. Only in the English lessons. (year 1, School D)

This experience was echoed across schools and year groups, particularly in the lower *vmbo* tracks, with pupils emphasizing the missed opportunity for language exposure within CLIL:

> We only hear English in the English lessons. Unfortunately. (year 2, School E)

Taken together, these accounts suggest that frustration arose when pupils perceived a mismatch between the promise of CLIL and the limited presence of English in day-to-day classroom practice.

Inconsistent and weakly structured English use

In addition to the limited amount of English used, pupils also reported frustration with how English was used during CLIL lessons. English was described as inconsistently ap-

plied, insufficiently structured, or reduced to symbolic or tokenistic functions, which undermined its perceived value for learning.

Some pupils acknowledged practical reasons for using Dutch in certain contexts, such as safety-related instruction, but questioned why English was also absent in situations where these concerns did not apply:

- The Technical Studies teacher doesn't use English, but that's because it can be really dangerous if you don't understand the instructions…
- But he doesn't speak English in the theory lessons either! (year 1, School C)

Others highlighted a recurring gap between stated intentions and actual classroom practice, noting that teachers often announced plans to teach in English but reverted to Dutch:

> They say they'll speak English but then they speak Dutch after all. (year 3, School D)

Unstructured language use was also experienced as unhelpful when English explanations were immediately repeated in Dutch, which pupils perceived as unnecessary:

> The teacher only speaks English about half the time. She first explains it in English and then again in Dutch, but that's completely unnecessary. (year 3, School D)

Finally, some pupils described English use as limited to brief, formulaic exchanges that occurred at the margins of lessons rather than forming part of substantive instruction. In these cases, English was used in predictable classroom routines, without requiring extended language use or engagement with subject content:

> Only the English and Religious Studies teachers speak English… then the teacher asks questions and we have to answer in English. For example: "Do you have a pencil?" (year 1, School F)

A clear pattern emerges: frustration arose when pupils perceived a mismatch between the promise of CLIL and the limited presence of English in day-to-day classroom practice.

Resistance to obligatory target language use

Some pupils, across different class years, expressed strong reservations about being required to communicate in English during CLIL lessons, particularly in peer–peer interaction. These views were voiced most clearly in lower pre-vocational streams, where pupils emphasized both the absence of clear rules regarding English use and their reluctance to engage in obligatory target language communication.

Several pupils stated explicitly that speaking English with classmates was neither expected nor enforced, and that they would resist such a requirement even if it were introduced:

> We never speak English with each other. And the teachers never say that we must. There is no rule that we have to, and even if there was, I wouldn't do it. (year 1, School E)

While teacher use of English was generally viewed as acceptable, expectations for pupils to use English were sometimes experienced as unreasonable, particularly when comprehension was already limited. In such situations, requests to use English were described as obstructive rather than supportive:

> You really don't understand something [and you ask in Dutch], then you get that irritating nonsense: "Speak English with me!" (year 3, School D)

Pupils also noted that expectations regarding English use were not consistently upheld by teachers or pupils, which weakened their legitimacy as shared classroom norms:

> There are teachers who say, "You must speak English", but we don't do that. Not exactly that we refuse, but nobody really does it, not the teachers and also not the pupils. (year 2, School C)

Finally, some pupils framed resistance to obligatory English use in terms of feasibility rather than willingness, pointing to large differences in English proficiency within classes:

> The ones who can speak some English could speak more, but not the others because they can't do it at all. (year 1, School F)

These perceptions indicate that pupils' resistance to obligatory target language use was shaped less by opposition to English itself than by concerns about feasibility, uneven proficiency levels, and the absence of shared and consistently enacted classroom norms.

Native-speaker teachers without Dutch proficiency

Issues concerning the balance of Dutch and English use were raised by pupils in nearly all focus groups. In particular, pupils expressed mixed views on the presence of native-speaker teachers, a requirement of the Dutch national Standard for pre-vocational bilingual schools. While the presence of a native speaker was generally not questioned in principle, many pupils emphasized that such teachers should also be able to speak Dutch, especially in lower *vmbo* contexts.

Several pupils argued that the absence of Dutch proficiency limited teachers' ability to support content understanding when pupils struggled, particularly in subjects such as mathematics:

> It's really important that a CLIL teacher can speak Dutch. Because if a pupil doesn't understand the subject at all… then you can't learn it at all if the teacher can only speak English. (year 2, School C)

Concerns were not limited to comprehension alone, but also related to pupils' trust in teachers' instructional competence. One pupil described how a lack of Dutch proficiency undermined confidence in the teacher more generally:

> One teacher couldn't speak Dutch… And we got Dutch worksheets from him, but he didn't even know what was on them and he couldn't really correct them. Then you don't have much faith in the teacher. (year 2, School E)

These concerns were articulated most explicitly by pupils in lower *vmbo* streams, who stressed that limited English proficiency made exclusive use of English particularly problematic:

> For lower *vmbo* kids like us, it's not good [if the teacher can't speak Dutch] … so it's better to give all of the explanation and instructions once in English and then again in Dutch. (year 2, School E)

At the same time, a few pupils expressed an opposing view, arguing that Dutch proficiency was not essential if the purpose of CLIL was understood as English-medium instruction:

> It's not important that the teacher can also speak Dutch… It doesn't matter if he can't speak Dutch, because it's supposed to be in English anyway. (year 1, School E)

The opinions expressed suggest that that pupils' evaluations of native-speaker teachers were shaped less by the teachers' first language status than by their perceived ability to support understanding and maintain instructional credibility within a lower pre-vocational CLIL context.

Pupils' conceptions of bilingual education

In addition to concerns about teacher language proficiency, some pupils articulated a more conceptual understanding of bilingual education, defining CLIL as involving the use of two languages rather than instruction conducted exclusively in English. From this

perspective, access to both Dutch and English was seen as a defining characteristic of bilingual education, rather than as a deviation from it.

Several pupils argued that the absence of Dutch undermined the very notion of bilingual education, particularly when it limited opportunities for clarification or explanation:

> They should speak good Dutch in order to explain things if necessary. Otherwise it's just monolingual education. (year 3, School D)

This interpretation was not uncontested. In some focus groups, pupils explicitly debated whether bilingual education required teachers to use both languages, revealing differing assumptions about the purpose of CLIL:

> A CLIL teacher should also be able to speak Dutch. – No, they shouldn't! – But it's 'bilingual!' If all the teachers don't speak Dutch and English, it's not bilingual any more. (year 1, School E)

These exchanges suggest that pupils' expectations of CLIL were shaped not only by classroom practice, but also by their underlying understanding of what 'bilingual education' entails.

Tapering-off of CLIL and exam-language misalignment

Some pupils also expressed frustration with the way CLIL was gradually reduced in years 3 and 4, as subjects were increasingly taught in Dutch in preparation for the Dutch-language final exams. From the pupils' perspective, this transition disrupted both continuity in CLIL and alignment between instructional language and assessment.

Several pupils questioned the fairness of being examined in Dutch for subjects that had been taught in English, arguing that assessment language should correspond to the language of instruction:

> Actually, it's not fair that the exams are in Dutch. I would rather do exams in English for the subjects that we've had in English. (year 2, School A)

This mismatch was described as particularly problematic when subject-specific terminology had been learned primarily in English, leading to feelings of insecurity and self-doubt:

> The subjects that we now have in English, like Biology and Math, we should do exams in English. Because I've seen one of the exam problems in Dutch and we haven't even had some of those terms, and then you feel like the dumbest one in the class. (year 2, School A)

In addition to concerns about assessment, pupils expressed disappointment that CLIL largely ended after year 2, despite their interest in continuing with English-medium instruction:

> And then in year 3 CLIL is basically over. That's really a shame. I would like to have another year. (year 2, School B)

Others linked this early tapering-off to perceived language loss, noting that reduced exposure made it difficult to maintain what they had previously learned:

> We used to have lots more CLIL subjects, but next year we have exams in Dutch, so now we don't have those subjects in English anymore. Then you forget a lot. (year 3, School A)

Overall, pupils' reports indicate that frustration was shaped by systemic constraints related to assessment and curriculum sequencing, which limited continuity in CLIL and weakened alignment between instruction and evaluation.

Theme: CLIL can be improved

The third theme reflects pupils' suggestions for improving the CLIL programs at their schools and highlights the conditions under which they believed CLIL could function more effectively in a pre-vocational context. Across the focus groups, more than 25 distinct recommendations were identified, ranging from expectations regarding teacher skills to lesson design and language use. These suggestions were coded and subsequently grouped into four overarching categories: essential pre-vocational CLIL teacher skills, lesson activities, L1/L2 language use, and teacher didactic, pedagogic, and interpersonal behavior.

Essential skills for pre-vocational CLIL teachers (according to pupils)

Across focus groups, pupils articulated a broad and coherent set of expectations regarding what they considered essential skills for pre-vocational CLIL teachers. While more than 25 individual recommendations were coded across groups, pupils' responses clustered around three core areas: language proficiency, pedagogical competence, and interpersonal sensitivity. Notably, pupils' initial and most immediate responses consistently concerned teachers' command of English, often coupled with the ability to adapt language use to pupils' level.

> CLIL teachers have to speak really good English, and give good explanations of their subject in English, using language that is not too difficult. They have to be able to adjust their language to our level. (year 2, School C)

At the same time, pupils emphasized that linguistic proficiency alone was insufficient. Effective CLIL teachers were described as those who could explain clearly, correct errors, and recognize when pupils needed additional support, while maintaining subject expertise.

> They have to be able to correct us if we make mistakes. (year 2, School D)

Alongside language and pedagogical skills, pupils placed strong emphasis on teachers' interpersonal qualities, particularly patience, emotional control, and understanding of the difficulty CLIL can pose for pre-vocational learners.

> Have a lot of patience and understand that we think it's really difficult sometimes. Be nice and don't get angry. Be helpful and kind. (year 1, School F)

Rather than presenting these skills as optional extras, pupils framed them as baseline conditions for CLIL to function effectively in a pre-vocational context.

Lesson activities: active, concrete, and varied formats

The largest category of improvement suggestions concerned lesson activities. Across focus groups, pupils repeatedly contrasted passive, worksheet-based lessons with activities that involved *movement, concrete materials, and varied audiovisual input*. What pupils emphasized was not novelty for its own sake, but sustained attention and active engagement through doing, seeing, and experiencing.

> Active lessons, that you don't just have to listen and sit still. More exciting activities, not just 'Go to the shop and buy chewing gum.' Don't just work from the book, for example, go outside and ask for directions in English. (year 3, School D)

Several pupils stressed the importance of hands-on and real-world materials, particularly in content subjects, arguing that concrete engagement supported both understanding and motivation.

> Let the pupils do real things, like if we have to collect leaves [for Biology] then real leaves, not just drawings. Things you really do something with. (year 1, School F)

In addition, pupils frequently mentioned the underuse of visual and audiovisual resources, especially given the availability of interactive whiteboards. Visual input was described as helping maintain attention and making lessons feel less monotonous.

> At the moment you usually just get a worksheet and you have to fill it in. They should use the resources that they have, for instance the smartboard [IWB]. We

> have smartboards in the classrooms, but that doesn't mean the teachers actually use them. (year 2, School E)

Finally, pupils highlighted films, music, and other contemporary media as particularly effective because these formats combined enjoyment with meaningful exposure to English.

> I would have the pupils learn the lyrics to a song, because that's easier to learn in English than a normal text. And then listen to the songs and understand the meaning. (year 2, School E)

Rather than requesting entertainment, pupils framed these activity types as ways of making English more visible, memorable, and worth engaging with within CLIL lessons.

Lesson activities: educational games as learning support

Within the broader set of activity-related suggestions, pupils frequently referred to educational games as a way of making challenging material more approachable. Importantly, games were not framed as a replacement for learning, but as a format that reduced pressure and supported understanding, particularly in difficult subjects.

> The atmosphere in the lesson is better if you just make it fun. For instance, make a game out of a difficult assignment, and then it's easier to learn because you're playing. (year 2, School B)

Pupils repeatedly emphasized that games were effective only when they were clearly linked to learning goals, and several explicitly rejected games that were merely entertaining.

> Play games once in a while where you really learn something, especially if it's a difficult subject, then you really feel like you're learning. (year 1, School A)

Games were thus presented as a pedagogical strategy rather than a diversion, with pupils highlighting their value for sustaining engagement while working through cognitively demanding content.

Lesson activities: internationalization and cultural orientation

A smaller but consistent subset of pupils' suggestions focused on the integration of culture and internationalization into CLIL lessons. Pupils expressed a desire for English to be connected more explicitly to people, places, and practices beyond the classroom, rather than remaining limited to decontextualized language tasks.

> Show something of British culture. Show what they eat, what they drink. (year 1, School F)

In addition to cultural facts and objects, some pupils pointed to direct international contact or experiential learning as particularly motivating, even if this could only be realized on a small scale.

> We would like to do something with kids from other countries. That would be fantastic. (year 3, School D)

These suggestions framed internationalization not as an abstract goal, but as a way of making English personally meaningful and socially situated for pre-vocational learners.

Teacher pedagogic behavior: interaction, group work, and differentiation

Another set of improvement suggestions concerned teachers' pedagogic choices regarding pupil interaction, group work, and differentiation. Across focus groups, pupils repeatedly expressed frustration about being discouraged from talking or working together, despite viewing peer interaction as helpful for understanding and engagement.

> We want to work together more. We're not allowed to do that often because we're too wild and loud and we talk too much. (year 1, School F)

Several pupils explicitly described peer explanation as more accessible than whole-class teacher explanations, particularly when teachers were focused on managing the class as a whole.

> I would like it more if another pupil explained things to me instead of the teacher, because the pupils can explain things so you really understand them… We never do this. Usually we're not allowed to talk with each other. (year 2, School B)

In addition, pupils pointed to differentiation as a way of addressing uneven understanding within the class, suggesting alternative instructional arrangements when some pupils fell behind.

> If some pupils understand the material but others don't, then they should get a different kind of lesson, maybe from a student teacher or trainee. (year 2, School B)

Rather than advocating unrestricted talk, pupils framed interaction and differentiation as structured support mechanisms that would help them keep pace with lessons and remain engaged.

L1/L2 language use: teacher input and strategic use of Dutch
Pupils expressed strong views on teachers' use of English, with several advocating extensive and sustained target language input, particularly at the start of lessons. For these pupils, consistent use of English was seen as a defining feature of CLIL and a condition for taking the approach seriously.

> The teachers should speak a lot of English, even in the beginning nonstop English until they know who really doesn't understand, and then only speak Dutch if it is absolutely necessary. (year 2, School C)

At the same time, other pupils emphasized the importance of strategic use of Dutch to ensure comprehension and prevent pupils from being left behind. Rather than rejecting English input, these pupils framed translation as a form of support.

> Translate to ensure that everyone really understands it, so nobody is left behind. (year 2, School C)

In sum, these suggestions indicate that pupils did not oppose extensive use of English, but expected teachers to balance immersion with responsiveness, using Dutch selectively when comprehension was at risk.

L1/L2 language use: opportunities for pupil speaking and interaction
In addition to teacher language use, pupils repeatedly highlighted the need for more opportunities to speak English themselves. Speaking was associated with active participation and with lessons that went beyond teacher-led explanation.

> Ask questions that the pupils have to answer in English. (year 2, School D)

Several pupils also linked increased pupil speaking to interaction and discussion, framing this as a defining feature of effective CLIL lessons.

> More speaking, and combine that with questions and discussions. That's called 'interactive', right? (year 3, School A)

While some pupils acknowledged that not everyone might feel comfortable speaking English frequently, increased pupil talk was nevertheless presented as a necessary component of CLIL rather than an optional add-on.

L1/L2 language use: rules for pupil target language use
A further subset of suggestions concerned explicit rules for pupils' own use of English in CLIL lessons. Pupils expressed ambivalent but reflective views: while speaking English

was widely recognized as difficult, many pupils nevertheless framed it as necessary for developing communicative competence, particularly if expectations were shared and consistently applied.

This tension emerged clearly in group discussion, where initial resistance was immediately countered by an explicit learning rationale:

- No!
- But really, it's good for our English if we have to speak more, because then we learn to communicate better in English.
- But it's harder!
- But if you have to, you will do it! (year 1, School D)

Several pupils emphasized that willingness to speak English depended less on individual confidence than on collective expectations, suggesting that uniform rules would reduce social pressure.

> If everyone had to do it, I would do it too. (year 1, School E)

Crucially, pupils repeatedly located responsibility for such rules with teachers, stressing the importance of consistency and modeling rather than coercion.

> It's good if the teacher has the rule that you have to speak English in class. The teacher should do that and the pupils too. It shouldn't be forced, but… if the teacher does it consistently, then the pupils will automatically do it too. (year 3, School D)

Taken together, these accounts suggest that pupils did not reject target language use requirements, but viewed them as effective only when they were collective, predictable, and consistently enacted by teachers themselves.

Teacher pedagogic behavior: interpersonal and learner-focused skills

The final category of suggested improvements concerned teachers' interpersonal and learner-focused behavior. Pupils emphasized that how teachers interacted with them mattered as much as what they taught, particularly in a CLIL context where linguistic difficulty could easily lead to frustration or insecurity. Central to pupils' accounts were expectations of patience, emotional control, and respectful treatment.

> Don't get angry if we don't understand it when you try to explain something. Don't get impatient if we don't understand it. (year 1, School E)

Beyond affective support, pupils valued teachers who were physically and socially present in the classroom. Walking around, initiating brief conversations, and checking in with pupils were described as signals that teachers noticed individual needs and were willing to offer support.

> Walk around, have a chat, don't just sit there. Walk through the class and ask if everything's ok and if anybody needs help. (year 1, School A)

Pupils also stressed the importance of making lessons personally meaningful, for instance by connecting content and language use to their own interests and experiences. Such personalization was seen as increasing both engagement and willingness to participate, particularly in English.

> Make it personal; choose examples that we can relate to, that are also relevant to us. For instance, gaming—compare something in a game to something that you're trying to explain. (year 2, School A)

What pupils described here was not an abstract teaching style, but a form of relational pedagogy in which feeling seen, respected, and taken seriously created conditions for engagement with both content and language.

L1/L2 language use: calls for increased teacher target language input

The final question posed to all focus groups directly addressed teachers' use of English in CLIL lessons. This question was added to the interview protocol after the first focus group indicated that teachers' use of the target language was perceived as minimal. With the exception of pupils from School A—who reported that their teachers already used English almost exclusively—*all* pupils across focus groups responded affirmatively when asked whether their CLIL teachers should speak more English.

Importantly, this endorsement was rarely unconditional. Pupils consistently qualified their responses by emphasizing clarity, teacher proficiency, and continued access to Dutch when comprehension was at risk.

> Yes, but slowly and clearly. (School E, year 1; year 2)

Several pupils also linked increased English use to teachers' own linguistic competence, expressing doubts about feasibility rather than principle.

> Yes, but first they have to be able to speak English well. (year 2, School C)

Finally, pupils stressed that increased English input should not come at the expense of understanding, reiterating the need for selective translation in cases of difficulty.

> Yes, but they still have to translate into Dutch if some children really don't understand. (year 2, School E)

These responses underscore that pupils' calls for more English reflected conditional support: greater target language exposure was welcomed, provided it remained comprehensible and responsive to learners' needs.

5.5D Teacher and pupil perceptions compared

In conclusion, while the three central themes—*CLIL is enjoyable*, *CLIL is frustrating*, and *CLIL can be improved*—were evident in both teacher and pupil accounts, the categories within these themes were not entirely congruent. Teachers tended to highlight professional, linguistic, and organizational dimensions of CLIL implementation, whereas pupils emphasized classroom dynamics, enjoyment, and practical lesson activities. For pupils, the vitality of CLIL appeared to depend less on any single factor than on a combination of conditions: teachers' confident use of English; varied, enjoyable, and communicative lesson design; and what might be described as teachers' *willing suspension of underestimation*—the belief that even less academically inclined learners can thrive in an English-medium environment.

Differences between schools and program levels further underscore this interplay. Pupils in the higher *vmbo* CLIL streams generally expressed satisfaction with the intensity of their teachers' use of English and the level of challenge offered, while those in lower *vmbo* tracks were more critical of limited English input, lower linguistic expectations, and reduced communicative variety. These contrasts point to the combined influence of program maturity, teachers' balancing of content comprehension with English use, and an expectations culture that may underestimate lower-*vmbo* learners, as perceived by pupils.

To facilitate systematic comparison in the Discussion chapter, the analytical categories have been adjusted and partially regrouped under slightly modified thematic headings. This reorganization enables a coherent analysis of similarities and differences across stakeholder perspectives while preserving the nuances of each group's contributions. Together, these results mark a transition from descriptive findings to interpretive understanding. The following section examines how these differential patterns can be understood in light of CLIL theory and principles of inclusive, affect-sensitive pedagogy.

5.6 Discussion and conclusion

The patterns identified above highlight how teachers' and pupils' perceptions of CLIL are shaped by the ways in which individual programs interpret and enact the approach, as well as by prevailing expectations of what pre-vocational learners are believed to be able to achieve through English.

This qualitative study investigated the perceptions of pre-vocational secondary school teachers (n = 53) and pupils (n = 267, including 70 focus group participants) engaged in bilingual education using the CLIL approach during the early years of implementation in the Netherlands. While CLIL has been widely studied in higher academic tracks, far less is known about how teachers and pupils in pre-vocational settings experience and evaluate it. By foregrounding these stakeholder voices, the study sheds light on both the opportunities and challenges of CLIL as perceived by participants who are central to its day-to-day implementation, yet whose perspectives have rarely been examined in depth.

Drawing on participants' own reflections and suggestions, the discussion includes recommendations for strengthening CLIL in Dutch pre-vocational education at the classroom, school, and policy level, with relevance for other CLIL contexts involving learners with lower academic aptitude. The discussion also situates the findings within the broader literature on stakeholder perceptions of CLIL in other educational contexts. Although this comparative element plays a more limited role, it helps to clarify where the Dutch pre-vocational experience aligns with or diverges from international patterns.

This chapter addresses two research questions, based on a multi-perspective qualitative analysis involving both pre-vocational CLIL teachers and their pupils. For RQ1a and b, teachers' and pupils' perceptions of CLIL classroom practice are presented under modified thematic headings, enabling systematic comparison of similarities and differences between stakeholder groups. These findings are then considered in relation to research from other CLIL contexts, with particular attention to studies involving learners with lower academic aptitude. Finally, in response to RQ2, recommendations for the improvement of CLIL programs at the classroom, school, and policy level are discussed, grounded in both the comparative analysis and the suggestions offered by the participating stakeholders themselves.

5.6A Summary of main findings

Before turning to the specific findings, it is useful to situate this chapter in relation to the earliest Dutch research on pre-vocational CLIL. The exploratory study by Denman et al. (2013) was among the first to describe CLIL in Dutch junior vocational secondary education, capturing both enthusiasm and uncertainty in its earliest stages. Conducted when *vmbo* CLIL was still a radical innovation, it found that about 70% of participating pupils

would choose the CLIL stream again, highlighting the motivational promise of bilingual education even for less academically oriented learners. Teachers and pupils valued the practical, interactive nature of CLIL and the increased use of English, while noting heavy preparation demands, scarce materials, and uneven language proficiency among both staff and pupils.

The present chapter extends that early work to a broader cohort of *vmbo* CLIL schools. Seventy-four per cent of pupils in the current study likewise indicated they would choose CLIL again, confirming the approach's sustained appeal. Yet the perceptions and emotions of pupils were explored much more fully here, revealing both enjoyment and self-confidence gains and critical reflections on teacher behavior, lesson activities, and program features across schools. Some new concerns, absent from the Denman et al. study, also emerged: pupils' awareness of a gap between teachers' actual use of English and the greater exposure they desired, reflecting heightened sensitivity to CLIL's linguistic aims, as well as teachers' worries about the effective learning of subject content in English. At the same time, as CLIL became more established, teacher and pupil expectations appear to have risen, possibly contributing to the sharper criticism voiced in the later study. Teachers continued to report professional and linguistic growth, but also ongoing constraints of time, resources, and proficiency demands. Together, these studies trace the evolution towards a more balanced understanding of the benefits, challenges, and expectations shaping CLIL in pre-vocational education.

English learning and pupil L2 proficiency

Both teachers and pupils in this study identified English language development as the most salient positive outcome of pre-vocational CLIL. Teachers reported visible progress in their pupils' vocabulary, oral proficiency, and willingness to use English in the classroom. Pupils themselves confirmed these perceptions: they valued the opportunity to improve their English, often naming it as the primary benefit of CLIL. 74% of those surveyed indicated that they would probably or definitely choose for a CLIL stream again. The findings suggest that for many pre-vocational pupils, English learning provides both intrinsic motivation (pride, enjoyment, sense of accomplishment) and extrinsic motivation (extracurricular communication, future studies or job prospects, travel). This aligns with international studies where pupils cite English improvement as a decisive factor in opting for CLIL (Doiz et al., 2014b; Lasagabaster, 2011).

At the same time, teachers also identified low pupil L2 proficiency as the greatest impediment to content learning. Subject teachers in particular voiced concern that their pupils' English was 'too low' to access disciplinary concepts, possibly leading to reduced subject learning or to wide disparities within classes. This ambivalence—seeing English gains but also experiencing English as a barrier—was also mirrored by pupils themselves. Some reported that their teachers underestimated them, for example by indicating that their

level was too low for reading stories or engaging with subject matter in English. Others expressed frustration when the amount of English in lessons was lower than expected, stating unanimously in the focus groups that their teachers should use *more* English, not less. Teacher self-reports of classroom English use help explain this disjuncture: some teachers, especially in lower *vmbo* tracks, reported using English only 20% (or less) of the time, while others reported much higher use. Pupils' accounts across schools broadly confirmed these self-reports, indicating that target language exposure varied considerably both within and across contexts.

This variation highlights a key paradox in pre-vocational CLIL: while teachers may reduce their English use in an effort to safeguard content learning, pupils interpret this as a lack of challenge or teachers' confidence in them, and their satisfaction appears to drop accordingly. The survey and focus group data both indicated a strong correlation between higher teacher use of English and higher pupil satisfaction, whereas lower English use was linked to frustration, echoing Tompkins' (2022) findings that socioeconomically diverse pupils exposed to less consistent target language input demonstrate weaker engagement and outcomes. Conversely, when pupils feel underestimated and undervalued—such as being denied opportunities to work in English—they may disengage or lose motivation, a process linked to self-sabotaging behaviors that diminish both self-esteem and academic involvement (Zhang, 2022). Such lowered expectations can have enduring negative effects: as shown in Rosenthal and Jacobson's (1968) classic study, teacher expectations can significantly shape pupils' intellectual development, with underestimated pupils often experiencing reduced achievement over time.

In comparison with research from other contexts, these findings both align and diverge. Studies of German *Hauptschule* pupils (Schwab, 2014, 2015) similarly noted clear English proficiency gains and strong motivation among weaker learners, but also reported comprehension problems and occasional cognitive overload. Other scholars caution that CLIL "may not work for all learners" if L2 input is prolonged and incomprehensible (Otwinowska & Foryś, 2017, p. 473). Yet the current study suggests that pre-vocational pupils do not consider CLIL unworkable because of these challenges; instead, they request *more* linguistic ambition from their teachers, combined with support to help them manage the difficulty.

These findings carry several implications. First, they reinforce the need for CLIL teachers, especially in pre-vocational contexts or mixed-ability groups, to maintain high expectations for pupils' English use while employing scaffolding techniques such as modelling, visual support, simplification, repetition, and peer collaboration. Reducing English input in order to 'protect' pupils or safeguard content learning may inadvertently lower motivation and satisfaction. Second, they underscore the importance of addressing teachers' own insecurities about pupil language levels. Rather than assuming that low proficiency

necessitates a retreat to Dutch, teachers might benefit from training in scaffolding strategies that sustain content learning *through* English. Third, the results support the value of *translanguaging* as a pedagogical resource, allowing teachers to draw strategically on both Dutch and English to maximize meaning-making and confidence (Garcia & Wei, 2014; Nikula, 2017b). Finally, the study suggests that pre-vocational CLIL may provide an important motivational pathway for pupils who otherwise have limited academic opportunities: by achieving visible progress in English, they gain both self-confidence and a sense of broader horizons.

CLIL didactics and lessons

In this chapter, *didactics* is used in the Dutch sense of the term, referring to the structuring and planning of lesson activities and delivery, whereas *pedagogy* denotes broader teaching strategies and approaches. This distinction is important for understanding teachers' and pupils' perceptions, since much of their commentary concerned the concrete organization and execution of CLIL lessons rather than pedagogical orientation.

Teachers and pupils alike emphasized that the didactic structuring of lessons was a decisive factor in shaping their experiences of CLIL. For teachers, CLIL was often associated with the opportunity to design lessons that were more engaging, creative, and authentic than in the regular stream. Some teachers described using multimodal resources, hands-on projects, or international themes to make subject content more meaningful. However, they also acknowledged substantial challenges: lack of ready-made, level-appropriate CLIL materials, insufficient preparation time, and limited opportunities for collaboration with colleagues often made it difficult to realize the dual focus on content and language that is central to CLIL, and contributed to teacher frustration. As one teacher noted in the questionnaire, "*A lack of preparation time means we have less CLIL focus in the lessons.*"

Pupils strongly confirmed that when CLIL lessons were varied, interactive, and connected to real-world activities, they found them motivating and enjoyable. They frequently cited '*fun*' and '*active*' tasks—such as group projects, creative assignments, international trips, and peer exchanges—as highlights of their CLIL experience. Yet the opposite was also true: when promised opportunities such as international trips or exchanges did not materialize, pupils' disappointment sometimes extended to a negative attitude toward the CLIL program as a whole. Similarly, when lessons relied heavily on teacher talk or routine exercises, pupils reported boredom and frustration. These reactions underline the extent to which pupils judged the value of CLIL through the lived reality of classroom activities and school-level implementation, rather than abstract program goals.

This emphasis on action-oriented, varied lessons is particularly salient in the *vmbo* context, where many pupils are characterized as 'doers' rather than 'listeners' (Korpershoek et al., 2016). Research has shown that *vmbo* learners tend to benefit from concrete,

hands-on tasks and are prone to disengagement when faced with prolonged listening or abstract instruction (Ledoux et al., 2010). Their relatively short attention span, combined with vulnerability to frustration and sometimes low academic self-esteem, means that lesson design plays an especially crucial role in sustaining motivation. The findings of the present study strongly reflect these tendencies: pupils valued CLIL most when it allowed them to be active, creative, and socially engaged, and they rejected it when lessons became monotonous, unenjoyable, or when promises were not fulfilled.

Taken together, teacher and pupil perceptions reveal both convergence and divergence. Both groups agreed that CLIL has potential to enrich lesson delivery, but teachers highlighted systemic obstacles (time, training, materials), whereas pupils focused on the immediate experience of activities. In other words, teachers tended to frame CLIL didactics as a structural challenge, while pupils framed it as a classroom experience. This pattern resonates with Kim and Graham's (2022) review of CLIL teachers' perception of implementation constraints, where contextual conditions often prevent teachers from fulfilling CLIL's dual aims, and with Cenoz et al.'s (2014) observation that CLIL effectiveness varies widely because of uneven didactic realization.

The findings also echo Banegas (2012), who argues that CLIL is most successful when adapted to local realities rather than applied as a rigid model. In the current study, pupils' desire for more engaging and relevant activities and teachers' call for greater methodological support both point to the need for school-level flexibility and investment in CLIL didactics. Furthermore, the mismatch between teachers' intentions and pupils' perceptions underscores the importance of teacher reflection: as Escobar Urmeneta (2013) notes, effective CLIL teaching requires continuous adjustment to learners' responses and feedback.

In the pre-vocational context, where pupils may already feel academically marginalized, the quality of lesson design appears especially crucial. Pupils' disappointment when activities were repetitive or lacked a communicative focus suggests that didactic shortcomings may erode trust in the CLIL program more broadly. Conversely, when lessons were dynamic and interactive, pupils not only enjoyed learning but reported feeling more engaged, perceiving their English improvement as more meaningful. Thus, didactics are not a neutral vehicle for CLIL but a central mediator of how pupils evaluate the entire program.

Teacher behavior and teacher affect

Teachers' behavior and attitudes emerged as a crucial influence on how pupils experienced CLIL. On the positive side, many teachers reported that CLIL gave them renewed enthusiasm, professional challenge, and a sense of creativity. Several described how working in English led them to rethink their teaching, adopt new strategies, and connect

more actively with pupils. One teacher noted: *"[There are] more positive challenges and more possibilities for interesting lessons with CLIL, this increases teacher motivation*." Some teachers pointed to the increased need for strengthening personal relationships with their learners in CLIL: "*The pupil-teacher relationship is even more important, especially with 'special needs' pupils*." Pupils confirmed that when teachers were enthusiastic, supportive, and approachable, this created a motivating classroom climate. They valued teachers who showed energy and personality in their teaching, who encouraged them to speak English, and who expressed confidence in their abilities. In such cases, pupils felt encouraged to take risks, to experiment with English, and to persevere even when subject matter was difficult.

At the same time, both groups pointed to challenges in teacher behavior and affect. Pupils repeatedly noted that some teachers showed limited empathy for their struggles, especially in Year 1 and in the lower *vmbo* tracks. They described situations where teachers seemed impatient with their difficulties or underestimated their abilities, which they perceived as discouraging and demotivating. Teachers, in turn, acknowledged frustration with pupils' low English level, inconsistent motivation, and behavioral issues, which sometimes made it difficult to sustain positive affect in the classroom. They also cited increased workload, lack of training, and insufficient collaboration with colleagues as factors that undermined their own energy and enthusiasm for CLIL. These findings resonate with recent conceptual work emphasizing cross-disciplinary collaboration as a core component of effective CLIL implementation (Dale & Mearns, 2023).

A few teachers questioned whether CLIL was even appropriate for these learners: *"Many pupils aren't even able to speak Dutch properly, so why are we teaching in English?"* which indicates doubts about the fundamental premises of CLIL. If teachers are not convinced of the likely success and benefits of bilingual education with CLIL, it seems likely that these doubts will be indirectly transmitted to the pupils and perhaps exacerbate their feelings of being underestimated.

For weaker scholastic aptitude learners, such as many pupils in the *vmbo*, teacher affect appears especially consequential. Research has shown that these pupils are highly sensitive to teacher attitudes: they often have relatively low academic self-confidence, are easily discouraged, and thrive when they feel recognized and supported (Korpershoek et al., 2016; Ledoux et al., 2010). In this context, lack of teacher empathy or negative assumptions about pupil ability may trigger disengagement, frustration, or self-sabotaging behaviors (Zhang, 2022). Conversely, teacher enthusiasm, humor, and belief in pupils' potential can serve as powerful motivators (Gordani et al., 2025). Pupils in this study frequently highlighted teachers who were "*really lively*" or who "*explain really well…and if you have questions he answers them in a really personal way*" as the reason they enjoyed CLIL and felt proud of their progress. This finding resonates with research on Foreign

Language Enjoyment (FLE), where Dewaele and MacIntyre (2014) showed that teacher enthusiasm, friendliness, and especially humor significantly boost enjoyment and reduce anxiety in foreign language classrooms. Later studies confirmed that supportive teacher behavior enhances both emotional engagement and willingness to communicate (Dewaele, 2019). Additionally, research into emotional contagion (Gordani et al., 2025) has shown that teachers' manifestation of positive emotions is highly correlated with learners' motivation and academic achievement in the foreign language classroom.

This interplay is particularly relevant to CLIL methodology. One of CLIL's central tenets is to provide a supportive environment where pupils can learn both content and language simultaneously through scaffolding and interaction (Coyle et al., 2010). Yet in this study, pupils' accounts suggest that not all teachers implemented CLIL in a way that made their learning struggles visible and acknowledged. When teachers reduced their English use or limited tasks because they believed pupils could not cope, pupils interpreted this as a deficit value judgement or lack of empathy, which in turn undermined motivation. By contrast, when teachers combined high expectations with encouragement, humor, and scaffolding, pupils perceived CLIL as both challenging and rewarding.

The findings align with Doiz et al. (2014b), who highlight the importance of affective variables in sustaining motivation in CLIL contexts. In pre-vocational education, where pupils often have shorter attention spans and greater vulnerability to frustration, teacher empathy and enthusiasm become even more critical. Without these affective supports, even well-structured CLIL didactics risk losing impact.

In sum, teacher behavior and affect strongly shaped pupil experiences of CLIL in this study. Teachers' enthusiasm, humor, and empathy amplified the benefits of CLIL, while their frustration, underestimation of pupils, or lack of visible support exacerbated challenges. For CLIL in *vmbo* contexts, investing in teachers' professional well-being, methodological training, and reflective practice appears vital, not only to sustain teachers themselves but also to ensure that weaker scholastic aptitude learners experience CLIL as motivating and empowering rather than frustrating.

Pupil affect and behavior

Pupils' accounts revealed that affective factors played a major role in their CLIL experience. Many expressed pride in their English progress, describing it as a source of motivation and enjoyment. As noted above, 74% of pupils indicated they would choose CLIL again, and quite a few compared themselves positively to their non-CLIL peers: *"CLIL is super! If you look at [the non-CLIL classes], we already know a lot more."* They also reflected positively on overcoming the initial challenges; as one pupil explained: *"In the beginning it was hard, but after a few months…you get better and better and you can do it…"*. These

moments of achievement contributed to a sense of self-confidence and FLE (Dewaele & MacIntyre, 2014).

Pupils also linked their affective responses directly to classroom dynamics. They enjoyed CLIL lessons most when they were interactive, creative, and allowed them to participate actively. As one pupil commented: *"We watch films and we have books with extra activities…That's good because you always have something to do. You get something extra, that's fun."* Conversely, a lack of 'do-activities' led to boredom and disengagement: *"We are not more active [in the CLIL lessons] or anything like that. It's all the same: just sit and listen."* Such preferences reflect broader tendencies of *vmbo* learners, who are often characterized as practical 'doers' with shorter attention spans, requiring varied and hands-on tasks to remain engaged (Korpershoek et al., 2016; Ledoux et al., 2010). In this sense, CLIL didactics that provided opportunities for active involvement amplified positive affect, boosting both motivation and satisfaction.

At the same time, negative affect was also evident. Some pupils, especially in the lower *vmbo* levels, reported feelings of frustration and anxiety when they struggled to understand subject matter in English. A Year 1 pupil remarked: *"[The teachers] don't seem to get it. They always get impatient if you don't understand."* Others described resistance to speaking English out of fear of making mistakes, echoing patterns of Foreign Language Anxiety (FLA) identified in other studies (Dewaele, 2019). Pupils also became discouraged when teachers underestimated them by reducing English input, interpreting this as a lack of faith in their abilities. As noted earlier, lowered expectations risk triggering disengagement or self-sabotaging behaviors and may ultimately depress academic outcomes.

Year 3 pupils highlighted another affective challenge: the compulsory shift back to Dutch for CLIL subjects that would become national exam subjects. Several expressed disappointment and even resentment about this policy change. As one pupil put it: *"We used to have lots more CLIL subjects, but next year we have exams in Dutch, so now we don't have those subjects in English anymore. Then you forget a lot."* This policy-driven shift undermined their sense of continuity and seemed to devalue their previous efforts.

Overall, the findings suggest that pupil affect in pre-vocational CLIL is highly sensitive to both classroom didactics and broader program structures. Enjoyment, motivation, and pride flourished when pupils were given opportunities to be active, to succeed in English, and to feel supported by teachers. Anxiety, frustration, and demotivation arose when they felt underestimated, when lessons lacked variation, or when external policies contradicted their progress. These findings align with Lasagabaster (2011), who emphasized motivation as a central benefit of CLIL, but also with Doiz et al. (2014b), who noted that affective variables are fragile and easily disrupted by contextual factors.

In pre-vocational contexts, where pupils may already experience low academic self-esteem and vulnerability to disengagement (Korpershoek et al., 2016), these affective dynamics are particularly consequential. The results of this study thus reinforce that CLIL's success with weaker scholastic aptitude learners depends not only on linguistic and didactic factors but also on how the program supports pupils' emotional resilience and sense of value.

L1 vs L2 use

The use of English as the target language was one of the most contested issues in the study, both in teachers' self-reports and in pupils' accounts. Teachers' estimates of their classroom English use varied enormously, with some reporting 20% or less, especially in lower *vmbo* tracks, and others claiming consistently high use across lessons. Pupils broadly confirmed these reports, for instance from School A, with the highest teacher self-report amount: *"We speak English with the teachers and also with each other. If we speak Dutch, the teachers correct us. Then they say, 'This is CLIL.'",* and the opinions of other focus groups were clear: they consistently wanted more English from their teachers. The lower *vmbo* level pupils also confirmed their teachers' less frequent use of the target language. As one Year 1 pupil commented, "*We think it's a pity that there's not more English. They don't speak English very often, really little.*" Similarly, a Year 3 pupil observed, "*The teachers are supposed to get the pupils involved with English. Now we just have to sit and listen. Or they ask something, and somebody answers in Dutch, and that was that.*"

Alongside these preferences, some pupils voiced frustration with inconsistent use of the L1 and L2. One focus group participant explained, *"In the lessons they speak Dutch and English all mixed up, and then we get completely confused."* This comment highlights that pupils expect teachers not only to increase their English use but also to provide clearer language boundaries: "*The teacher starts a sentence in English but then switches: 'Good morning,* wij gaan aan het werk'." *That doesn't set a good example.*" For pupils, the teacher is a crucial language model, and unpredictable code-switching between Dutch and English undermined both comprehension and their perception of what counted as legitimate language learning.

These findings reflect long-standing debates in Dutch foreign language education. The *Doeltaal - Voertaal* principle (Westhoff, 1998) emphasized maximizing the target language as the medium of instruction, whereas the more recent *Doeltaal - Leertaal* approach (Dönszelmann, 2020, 2024) advocates a more balanced stance: using the target language not just as the vehicle but as the object of learning, while strategically incorporating L1 to scaffold comprehension (Gierlinger, 2015). Internationally, this aligns with critiques of the strict 'monolingual principle' (Gallagher & Colohan, 2017), which argue that maximizing L2 at all costs may disadvantage weaker learners, particularly in content-based instruction.

Pupils in this study, however, did not reject the presence of English even when it posed difficulties, as long as the teachers were willing to spend extra time and effort explaining, repeating, and ensuring comprehensibility. On the contrary, they often equated more English with higher quality CLIL and interpreted limited English use as underestimating their ability. At the same time, their criticism of confusing code-switching points to the need for teachers to signal language choice clearly and to model target language use consistently. These findings suggest that a productive response lies not in reducing L2 input, but in making it comprehensible and accessible through scaffolding strategies such as visuals, structured tasks, and peer collaboration (Mahan, 2022). Yet teachers voiced real dilemmas: subject specialists in particular feared that weaker pupils were not grasping essential content when English predominated, leaving them torn between loyalty to their subject specialty and the CLIL imperative to maintain target language exposure.

This tension—between maximizing L2 input, safeguarding content comprehension, and addressing the needs of weaker learners—sits at the heart of CLIL in the *vmbo* context. Pupils' clear preference for more English, paired with their call for clarity and consistency, suggests that reducing L2 may be a misplaced form of protection. Instead, the findings point toward *Doeltaal – Leertaal* as a productive framework: not abandoning the target language, but embedding scaffolding, translanguaging, and strategic L1 use to ensure that pupils can engage with both language and content. In some cases, this may mean that teachers' own level of English needs improvement to enable a more flexible and effective use of the target language. In this way, teachers can reconcile the demands of CLIL methodology with the realities of teaching pupils who have lower scholastic aptitude and low target language proficiency but generally high motivation to learn English.

5.6B Integrating findings across data sources

A key characteristic of this study lies in its multi-perspective, mixed-methods design, which combined teacher questionnaires, teacher self-reports of target language use, pupil questionnaires, and pupil focus groups across six *vmbo* schools. Taken together, these data sources provide a nuanced and triangulated picture of CLIL practice in pre-vocational education. Several points of convergence and divergence stand out.

First, English learning and target language use were consistently highlighted across instruments. Teachers and pupils alike identified English improvement as the major benefit of CLIL, and both groups acknowledged the wide variation in teacher use of English across schools and levels. The teacher self-reports provided quantitative evidence of this variation, and focus group pupils confirmed these differences, linking them directly to their satisfaction. Questionnaire data added weight: although a clear majority indicated they would choose CLIL again, in focus groups they often qualified this by emphasizing the importance of teachers speaking English more consistently. This triangulation sug-

gests that while overall enthusiasm for CLIL is high, its sustainability appears to depend heavily on consistent teacher use of the target language.

Second, didactic quality emerged as another cross-cutting theme. Teachers in questionnaires stressed systemic barriers such as time, materials, and lack of collaboration, which limited their ability to realize CLIL's dual focus. Pupils, in contrast, evaluated lessons primarily through their lived experience: whether activities were fun, active, and varied. Focus group data revealed how the perception of broken promises (e.g., excursions that did not take place) could erode trust in the entire CLIL program. Here, teacher and pupil perspectives converged on the importance of lesson design, but diverged in emphasis: teachers framed didactics as a structural challenge, while pupils judged them experientially. Both perspectives, however, point to uneven realization of CLIL's core principles of balancing content and language learning, offering authentic communicative opportunities, and maintaining program-wide coherence.

Notably, teachers with greater CLIL experience tended to articulate these challenges in more specific terms, including explicit calls for further professional development, whereas less experienced teachers mentioned such needs less frequently. This pattern is consistent with broader research on teacher learning, which suggests that increasing experience can lead to greater awareness of instructional complexity and more differentiated perceptions of professional development needs (Borko, 2004). The earliest study of CLIL in several *vmbo* schools, *Proud to be t-vmbo* (Tanner & de Graaff, 2011), notes that schools were primarily focused on improving teachers' English proficiency rather than providing systematic training in CLIL methodology. In light of this, the articulated need for professional development in the current study aligns with findings in the CLIL literature more broadly, where teachers across contexts consistently report a need for more structured pedagogical and collaborative training (Kim & Graham, 2022).

Third, affective factors surfaced consistently, though from different angles. Teachers reported professional enjoyment and renewed motivation in CLIL, but also frustration at pupils' uneven proficiency and behavior. Pupils highlighted enjoyment, pride, and motivation when supported by enthusiastic teachers, but also frustration, anxiety, and occasional feelings of being underestimated. The focus group data gave voice to these affective nuances, complementing the broader trends in the pupil questionnaire. This integration reveals the centrality of teacher affect: enthusiasm and empathy acted as amplifiers of positive pupil experiences, while frustration and underestimation intensified challenges.

Finally, the integration across data sources highlights important gaps in perception. Teachers often emphasized the obstacles posed by pupils' low English level, whereas pupils interpreted reductions in English input as underestimation and expressed a desire for

more English, including stricter classroom rules about speaking English. Similarly, teachers frequently cited workload and organizational issues as limiting CLIL's implementation, but pupils focused on immediate experiences of lessons and activities. Such divergences underscore the importance of including pupil voices: without them, the interpretation of CLIL in pre-vocational settings would remain incomplete, dominated by teacher concerns.

In sum, the data sources converge in demonstrating that *vmbo* CLIL is both highly valued and unevenly realized. Pupils' satisfaction, motivation, and affect are closely tied to the consistency of English use, the quality of didactic design, and the affective stance of teachers. Teachers' enthusiasm is genuine, but their concerns about feasibility and workload are equally real. By integrating across these perspectives, this study highlights both the opportunities and the fragility of CLIL at this level: its success depends on aligning teacher practice, institutional support, and pupil expectations, while improving the fidelity to CLIL principles.

5.6C Situating the findings in institutional and policy context

Alongside classroom-level dynamics, several structural and policy-level factors shaped the implementation of CLIL in *vmbo*. Teachers frequently cited lack of time, insufficient training, and weak coordination across departments as major barriers. These concerns resonate with the official Standard for bilingual education at *vmbo* level that was in force at the time, which required schools to provide at least 30% of total instructional time in the target language, ensure subject teachers achieved at least CEFR B2, integrate authentic and challenging materials, facilitate pupils' communication in the target language, embed internationalization into the curriculum, and encourage cross-curricular collaboration supported by school leadership. A related tension emerged in teachers' views on pupil selection: although the national standard framed pre-vocational CLIL as inclusive, teachers' suggestions to introduce selection procedures appeared closely linked to their concerns about pupils' low English proficiency and the resulting challenges for subject learning in CLIL.

The present study reveals significant variation in what can be termed implementation fidelity—that is, the degree to which schools adhered to the intended model of CLIL in the first few years. While schools formally adopted bilingual programs, their adherence to the standards was inconsistent. This was especially visible in three areas:

Target language use

Policy expectation: The Standard implied that the target language should be used for most of the instruction, though no fixed percentage was stipulated per lesson.

Practice: Some teachers reported using English as little as 20% of the time, particularly in lower tracks. Pupils confirmed these discrepancies, and, particularly in some higher years, voiced disappointment at the low amount of target language use. Many also expressed confusion when teachers switched between Dutch and English, particularly within a single utterance. This lack of clarity undermined the teacher's role as a language model.

Internationalization
Policy expectation: International exchanges, projects, and study visits were to be central to the CLIL EIO curriculum.

Practice: Pupils valued such opportunities highly, but when promised activities did not materialize, they expressed disappointment and sometimes questioned the worth of the program itself. Additionally, some pupils expressed the opinion that the in-class EIO activities were 'boring' and irrelevant.

Teacher qualifications and collaboration
Policy expectation: Subject teachers were to hold CEFR B2 in English and receive training in CLIL didactics, ideally working closely with English teachers in cross-curricular teams.

Practice: Many teachers reported little or no CLIL-specific preparation, and many perceived a lack of collaboration among CLIL team members, especially between subject and English teachers. As a result, subject teachers often felt ill-equipped or uninclined to balance content and language goals, especially when pupils' comprehension faltered. Furthermore, pupils in some focus groups were critical not only of the lack of attention to language in the subject lessons, especially corrective feedback, but also of various aspects of their teachers' English, which suggests that a B2 level may not be sufficient for CLIL teaching, particularly for the higher *vmbo* track pupils who need to achieve a B2 level themselves in some skills.

Finally, national policy placed constraints beyond the control of individual schools. Pupils in Year 3 expressed frustration at being required to switch back to Dutch for CLIL subjects that would become national exam subjects in Year 4. Although this policy reflected broader assessment regulations, it undermined the continuity of pupils' bilingual experience and, in their eyes, devalued the effort they had invested in learning through English and added an extra challenge to re-learn the subject terminology and concepts in Dutch.

Overall, these findings suggest that many of the challenges identified by teachers and pupils stemmed less from individual shortcomings than from systemic limitations in organizational support and early-stage policy design. In the early years of CLIL at the *vmbo* level, schools were navigating new expectations with limited resources and experience, which appears to have contributed to uneven implementation. While these

findings highlight the fragility of CLIL in such contexts, it is also important to note that since this research was conducted, there has been substantial growth and improvement in *vmbo* bilingual education. More schools now offer CLIL programs, teacher professional development has been expanded, and a new Standard has been written to refine and strengthen the framework. These developments are addressed in the next chapter.

5.6D Comparison with other CLIL contexts

A secondary aim of this study was to situate *vmbo* teachers' and pupils' perceptions within the broader international research on CLIL, particularly in contexts serving lower scholastic aptitude learners. While the international CLIL literature has historically focused on higher academic tracks, an increasing number of studies have explored CLIL in vocational, pre-vocational, and lower-secondary settings, offering valuable points of comparison.

Several findings from this study mirror those reported elsewhere. Most notably, teachers and pupils agreed that English proficiency gains were the clearest benefit of CLIL. This echoes research in a German *Hauptschule* setting, where Schwab (2014, 2015) also documented improved English outcomes and strong pupil motivation despite low initial proficiency. Similarly, Jäppinen (2005) found in Finnish lower-secondary CLIL that academically weaker learners could make meaningful progress in both language and content when appropriate scaffolding was in place. In both this study and those contexts, pupils expressed pride in their progress and motivation to continue with CLIL, challenging the assumption that CLIL is only viable for stronger academic learners.

Pupils' affective responses also resonate with other research. In line with Lasagabaster (2011) and Doiz et al. (2014b), this study found that motivation and enjoyment were central to the CLIL experience but were easily disrupted by inconsistency, low teacher expectations, or lack of support. The enthusiasm of teachers, their willingness to use humor, and their belief in pupils' potential were crucial motivators here, as in other studies highlighting the role of teacher affect for academically less confident learners (Dewaele, 2019; Dewaele & MacIntyre, 2014).

At the same time, the findings diverge in important ways. In some contexts, researchers have suggested that weaker learners may find CLIL cognitively overwhelming if the target language demands are too high (e.g., Otwinowska & Foryś, 2017). By contrast, *vmbo* pupils in this study consistently asked for *more* English, not less, and criticized teachers who reduced their English input or switched unpredictably between languages. This suggests that although comprehension remains a challenge, reducing target language exposure may undermine motivation and create perceptions of underestimation. Pupils in this study were willing to face the linguistic challenges, provided they received appropriate support through strategic translanguaging and scaffolding.

Another difference lies in organizational support. Pérez Cañado (2016) notes that teachers across Europe report implementation bottlenecks plaguing their CLIL programs, such as teacher training in CLIL methodology, resources, and ongoing professional development. The present study strongly confirms this observation but also shows how such systemic issues are experienced at pupil level: when exchanges or international projects promised by the CLIL program failed to materialize, pupils expressed disappointment and even questioned the legitimacy of the program. This particular sensitivity to broken promises appears especially pronounced in *vmbo* contexts, where pupil trust in school may already be fragile.

When compared to research in more academically selective secondary schools, several differences emerge. In pre-university contexts in the Netherlands (Admiraal et al., 2006) or in Sweden (Sandberg, 2017), pupils often emphasize how CLIL supports their long-term academic and professional ambitions, such as university entry, study abroad, or careers requiring high levels of English. By contrast, the *vmbo* pupils in this study rarely framed CLIL in terms of future opportunities, except in a few initial reasons for choosing the CLIL stream (which included parental advice). Their own focus was overwhelmingly on the here-and-now: whether lessons were engaging, whether teachers believed in them, and whether they felt successful in English today. This difference may reflect the more immediate vocational orientation of *vmbo* learners and suggests that their engagement is tied less to abstract future goals and more to concrete, enjoyable, and affirming classroom experiences.

Behavior and classroom dynamics also contrast sharply. Studies in higher academic streams often describe pupils as academically disciplined and self-regulated, with relatively few references to behavioral disruption (Coyle et al., 2010; Dalton-Puffer, 2007). In the present study, however, both teachers and pupils frequently referenced behavioral challenges: short attention spans, restlessness, uncooperativeness, and frustration when lessons were not sufficiently varied or comprehensible. These findings align with research on *vmbo* pupils as 'doers' who learn best through active, hands-on tasks, as described above. Thus, while CLIL may succeed in both contexts, the didactic strategies that underpin success differ markedly: what works for pupils in pre-academic tracks (lectures, extended reading, self-study) may fail with *vmbo* pupils unless lessons are interactive, scaffolded, varied, and perceived as enjoyable.

Finally, preferred activities also differ. Higher-track CLIL pupils often report appreciating cognitively demanding tasks that challenge their linguistic and academic skills, such as debates, essay writing, or extended project work (Lasagabaster, 2011). In contrast, *vmbo* pupils in this study repeatedly suggested improvements centered on activities that were fun, active, creative, and relevant, including role-plays, games, excursions, and teacher anecdotes. This contrast underscores the importance of tailoring CLIL didactics to learner

profiles: activities that stimulate motivation in academically stronger learners may not engage pupils in vocational tracks.

5.6E Implications and stakeholder recommendations for inclusive CLIL pedagogy

Viewed across contexts, these comparisons suggest that CLIL in weaker-learner contexts is not inherently unworkable, as sometimes assumed. On the contrary, evidence from *vmbo*, *Hauptschule*, and other lower-secondary settings shows that pupils can thrive affectively and linguistically at their own level, provided that implementation fidelity is high and scaffolding and translanguaging are systematically employed. What these contexts tend to share are vulnerabilities: weaker learners are more sensitive to teacher affect, to disengagement and short attention spans, to consistency in language use, and to organizational follow-through. Where CLIL principles are applied unevenly, these pupils risk losing motivation quickly. But where CLIL is implemented with care, they gain not only increased linguistic competence but also a sense of accomplishment and international orientation that might otherwise remain out of reach.

The present section addresses the second research question (RQ2), which concerned the recommendations of CLIL teachers and pupils for strengthening bilingual education in the pre-vocational sector. In addition to reporting on their experiences, both teachers and pupils offered numerous suggestions for strengthening CLIL in the *vmbo*. These recommendations highlight areas of consensus as well as divergence, providing insight into how different stakeholders envision the future of bilingual education in pre-vocational contexts.

Classroom-level improvements

Pupils were particularly forthcoming with ideas for improving classroom practices. Beyond their frequent calls for more consistent use of English, they articulated a wide range of didactic and pedagogical improvements, many of which showed striking insight into effective teaching and learning. Suggestions included more individual feedback from teachers, the use of authentic materials and realia, more variety in the lesson activities, and cross-curricular projects that would link English to other subjects and to the real world. These ideas went beyond calls to simply 'make lessons more fun' and demonstrated that pupils can be keen observers of teaching practice with a sense of what helps them learn best. Teachers' recommendations partly overlapped with these insights: they also recognized the need to strengthen CLIL didactics, but framed this primarily in terms of their need for training, scaffolding strategies, and resources to implement more ambitious lessons effectively.

Organizational and school-level improvements

At the school level, both groups identified problems of coordination and support. Teachers repeatedly stressed the need for better collaboration between English teachers and subject specialists, more planning time, and clearer school-wide CLIL policies. Some teachers expressed doubts that the lower *vmbo* pupils could handle the challenges of CLIL and recommended limiting the CLIL program to the higher *vmbo* groups, and some teachers even questioned whether CLIL should be offered to *vmbo* pupils in the lower *vmbo* streams or those with an uneven mastery of Dutch. Pupils, especially in the lower *vmbo* streams, voiced the perception that teachers were giving up on them too easily: "*…sometimes they just translate everything right away because they don't think we can understand it.*" They expressed disappointment when promised activities (e.g., international projects or excursions) did not take place, recommending more reliable follow-through and communication. Both perspectives converge on the importance of coherent and consistent implementation at the organizational level: schools need to have a unified policy about target language use and teachers need structural support to collaborate, while pupils need assurance that the program will deliver on its stated aims.

Policy-level improvements

At the policy level, teachers were especially vocal. They called for more systematic professional development in CLIL methodology, lighter teaching loads, and more institutional investment in bilingual education. A few teachers suggested introducing selection criteria for CLIL, an idea which runs counter to the non-selective, inclusive nature of CLIL at *vmbo* level. Pupils, by contrast, did not engage directly with workload or policy issues; instead, they criticized the Year 3 switch back to Dutch for exam subjects, which they perceived as undermining the value of their previous effort in CLIL. Although this is a national requirement beyond school control, it shows pupils' acute awareness of how policy decisions shape their learning trajectory.

Alignment and divergence

There is clear alignment between teacher and pupil recommendations in their emphasis on strengthening CLIL didactics and ensuring reliable program delivery. Both groups recognize that high-quality CLIL depends on consistent language use, engaging lesson design, and structural coherence across the school. There is also alignment to a certain degree concerning increasing the use of English in the classroom: teachers seem to want to use English more and introduce more communicative activities, but perceive hindrances in the pupils' language level and their reluctance to participate. Pupils, on the other hand, unanimously want more English, more corrective feedback, and more enjoyable communicative activities—even if they have to be 'forced' to use the target language.

At the same time, divergences emerge in how these recommendations are framed. Pupils tended to focus on concrete, immediate improvements (e.g., authentic materials,

projects, excursions, and individual feedback), while teachers emphasized systemic enablers (e.g., training, planning time, collaboration, possible selection of pupils). Similarly, only teachers raised concerns about workload, whereas only pupils criticized the Year 3 language-policy shift.

Taken as a whole, these recommendations reveal both the promise and the precariousness of CLIL at *vmbo* level. Pupils' voices, often underutilized in discussions of CLIL implementation, proved especially valuable: far from being superficial, their suggestions displayed significant educational merit and alignment with best practices in CLIL didactics and pedagogy. Teachers remain for the most part committed to CLIL but require stronger institutional and systemic support to realize its full potential. The overlap between teacher and pupil voices suggests that targeted improvements in didactic training, collaboration, and program fidelity would benefit both groups simultaneously, while divergences highlight the need for ongoing dialogue that takes pupils' insights as seriously as those of their teachers.

5.7 Limitations, recommendations for future research, and conclusion

While this study provides important insights into the perceptions of teachers and pupils involved in CLIL at the *vmbo* level, several limitations must be acknowledged. These limitations frame the scope of the findings and highlight directions for further research.

Early stage of CLIL implementation
The research was conducted during the early years of CLIL in the *vmbo*, when many schools were still experimenting with bilingual education and had relatively little experience in implementing CLIL principles. Since then, both the number of schools offering CLIL and the maturity of programs have increased, and a new Standard has been published. The findings therefore capture an early snapshot of CLIL in development, which does not fully represent current practice.

Self-report measures
Teachers' estimates of their own target language use were based on self-report, which may not always reflect actual classroom practice. Importantly, these estimates were largely congruent with pupil questionnaire responses and focus group comments, suggesting a reasonable degree of convergence across data sources. More robust methods such as classroom observations or audio recordings would nevertheless have provided stronger evidence. However, the 'observer's paradox' (Labov, 1972) may result in unreliable measurements in lesson observations, as teachers' target language use might be higher than usual when they know they are being observed or recorded.

Sample characteristics and generalizability

The study included six schools and fifteen class groups, with 53 teachers and 267 pupils (70 in focus groups). While this is a substantial sample for a qualitative study and included nearly half of the *vmbo* CLIL programs in English at the time, it is not necessarily representative of all CLIL *vmbo* schools in the Netherlands. The findings are therefore best understood as illustrative of trends and challenges in the early years of *vmbo* CLIL, rather than generalizable to all contexts. Similarly, while the findings may resonate with stakeholders in other lower scholastic aptitude contexts, they cannot be generalized to these contexts.

Methodological scope

The mixed-methods design enabled triangulation across questionnaires, self-reports, and focus groups, but it also had limitations. The teacher questionnaire, in keeping with the teachers' condition of brevity, was relatively short and descriptive, offering limited depth into teachers' beliefs. Conversely, the pupil focus groups generated rich qualitative data, but these discussions may have been influenced by group dynamics, peer pressure, or pupils' willingness to voice critical opinions to a stranger. It was also not possible for the researcher to ensure that those pupils (chosen by the teacher) expressed opinions that were representative of their class as a whole.

Contextual constraints

Several factors external to the study also shaped findings. National policy requiring a shift to Dutch in Year 3 for exam subjects was beyond school control, yet strongly influenced pupil perceptions of continuity. Similarly, the Standard at the time did not specify a minimum percentage of L2 use per lesson, leaving teachers to interpret 'most of the time' in their own way. Such contextual factors limit the extent to which observed variation can be attributed solely to individual teacher or school choices.

Researcher positionality

Finally, as in all qualitative research, researcher positionality may have influenced both data collection and interpretation. Although multiple data sources were used to triangulate findings, complete neutrality is unattainable, and interpretations inevitably reflect the researcher's own framing of CLIL in the *vmbo* context.

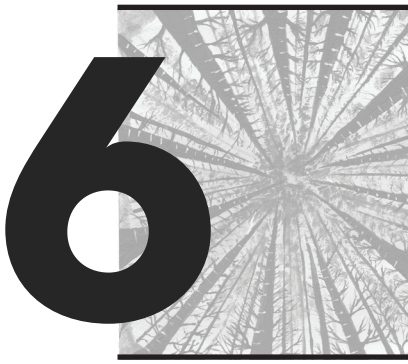
Summary

Considered together, these limitations situate the findings within a specific time, place, and methodological frame. The results should be seen as highlighting key patterns and tensions in the early implementation of CLIL in Dutch *vmbo*, offering a foundation for future research to build upon and refine.

5.7A Conclusion

In summary, this chapter has examined how teachers and pupils in Dutch pre-vocational CLIL programs experienced and evaluated bilingual education during its formative years. The findings reveal pronounced variation among schools in the organization and delivery of CLIL: the proportion of English used in lessons, the depth of CLIL pedagogy and communicative task design, the inclusion of international activities, and the level of collaboration between English and subject teachers appeared to differ considerably. Such programmatic variation also surfaced in participants' perceptions. Both teachers and pupils in the more theoretical *vmbo* streams generally tended to describe CLIL as stimulating and rewarding, whereas those in the more practical streams, while guardedly positive, expressed greater uncertainty about language demands, workload, CLIL pedagogy, and the balance between content and language. School-specific factors further shaped these experiences. In particular, School A, which hosted established *havo* and *vwo* CLIL stream alongside its *vmbo* program, appeared to benefit from the ripple effect of nearby expertise and a more cohesive bilingual culture of teaching and learning.

Importantly, the study also underscores the value of listening closely to learner voices: pupils' perceptions revealed a clear gap between teachers' actual use of English and the greater target language exposure they desired, with some describing feelings of being underestimated when this gap widened. Overall, these results underline that pre-vocational CLIL in the Netherlands is far from uniform; it reflects multiple local interpretations of the model, shaped by teacher proficiency, institutional history, pupil profile, and program affordances. The diversity observed here provides an essential foundation for the broader synthesis and recommendations presented in the next chapter, where the implications for inclusive and sustainable CLIL practice across educational levels are discussed.



6

CLIL at *vmbo*: Conditions, Constraints, and Conclusions

6.1. Background

6.1A Rationale and relevance

Over the past three decades, bilingual education in Europe has been dominated by CLIL models implemented in academically selective contexts. Dutch bilingual *tweetalig onderwijs* (TTO) programs, which apply CLIL methodology, have achieved notable linguistic and motivational success among pupils in the upper secondary tracks (*vwo* and *havo*). Yet these successes raise a critical question: if CLIL demonstrably benefits learners in higher tracks, should it not also be accessible to those with lower scholastic aptitude? This dissertation addressed that ethical and pedagogical question by investigating whether CLIL is linguistically feasible and motivationally sustainable for pupils—and, to a more limited extent, for their teachers—in the early years of the pre-vocational (*vmbo*) stream of Dutch secondary education.

The research did not attempt to evaluate all aims of Dutch bilingual education—such as subject content attainment, maintenance of Dutch proficiency, or the full range of European and International Orientation (EIO) activities.[15] Rather, the research focused on two interrelated dimensions: (1) the development of English proficiency and (2) the attitudes and perceptions surrounding CLIL teaching and learning. These dimensions represent the most immediate and measurable outcomes of early *vmbo* CLIL, and together they offer insight into how bilingual education functioned for a group long excluded from such opportunities. By examining both learner and teacher perspectives, the study situates pre-vocational CLIL not merely as a pedagogical innovation but as an inquiry into educational equity, inclusion, and professional adaptation in a streamed national system.

6.1B Theoretical underpinnings

The study was anchored in complementary foundations of applied linguistics, cognitive theory, and pedagogical contextualization, each contributing to an understanding of the challenges involved in adapting CLIL for diverse learners in pre-vocational education.

At the psycholinguistic and applied linguistic level, this study draws first on theories that explain how bilingual and second-language learning interact with cognitive development and learner aptitude. Central among these are Cummins' (1979b) Threshold Hypothesis and Sweller's (2011; 2017) Cognitive Load Theory, both of which address the balance between linguistic and cognitive demands in dual-focus instruction. Later interpreted as a limitation for weaker learners, this view assumed that pupils with limited L2 proficiency might not benefit fully from bilingual education. However, Genesee (2007) demonstrated that bilingual instruction can also support learners who are less academically advantaged, provided that instruction is well scaffolded and the language input is

15 An initial attempt to measure pupils' Dutch proficiency was ultimately excluded because the testing conditions and collected data did not permit reliable interpretation.

meaningful. Zydatiß (2007, 2012) refined this notion by suggesting that bilingual learners operate between two thresholds, a minimal and an optimal one—thereby allowing for developmental progression within bilingual programs rather than a single pass/fail cutoff. Together, these frameworks justify examining whether pre-vocational CLIL can succeed when linguistic and cognitive load are carefully managed and pedagogical scaffolding is prioritized.

At the pedagogical level, the CLIL model itself provides a comprehensive framework for integrating content and language learning. Coyle's (2006b; Coyle et al., 2010) *4Cs model*—Content, Communication, Cognition, and Culture—offers a coherent theoretical lens for understanding how learning tasks can simultaneously promote subject mastery and language development. Her related triad of *language of learning*, *language for learning*, and *language through learning* captures the interplay between linguistic input, classroom discourse, and cognitive engagement. Within this framework, effective CLIL instruction depends on teachers' ability to sequence tasks, model academic discourse, and scaffold both content and linguistic objectives in balance. For lower-aptitude learners, these principles underscore the importance of explicit language support, gradual release of responsibility through scaffolding (Hammond & Gibbons, 2005), and sustained interactional practice as means of managing both cognitive and linguistic load.

At the contextual and inclusive level, emerging scholarship on CLIL in the late 2000s and early 2010s highlighted the need for adaptation and contextualization in response to learner diversity. Rather than prescribing a fixed model, researchers emphasized that CLIL's success depends on how sensitively it is adjusted to local educational cultures and pupil profiles (Banegas, 2012; Dalton-Puffer, 2008b; Lasagabaster & Sierra, 2010; Marsh et al., 2009b). This view repositions CLIL as a flexible, accessible approach rather than an elite pathway, arguing that its viability depends on realistic linguistic expectations, responsive pedagogy, and attention to contextual variation. More recent contributions to the international discussion on diversity, equity, and inclusion within bilingual education (e.g., Genesee & Fortune, 2014; Pérez Cañado, 2023; Somers, 2017) further underscore that appropriately adapted CLIL environments can offer meaningful opportunities for learners with widely differing profiles, including those with lower scholastic aptitude or multilingual home backgrounds. Within this perspective, progress toward sustainable *vmbo* CLIL pedagogy is characterized by efforts to foster engagement and meaningful participation of all pupils while supporting them through attentive, responsive interaction as they work toward attainable language and content goals.

6.1C Context of the Study

As outlined in Chapters 1–4, the Dutch secondary education system is highly streamed, with pupils placed after primary school into one of several tracks reflecting their scholastic aptitude. The pre-vocational track (*voorbereidend middelbaar beroepsonderwijs*, *vmbo*) serves learners who benefit from more applied, practice-oriented forms of learning and who typically progress to senior vocational education (*mbo*). This structural differentiation simultaneously provides both the rationale for and the principal challenge to implementing Content and Language Integrated Learning (CLIL) in inclusive, non-selective contexts.

Dutch bilingual education (*tweetalig onderwijs*, TTO) applies the CLIL approach across subjects, combining content learning with systematic attention to language. National guidelines stipulate that at least 30% of the *curriculum* (not necessarily 30% of class time or teacher talk) be delivered through English. Whereas early TTO initiatives began in the 1990s in the academically selective *vwo* track and later extended to *havo* (both with a minimum of 50% of the curriculum through English), the first *vmbo* pilots were launched only around 2009. Their late arrival reflects a long-standing assumption that bilingual education might prove too demanding for less academically advantaged learners. In response, the *vmbo* TTO Standard was deliberately adapted to emphasize accessibility and inclusion: reduced exposure targets, differentiated CEFR outcomes, flexible lesson design, and a mandatory European and International Orientation (EIO) component encouraging intercultural awareness and project-based learning.

A further defining feature of Dutch bilingual education is that language-attainment standards are adapted to the educational track. In *vmbo* CLIL, pupils follow a shorter bilingual trajectory than their peers in the higher academic tracks and work toward correspondingly differentiated outcomes on the CEFR scale, designed to support the inclusivity of CLIL at this level. Within the *vmbo* stream itself, the required attainment levels are formally specified in the *vmbo* TTO Standard (Nuffic, n.d.-b) and vary by subtrack: the *vmbo-t* program prescribes higher linguistic goals than the *kader* or *basis* programs. This differentiation of language-attainment standards reflects a pedagogically realistic policy choice rather than a limitation: it acknowledges cognitive diversity, aligns expectations with attainable progress, and prevents unrealistic comparison with more academically oriented learners.

In addition to these structural features, *vmbo* CLIL operates within a wider educational climate in which the *vmbo* sector continues to contend with a persistently negative public image. Despite broad acknowledgement from employers and educational experts that its practice-oriented competencies are essential to Dutch society, *vmbo* is still widely perceived in the public sphere as a less desirable pathway. Examining CLIL in this reputational context adds an important societal dimension to the research: its emphasis on functional English proficiency and international orientation offers *vmbo* pupils forms of

linguistic and intercultural capital increasingly valued in education, employment, and civic life.

Within this evolving policy landscape, the present research examined the formative stage of *vmbo* CLIL implementation, with data gathered between 2012–2014. The participating CLIL schools—six in total, representing all-CLIL schools as well as schools with a parallel 'regular' Dutch stream—differed widely in experience, teacher proficiency, and institutional resources, offering a cross-section of early practice. The teacher group spanned English and subject specialists with highly variable linguistic backgrounds, while the convenience sample of 603 pupil participants (CLIL n=313; non-CLIL n=290) represented all four *vmbo* levels and diverse linguistic and cultural profiles. Taken together, these settings provided a uniquely heterogeneous yet coherent sample through which to explore how CLIL could take shape in the pre-vocational sector of Dutch education.

6.1D Overview of the research design and chapter logic

Chapter 1 established the rationale for investigating Content and Language Integrated Learning (CLIL) in Dutch pre-vocational secondary education (*vmbo*). It outlined the ethical and pedagogical motivation for extending bilingual opportunities to learners with traditionally more limited access to selective *tweetalig onderwijs* (TTO) programs and formulated the overarching research aim: to examine the linguistic and affective feasibility of CLIL for this population. The chapter also introduced the mixed-methods design that integrates longitudinal quantitative data with qualitative insights from teachers and pupils.

Chapter 2 addressed attitudinal factors through the lens of the Model of Planned Behavior (Ajzen, 1991). Using data from nearly 500 pupils across CLIL and non-CLIL classes, it examined how *cognition*, *affect*, *subjective norm*, and *perceived behavioral control* predicted learners' intention to make an effort to learn English. The findings established a psychological baseline for understanding motivation in early-stage *vmbo* CLIL.

Chapter 3 compared CLIL and non-CLIL pupils' growth in receptive and controlled productive English proficiency—respectively, for vocabulary and grammar—over three years. These analyses provided the first longitudinal evidence on whether CLIL pupils at the lower-academic tracks made comparable or superior linguistic gains relative to their *vmbo* peers in regular programs.

Chapter 4 turned to productive skills, investigating writing proficiency, oral proficiency, and willingness to communicate in English. Together, Chapters 3 and 4 captured how CLIL exposure translated into measurable linguistic outcomes, complementing the attitudinal dimensions explored in Chapter 2.

Chapter 5, drawing primarily on qualitative methods, examined the perceptions of teachers and pupils regarding the strengths, weaknesses, and future potential of CLIL in the pre-vocational context. This chapter added the stakeholder dimension necessary to interpret the quantitative findings in light of classroom realities, particularly teacher language use, pupil affect, and program diversity across schools.

Taken together, these five chapters, discussed in more detail in 6.2 Chapter Core Findings, constitute a coherent mixed-methods investigation of CLIL's linguistic and affective dimensions in a non-selective setting. Rather than attempting to evaluate the full range of TTO aims, they provide a multi-perspective picture of English proficiency development and stakeholder perceptions during the formative years of pre-vocational CLIL implementation in the Netherlands.

6.1E Significance

This dissertation represents the first large-scale, mixed-methods investigation of CLIL implementation in Dutch pre-vocational secondary education. Conducted during the formative years of *vmbo* TTO, it captured both linguistic development and the attitudinal/perceptual dimensions of CLIL among teachers and pupils at a time when few comparable data existed. Despite the passage of more than a decade since the start of this research, it remains the most comprehensive empirical account of *vmbo* CLIL in the Netherlands, and subsequent work appears to have remained scarce and largely small-scale. As such, the present study functions as an effective baseline or zero-measurement for evaluating the longer-term development of inclusive, accessible, and sustainable CLIL practice in the lower academic tracks. This significance is underscored by broader reviews of the field: Pérez Cañado and colleagues in the ADiBE project (Pérez Cañado, 2023) have repeatedly highlighted persistent gaps in large-scale, longitudinal, and context-diverse CLIL research, particularly beyond elite settings, while Cenoz (2015) has called for more differentiated, context-sensitive designs that address learner variability rather than relying on uniform models.

By integrating quantitative growth data on English proficiency with qualitative insights into stakeholder perceptions, the study illuminates how CLIL can extend bilingual learning opportunities to traditionally underserved populations while foregrounding the practical constraints that shape implementation. The sections that follow synthesize the principal findings of each empirical chapter (2–5), tracing how attitudes, proficiency growth, and stakeholder perceptions intersect across diverse *vmbo* contexts.

6.2 Chapter core findings

6.2A Chapter 2 core findings: attitudinal factors and the intention to learn English

Chapter 2 examined how pre-vocational pupils' attitudes towards learning English evolved over three measurement points across two years in CLIL and non-CLIL settings, applying the Model of Planned Behavior (MPB) (Ajzen, 1991) to nearly 500 pupils in *vmbo* years 1–3. The analyses combined growth-focused comparisons (using change scores across the three time points) with structural modelling based on the MPB framework. Confirmatory factor analyses established a good model fit and demonstrated measurement invariance across CLIL and non-CLIL groups.

Within this MPB model, *cognition*—the perceived usefulness of making an effort to learn English—emerged as the strongest predictor of learners' *intention* to make that effort. The structural model further showed that *affect* and *subjective norm* were strongly correlated with both *cognition* and i*ntention*, but they did not explain additional variance once *cognition* was controlled. *Perceived behavioral control (PBC)* was weakly correlated with *intention*, and its direct path in the structural model was not significant, indicating that perceived obstacles had minimal independent impact on pupils' attitudinal outcomes. Correlations and structural equation modeling (SEM) paths come from different analytical steps; however, both address related aspects of the model and lead to the same overall interpretation.

CLIL pupils scored significantly higher than their non-CLIL peers on four of the five MPB constructs—*cognition*, *affect*, *PBC*, and *intention*. These differences were identified using multilevel analyses of class-level variance. Roughly 40% of between-class variance in *cognition*, *affect*, and *intention*, and 25% in *PBC*, was associated with CLIL classes. Importantly, the groups were comparable at the start of secondary school (cohort 1): neither *cognition* nor *affect* differed significantly at baseline, indicating that the CLIL advantage developed during schooling rather than through pre-selection. The difference in *affect* peaked in year 2 and diminished again by the end of year 3, after English-medium exposure was reduced. A small but sustained advantage in *PBC* suggested that gains in self-efficacy and autonomy persisted even as program intensity decreased.

Taken together, these results indicate that CLIL participation strengthened both the perceived value and enjoyment of English learning, especially during the more intensive first half of the program. The affective advantage waned once exposure declined, but the motivational benefits linked to perceived usefulness and control remained. Overall, this pattern suggests that CLIL participation strengthened the attitudinal components linked to pupils' intention to learn English, even among less academically advantaged learners.

6.2B Chapter 3 core findings: vocabulary and grammar

The second empirical study examined whether CLIL pupils in pre-vocational education developed greater receptive command of English vocabulary and grammar than their non-CLIL peers. The two-year longitudinal study used tests of vocabulary breadth and contextual grammar at three measurement moments and applied multilevel growth modelling to account for the nested data structure (pupils within classes and schools). Educational track, gender, and home language were entered as covariates to isolate the association between CLIL participation and receptive linguistic growth.

Results showed that CLIL pupils scored significantly higher than their non-CLIL peers in both vocabulary and grammar after controlling for class and school differences. From the outset, these differences amounted to 5.18 points (max = 48) for vocabulary and 2.0 points for grammar (max=36), and CLIL pupils continued to extend this lead over two years. Growth analyses revealed that the vocabulary advantage was consistently stronger, with CLIL pupils gaining an additional 0.016 points per week compared with non-CLIL pupils, explaining 0.8 percent of total variance in vocabulary gains and nearly 7 percent of variance between classes. In contrast, grammar growth was more modest, accounting for 0.5 percent of total variance and 3.7 percent of growth variance. The strongest gains occurred in the first two years of the program, coinciding with the period of maximal CLIL exposure.

Minor but expected effects were also observed for school year and *vmbo* track, with higher-track pupils and those in year two scoring slightly better overall. As expected, pupils in higher tracks progressed faster overall, yet CLIL's relative benefit for grammar was somewhat greater in lower-track classes. Grammar scores for both groups in the year 3 cohort declined slightly, coinciding with reduced CLIL intensity.

Overall, CLIL participation was associated with small but statistically reliable advantages in both vocabulary and grammar. Vocabulary growth, in particular, was strongest during years with greater English-medium exposure and slowed as exposure declined, indicating a clear temporal pattern in pupils' receptive development.

6.2C Chapter 4 core findings: writing, speaking, and Willingness to Communicate

The third study extended the investigation to productive English skills and willingness to communicate (WtC), examining whether CLIL pupils in the pre-vocational track developed stronger writing and speaking proficiency and fluency than their non-CLIL peers. Writing and speaking were each analyzed for both overall proficiency and fluency (total number of English words produced), using multilevel regression analyses that accounted for the nested data structure of pupils within classes and schools. Complementary self-

report questionnaires on WtC explored pupils' confidence and willingness to use English in both familiar (WtC-'School') and unfamiliar (WtC-'London') contexts.

Results revealed a significant CLIL advantage for writing proficiency and fluency, though effect sizes were small: CLIL explained about 2 percent of relative growth in writing proficiency and 1 percent in writing fluency. Speaking proficiency likewise showed a small but consistent CLIL advantage in both analytic approaches (ANCOVA and change-score). Speaking fluency, however, presented a more nuanced pattern: the ANCOVA model showed that CLIL pupils produced on average about 20 more words per task, while the change-score analysis followed the same positive trend but did not reach significance. Following standard practice for non-randomized groups (van Breukelen, 2013), the change-score result was treated as the more conservative estimate. Even so, the pattern indicates a fluency benefit for the youngest CLIL pupils (cohort 1), whose oral output increased most strongly over time. A marginal moderation effect further suggested that pupils with lower primary-school Cito aptitude scores benefitted slightly more in speaking fluency than their higher-aptitude peers.

CLIL pupils also rated themselves higher in WtC in the school context, though not in the unfamiliar (London) context. Both WtC-'School' and WtC-'London' scores increased significantly for all pupils, yet growth in WtC did not differ significantly between CLIL and non-CLIL groups after adding interaction terms. CLIL pupils maintained higher self-reported confidence in school-based communication, but no additional gains in unfamiliar contexts. Minor contextual influences, such as *vmbo* subtrack, grade, and teacher L2 use, accounted for variation between schools. The study also yielded slightly different outcomes depending on the analytic approach (ANCOVA versus change-score analysis), a methodological point addressed later under Section 6.7.

Overall, the findings show that pre-vocational CLIL pupils achieved statistically significant gains in productive English proficiency relative to their non-CLIL peers, particularly in writing and speaking, while fluency gains were more modest and context-dependent. Although ANCOVA and change-score results differed slightly in strength, both analyses point in the same direction of advantage. The overall pattern suggests that CLIL was associated with small but consistent benefits in productive skills and in pupils' confidence when communicating in English at school, even though the statistical effects remained modest across analytic approaches.

6.2D Chapter 5 core findings: teacher and pupil perspectives

Chapter 5, based on questionnaires and focus groups with CLIL teachers and pupils across six *vmbo* schools, explored stakeholder perspectives within CLIL programs only. Unlike the preceding quantitative studies, it did not include non-CLIL comparison groups, focusing instead on the experiences and reflections of those directly involved in CLIL implementa-

tion. These qualitative insights deepen the statistical findings by showing how pupils and teachers perceived the strengths and limitations of *vmbo* CLIL in everyday practice.

Both pupils and teachers most often cited pupils' progress in English as the most positive aspect of CLIL. Pupils acknowledged the difficulty of learning through English, yet generally viewed well-implemented lessons as fun, different, and active, associating them with confidence, interaction, and enjoyment. Some, however, felt that lessons lacked sufficient target language input or communicative focus. Nearly three-quarters (74%) indicated they would choose CLIL again, suggesting that its perceived benefits outweighed its challenges. Teachers likewise described CLIL as engaging and motivating, though several noted inconsistencies in language use, workload pressures, and a need for greater professional collaboration.

Comprehension of language and subject matter emerged as a recurring tension. Several teachers questioned the feasibility of CLIL for pupils in the lower *vmbo* tracks, while many pupils argued that a better balance between English and Dutch was needed to ensure understanding. Pupils valued teachers who communicated clearly and used English extensively or alternated strategically between languages, whereas inconsistent English use or limited explanation in either language sometimes led to frustration or disengagement. These perceptions mirrored teachers' own variation in target language use.

Teacher estimates of classroom English use ranged widely—from nearly continuous use in some programs to less than 20% in others—reflecting local differences in proficiency, policy, and emphasis on subject comprehension. Pupils confirmed these patterns: some were motivated by consistent English exposure, while others found fluctuating language practices confusing or demotivating.

Both pupils and teachers underscored the importance of interaction and task variety. Pupils linked communicative and hands-on activities to enjoyment and progress, while teachers valued the variety and authenticity of CLIL tasks despite the additional planning they required. At the same time, several pupils noted that language feedback was limited and that attention to form in content subjects was rare, potentially constraining opportunities for language development.

Beyond language and methodology, many pupils stressed the importance of the teacher–pupil relationship. They associated effective CLIL teaching with patience, personal attention, and teachers' awareness of how challenging it could be to learn through English. Teachers likewise emphasized empathy, noting that behavioral issues often stemmed from frustration when comprehension broke down. These relational factors appeared to shape pupils' confidence as much as task design or language input, underscoring the interpersonal dimension of CLIL classrooms later discussed in Section 6.3C.

Taken together, these perspectives reveal that *vmbo* CLIL's effectiveness and appeal are shaped by the clarity, empathy, and linguistic responsiveness of teachers, alongside the extent and quality of English used. These findings provide a qualitative bridge between the proficiency and attitudinal patterns identified in Chapters 2–4 and the broader themes of cognition, affect, and program diversity discussed in 6.3A–D.

6.3 Overarching analytical themes

This section draws together findings from the four empirical studies to interpret how CLIL operates linguistically, affectively, and contextually in the pre-vocational stream. It identifies four interrelated themes—inclusion and paradox**,** cognition and motivation**,** affective alignment and agency**,** and program diversity and local interpretation—that illuminate the mechanisms through which *vmbo* CLIL supports or constrains pupils' progress and engagement. Together, they illustrate how CLIL can foster language growth and learner engagement under diverse local conditions, while also revealing the tensions that shape its inclusiveness, sustainability, and pedagogical coherence.

6.3A Inclusion and paradox

Across the four studies, pre-vocational CLIL emerges as both an ambitious extension of bilingual education and a revealing test of its structural limits. Introduced into a system long characterized by early streaming and academic stratification, *vmbo* CLIL aimed to extend the motivational and linguistic benefits of bilingual learning to pupils traditionally excluded from it. The findings confirm that this goal was not misplaced: CLIL pupils showed more growth over time than their non-CLIL peers in most assessed areas of English vocabulary, grammar, and writing, and speaking, and expressed more positive attitudes toward learning English. Teachers and pupils alike generally viewed the CLIL environment as engaging and effective for language learning. Yet these same results also expose an underlying paradox. Implementing a cognitively and linguistically demanding pedagogy within a population selected precisely for its more practical orientation and lower scholastic aptitude requires thoughtful pedagogical adaptation.

This tension was evident in classroom practice. Teachers often grappled with the need to balance accessibility and authenticity, in other words maintaining meaningful content instruction while using English extensively with learners whose receptive skills were still developing. A few pupils also remarked that language feedback was rare and that subject lessons offered little explicit attention to English itself, apart from its use as a medium of instruction. One pupil expressed this clearly: *"English is about English, Religious Studies is about Religious Studies, not about English. It's not like you suddenly talk about the English language during a Biology lesson or something like that."* While CLIL is intended to integrate content and language learning, some empirical studies note that in practice many lessons

become primarily content-oriented, with limited explicit focus on language form (Lim Falk, 2008, as cited in Dalton-Puffer, 2011, p. 5; Villabona & Cenoz, 2022). In such cases, the linguistic dimension remains largely implicit, reducing the 'dual focus' central to CLIL. The limited provision of corrective feedback may therefore have reduced opportunities for noticing and reformulating language—processes central to interlanguage development (Lyster, 2007; Swain, 1985)—and could help to explain the modest proficiency gains observed in Chapters 3 and 4.

Pupils, in turn, oscillated between pride and frustration: they celebrated their growing command of English but were acutely aware of comprehension gaps and inconsistent teacher language use. The results show that *vmbo* CLIL can indeed foster linguistic growth (Chapters 3–4) and positive engagement (Chapter 5), but that success appeared to depend on varied, meaningful, and enjoyable communicative activities, careful scaffolding, flexibility in balancing language and content, and recognition of these pupils as 'doers': *"Pupils are well-engaged with CLIL, especially with 'hands-on' activities."* This is not a deficit narrative but a call for contextualized pedagogy that recognizes that linguistic simplicity does not equate to cognitive simplicity, and that inclusion in CLIL may require adaptation of input, task design, and teacher support. In this sense, pre-vocational CLIL embodies both the potential of educational equity and the complexity of realizing it within the constraints of a streamed, high-stakes educational system.

Taken together, the quantitative and qualitative findings reinforce this duality. The significant evidence of English-language gains (Chapters 2–4) demonstrates that CLIL can raise linguistic proficiency and pupils' attitudinal engagement without academic selection, while the focus-group perspectives (Chapter 5) reveal how fragile those gains remain when institutional and linguistic supports are uneven. Both strands converge on the same conclusion: *vmbo* CLIL's success lies not in replicating the conditions of high-track bilingual education but in re-imagining them for a more heterogeneous learner population—one whose progress depends on pedagogical empathy (Chapter 5) as well as on pupils' perceived relevance of English to their own lives and futures (Chapter 2).

6.3B *Cognition*, motivation, and temporal dynamics

A recurring pattern in the attitudinal analysis (Chapter 2) is the central role of *cognition*—understood in the Model of Planned Behavior as the perceived usefulness or importance of learning English—in shaping pupils' engagement and sustained effort. As noted earlier, both zero-order correlations and structural modelling were used to examine relations among the MPB constructs. In the structural model, *cognition* was the strongest and only significant predictor of *intention*, while *affect* and *subjective norm* showed strong zero-order correlations with intention but did not contribute uniquely once *cognition* was controlled. This pattern suggests that *vmbo* pupils' motivation to learn English rests less on emotional enjoyment or external pressure than on a growing awareness of its

instrumental and social value. The finding is particularly meaningful in a population often described as more practically than academically oriented: perceiving English as a tool for short-term goals or long-term success may serve as a stable motivational anchor even when self-efficacy is fragile or extrinsic incentives wane. Still, many pupils expressed that their day-to-day engagement depended less on perceived utility than on the immediate enjoyment and variety of classroom activities, as reflected in their frequent references to CLIL lessons being "fun," "different," or "active" (Chapter 5).

These patterns suggest that maintaining motivation and progress may depend on how well schools and teachers preserve variety, challenge, and sustained exposure in the later years of the program. The attitudinal differences (Chapter 2) between CLIL and non-CLIL pupils widened during the first two years, most notably in *affect* and *perceived behavioral control*, but tended to plateau or recede by the end of the second year. A similar trajectory appeared in proficiency growth: vocabulary and grammar gains (Chapter 3) were significant but tapered over time, and positive attitudes toward learning English did not increase in the upper years. This indicates that while CLIL can initially stimulate enthusiasm and effort, momentum may decline if exposure and classroom variety decrease, suggesting that sustaining motivation may require deliberate renewal.

At the same time, the coexistence of measurable proficiency gains and only moderate attitudinal divergence points to a nuanced interplay between cognitive and affective factors. Pupils' sense of the usefulness of English (*cognition*) may help sustain their effort even when enjoyment (*affect*) fluctuates (Chapter 2), a finding consistent with Dörnyei's (2005, 2009) dynamic models of L2 motivation. In *vmbo* CLIL, motivation thus appears cyclical rather than cumulative: initial enthusiasm feeds effort and achievement, which in turn reinforce pupils' sense of usefulness, though this cycle seemed to weaken as exposure decreased or as opportunities for meaningful communication declined: "*...they say that there will occasionally be a lesson in English but they never get around to it, they forget to do it...*" (year 3).

Taken together, these findings portray a developmental curve of early CLIL attitudinal engagement—rapid initial growth followed by stabilization—that underscores both the power and the fragility of affective engagement in lower-track bilingual programs. Sustaining *vmbo* pupils' motivation requires not only meaningful tasks and communicative variety, but also the explicit linking of English learning to attainable personal and vocational goals, as suggested by teachers' and pupils' reflections in Chapter 5.

6.3C Affective alignment and agency

In the *vmbo* context, inclusive CLIL emerged as a relational as well as a linguistic enterprise. As one teacher observed, "*In CLIL, there is a closer relationship between teacher and pupil.*" Across the studies, this relational dimension proved fundamental: pupils' enjoy-

ment of English and their sense of agency repeatedly appeared to sustain motivation. The attitudinal results (Chapter 2) showed that affective engagement—liking English and believing it matters—was almost as influential as perceived usefulness (cognition) in predicting pupils' intention to invest effort. Similarly, the qualitative findings (Chapter 5) revealed that pupils most often described CLIL lessons as "fun," "different," or "active," linking enjoyment with confidence and willingness to participate. This enjoyment was not trivial: it provided emotional scaffolding for perseverance in a program that often challenged pupils' linguistic confidence and self-belief ("*I don't understand anything… I have no idea what's going on.*")

Affective benefits, however, depended strongly on alignment between pupil expectations and teacher behavior. Pupils reported feeling most engaged when teachers were patient, encouraging, and communicative in English, showing understanding for how difficult CLIL could sometimes feel. As one pupil advised, "*Have a lot of patience and understand that we think it's really difficult sometimes. Be nice and don't get angry. Be helpful and kind.*" Others, however, felt that teachers underestimated their abilities, sometimes avoiding English materials or choosing easier tasks, whereas pupils felt they could cope: *"But they underestimate us; we can do it. They just have to give us the chance."* These contrasting perceptions point to a delicate balance between empathy and challenge. Affective alignment—teachers' belief in pupil potential and pupils' sense of being challenged yet supported—appeared central to sustaining motivation and agency. This pattern aligns with principles of inclusive pedagogy as described by Florian and Black-Hawkins (2011), who emphasize accessibility, participation, and responsiveness to learner diversity. These principles help to explain why relational support and sensitive task design emerged as critical conditions for feasible CLIL in the *vmbo* context**,** and why teachers sometimes struggled to balance accessibility with linguistic challenge.

Teachers, for their part, described this same tension from the opposite perspective, balancing comprehensibility with content mastery and sometimes doubting pupils' ability to follow extended English input—an understandable caution, but one that some pupils perceived as reducing the very engagement CLIL seeks to cultivate. This dynamic reflects the broader paradox noted earlier: the learners who stand to gain most from CLIL's motivational and linguistic potential—relative to their starting point—are also those for whom it poses the steepest cognitive and linguistic demands. While CLIL's success relies on emotional investment and trust, that investment may falter when teachers, in protecting pupils from difficulty or prioritizing subject comprehension, inadvertently constrain linguistic growth and agency.

This tension may be especially consequential in the multilingual *vmbo* environment, where roughly one-third of pupils speak a home language other than Dutch and may have less exposure to English outside school. The *vmbo* data reinforce Dewaele and Ma-

cIntyre's (2014) observation that enjoyment and anxiety are interdependent: pupils who feel encouraged, respected, and heard experience greater enjoyment and lower anxiety. For inclusive CLIL, affective alignment—teachers' belief in pupil potential and pupils' sense of being challenged yet supported—appears to be a precondition for effective learning. Cultivating such classrooms requires reflection on discourse, scaffolding, and feedback practices to ensure that efforts to safeguard subject understanding do not limit teachers' target language use or pupils' linguistic development.

The extent to which schools achieve this balance varies widely, reflecting the diversity of *vmbo* CLIL models and professional expertise—a theme explored in the next section.

6.3D Program diversity and local interpretation

The findings across chapters reveal that early pre-vocational CLIL in the Netherlands was not a single model but a spectrum of local interpretations. Even among the six participating schools, the variation in CLIL approach, organization, teacher collaboration, and English exposure was substantial. Some programs followed the national *vmbo* TTO Standard closely, integrating English into roughly 30 % of the curriculum and embedding a clear European and International Orientation (EIO) component. Others offered fewer CLIL subjects taught partly in English, with limited progression frameworks or explicit attention to CLIL methodology. One program used a modular structure, alternating between four- to six-week CLIL blocks and regular Dutch blocks during the first two years, sometimes leading to pupil ambivalence: *"I like the weeks when everything is in Dutch, but not the CLIL weeks."* This unevenness reflects both the autonomy of Dutch schools and the experimental nature of early *vmbo* CLIL implementation.

Teachers' self-reports underscored this diversity of practice. Subject teachers, in particular, differed widely in their English use—from limited use in specific parts of lessons to near-continuous L2 instruction—depending on school language policy, individual proficiency, priorities for subject content, and perceived pupil comprehension. In some schools, English teachers and subject specialists collaborated closely, fostering an emerging bilingual culture that supported shared planning and cross-curricular reinforcement. In others, CLIL remained a tentative initiative lacking whole-school commitment: *"Many pupils aren't even able to speak Dutch properly, so why are we teaching in English?"* These contrasts appeared to shape pupils' experiences, as reflected in their accounts in Chapter 5. In the most cohesive programs, pupils described CLIL as motivating, linguistically consistent, international, and community-building. In less structured settings, they reported confusion about goals, uneven challenge levels, and frustration with inconsistent English use.

Such program diversity presents both opportunity and risk. It enables schools to tailor CLIL to their pupil populations, curricular priorities, and staffing realities, but also com-

plicates efforts to monitor quality and ensure continuity across *vmbo* levels. The data suggest that local ownership, while fostering innovation, may inadvertently perpetuate inequality in bilingual provision, with some pupils enjoying rich, sustained exposure and others receiving only sporadic English use. These findings highlight the challenge of balancing contextual flexibility with clearer professional benchmarks for language use, including teachers' proficiency, extent of target language use, and linguistic adaptability, as well as for methodological integration and intercultural learning. Balancing local autonomy with clearer professional benchmarks remains the core challenge for sustainable, inclusive CLIL, a challenge that echoes in international efforts to reconcile flexibility with quality—the focus of the following section.

6.4 Cross-context comparison

Viewed against the broader European CLIL landscape, the results of this dissertation both converge with and depart from established patterns in the literature. In line with earlier studies from academically selective settings (e.g., Admiraal et al., 2006; Dalton-Puffer, 2007; Lasagabaster & Sierra, 2009), the *vmbo* pupils in the present research showed measurable linguistic gains and generally positive attitudes toward learning English. This reinforces the claim articulated by Coyle (2006a, 2013) and Pérez Cañado (2016) that CLIL can foster motivation, engagement, and self-confidence even when academic outcomes vary. Yet the pre-vocational context introduces complexities rarely observed in higher-track CLIL. The heterogeneity of language proficiency, literacy levels, home language backgrounds, and classroom engagement in *vmbo* settings revealed that while CLIL's *affective promise* remains strong, its *pedagogical calibration* requires far greater contextual sensitivity.

In contrast to findings from high-exposure programs in Spain, Austria, and Finland (Dalton-Puffer, 2007; Hüttner et al., 2013; Seikkula-Leino, 2007), *vmbo* CLIL operates under markedly reduced English-input conditions and with teachers whose language proficiency and methodological preparation vary widely. These structural constraints mirror those noted by Tammenga-Helmantel and Mossing Holsteijn (2016), who reported that trainee teachers in the Netherlands often limit their L2 use in *vmbo* classes out of concern for pupil comprehension. The present results extend that observation to *vmbo* CLIL teachers: despite positive attitudes toward CLIL, many self-reported using English far less than intended, particularly in content subjects. This highlights a recurrent cross-context tension between the *principle* of target language immersion and the *practice* of linguistic accommodation to perceived learner needs.

A further point of divergence from mainstream CLIL research lies in the persistence of strong affective responses among *vmbo* pupils. Studies in more academic contexts (e.g.,

Doiz et al., 2014b; Lasagabaster & Sierra, 2009) have reported a leveling of motivation as novelty fades, whereas in the *vmbo* data, enthusiasm and enjoyment remained salient even after two years. This pattern resonates with Dewaele & MacIntyre's (2014) work on Foreign Language Enjoyment, suggesting that for less academically advantaged learners, emotional engagement may serve as a primary driver of persistence. It also supports Coyle's (2013, p. 244) call to "listen to learners", as the qualitative findings in Chapter 5 revealed that pupils' perceptions of teacher empathy and encouragement were central to sustaining motivation and self-belief.

When compared with international efforts to diversify CLIL (e.g., Bauer-Marschallinger, 2023; Tedick, 2022), the Dutch *vmbo* model stands out as both pragmatic and experimental. Its flexible implementation aligns with Pérez Cañado's (2020, 2023) advocacy for inclusive CLIL responsive to context, yet the *vmbo* experience also illustrates the fragility of such inclusivity when systemic support and teacher training remain uneven. While studies in Norway and Germany (Bakken & Brevik, 2023; Rumlich, 2014) similarly report tensions between accessibility and rigor, the present research underscores how these tensions become magnified in streamed systems where linguistic and academic stratification are institutionalized.

Overall, the *vmbo* findings confirm that CLIL's benefits are not confined to elite learners, but they also challenge the assumption of automatic transferability across contexts. Effective and sustainable CLIL in pre-vocational education depends on contextual adaptation, professional capacity-building, and recognition of pupil voice as integral components of quality assurance. In this sense, the Dutch *vmbo* experience contributes uniquely to the international evidence base: it provides a rare longitudinal and multi-perspective portrait of inclusive CLIL in its formative phase and serves as a reminder that bilingual education's success hinges as much on empathy, adaptability, and shared purpose as on exposure hours or test scores.

The convergence of international evidence and the present *vmbo* findings underscores a central insight: the success of inclusive CLIL cannot be measured solely by linguistic outcomes, but by the system's capacity to make bilingual education sustainable, equitable, and meaningful for all learners**.** While *vmbo* CLIL demonstrates that motivation, enjoyment, and language development are attainable even among pupils of lower scholastic aptitude, it also exposes the fragility of these gains in the absence of structural continuity, teacher preparation, and institutional vision. These patterns highlight an urgent need for pedagogical calibration, ensuring that the principles of CLIL are not diluted in pursuit of accessibility, yet not applied so rigidly as to exclude those with weaker language foundations. The next section turns to these implications, outlining how classroom practice, school organization, and educational policy might better align to support the linguistic, affective, and professional dimensions of CLIL in pre-vocational education.

6.5 Implications for practice and policy

These implications draw on the empirical findings reported in Chapters 2–5, on national TTO policy frameworks, and on insights gained through *vmbo* TTO visitation and accreditation procedures, supplemented by contextual insights from more than fifty research visits undertaken during the broader data-collection period.

6.5A Classroom practice

Findings from Chapters 3–5 suggest that classroom practice plays an important role in sustaining both linguistic progress and learner engagement. Pupil perspectives indicate that lessons perceived as varied, meaningful, and enjoyable fostered confidence and persistence, particularly among learners with low linguistic self-confidence. In contrast, pupils who described lessons as *"boring"* or overly teacher-dominated often associated CLIL with limited interaction and slower progress. The balance between comprehensibility and challenge therefore remains crucial in *vmbo* CLIL.

Results from Chapters 4 and 5 also suggest that the amount and quality of target language use may influence pupils' language development and classroom experiences. Pupils who described active, communicative lessons highlighted the benefits of frequent English use: *"We speak English a lot … we also speak a lot of English outside of the lessons … The more you speak English, the easier it gets."* Others contrasted this with more passive classes: *"We are not more active [in the CLIL lessons] or anything like that. It's all the same—just sit and listen."* These differing experiences illustrate how classroom interaction appears to shape confidence and opportunities for language use, whereas insufficient interaction may restrict them. Some pupils also observed that explicit language feedback was rare and that attention to form within subject lessons was minimal, which may have limited opportunities for noticing and reformulation—processes regarded as central to interlanguage development (Lyster, 2007; Swain, 1985). As one pupil explained, *"Sometimes you hear that you're doing something wrong, but you don't know what and they don't tell you, and you think, 'Everything I do is wrong.'"*

A related tension concerned the balance between target language immersion and the use of Dutch for clarification. Many pupils in the lower *vmbo* focus groups argued that teachers should occasionally use Dutch when English impeded comprehension, whereas pupils in the higher-level programs—particularly in School A—advocated consistent English use: *"All CLIL teachers require us to speak English with each other. And if someone doesn't, then we [the pupils] help each other to keep speaking English."* These contrasting preferences underline that perceptions of effective language use depend on pupils' proficiency levels and program intensity. They also point to the need for teacher flexibility in shifting between target language authenticity and pupil accessibility.

Grounded in CLIL pedagogy and consistent with pupil reflections in Chapter 5, effective scaffolding combines clarity of input, structured support, judicious translanguaging, and gradual autonomy within communicative tasks. Both teachers and pupils valued CLIL most when lessons in English encouraged interaction, feedback, and enjoyable activities—making *'fun'* not a superficial goal but a motivational catalyst that supports sustained effort.

6.5B Teacher expertise and professional support

At the school level, leadership, coherent program organization, and collegial coordination appear decisive for CLIL quality and inclusiveness. Chapter 5 revealed considerable variation among programs in the balance between English and Dutch, the implementation of international activities, and the degree of collaboration between English and subject teachers. Schools where collaboration was embedded in the program through shared planning, cross-curricular projects, or well-organized and inclusive trips and exchanges tended to report higher satisfaction among both teachers and pupils.

A related question concerns the language proficiency requirements for CLIL teachers. The *vmbo* TTO Standard stipulates that schools include at least one native-speaker English teacher, yet the findings suggest that pupils valued linguistic clarity and bilingual flexibility as much as, if not more than, native-like command. In one school, several pupils expressed frustration when a native-speaker teacher could not provide explanations in Dutch and felt they had learned less subject content as a result. Conversely, teachers who combined strong English proficiency with sufficient Dutch to clarify complex points were perceived as more effective and approachable. These experiences highlight a practical dilemma for schools: while recruiting native-speaker teachers can enrich language input and intercultural awareness, ensuring that all teachers possess adequate Dutch proficiency seems equally important for accessibility and inclusion. At the policy level, this tension underscores the need for clearer national guidance on how bilingual staffing standards can be balanced with local realities of teacher supply and pupil needs.

Ongoing professional development that extends beyond language proficiency remains essential, particularly for subject specialists who may have entered CLIL without prior methodological training. Such training should not only address CLIL principles and scaffolding strategies, but also practical techniques for integrating light corrective feedback and adopting a counterbalanced approach that combines meaning-focused instruction with form-focused attention (Lyster, 2007). Such preparation could take place through team participation in CLIL methodology courses (Pérez Cañado, 2016), whether domestically or abroad, or through sustained mentoring and reflective dialogue within the school. Investment in professional learning communities can strengthen program coherence and help prevent the fragmentation that characterized some early-stage *vmbo* CLIL initiatives.

6.5C Policy and systemic support

At policy level, the unevenness observed across *vmbo* programs reflects a broader structural gap: the absence of national frameworks for teacher preparation, assessment, and *vmbo*-specific CLIL pedagogy. Without these, implementation remains dependent on local expertise, enthusiasm, and resources—conditions that favor well-supported schools but risk inequity elsewhere. National policy could encourage continuity by offering more accreditation routes for CLIL teacher training and by providing model lesson frameworks that demonstrate attainable language–content integration for mixed-ability learners.

A further systemic challenge concerns teacher recruitment and staffing standards. The *vmbo* TTO Standard's expectation that schools include at least one native-speaker English teacher may be difficult to realize sustainably, particularly outside major cities, and may not always align with pupils' linguistic and accessibility needs. As noted in Section 6.5B, bilingual proficiency and the ability to provide occasional support in Dutch can be at least as important as native-level English skills. National guidance could clarify how such staffing requirements can be adapted to different school contexts without compromising program quality or inclusivity.

Equity also extends to teacher training and international opportunities. Professional development in CLIL methodology remains unevenly available and often costly for schools. Subsidizing participation in accredited CLIL courses or offering publicly funded regional training networks could ensure that financial barriers do not restrict access to high-quality preparation, particularly for smaller or less well-resourced *vmbo* institutions.

Similarly, exchange programs and study trips play an important motivational and intercultural role, yet participation opportunities vary widely between schools. Financial constraints, logistical challenges, and differing local policies on pupil travel can limit access, creating disparities that contradict CLIL's inclusive ethos. Broader structural support through partial funding schemes, digital exchanges, or regionally coordinated initiatives could help ensure that all *vmbo* CLIL pupils benefit from authentic international contact and exposure to English beyond the classroom.

While Nuffic's established system of school visits and evaluation ensures high standards for accredited programs, the trajectory toward TTO accreditation remains lengthy, complex, and costly. Schools would benefit from clearer preparatory guidance, mentoring, and additional interim evaluation stages to support quality development before formal accreditation. Such scaffolding could help newer programs progress more confidently and maintain consistency across the *vmbo* CLIL landscape.

Equally, inspection and evaluation processes should emphasize quality and consistency rather than quantity of English use. Revising inspection criteria to ensure adherence to

the national standard regarding target language use and communicative focus would help maintain clarity while allowing flexibility for local adaptation.

A complementary policy consideration concerns the target language-attainment levels prescribed in the *vmbo* TTO Standard. Sustaining *vmbo* CLIL within differentiated attainment frameworks is essential to both its equity and its long-term viability. Success at this level should continue to be defined by feasible linguistic progression, consistent with each subtrack's curricular standards. However, in line with the findings from the qualitative chapters and the recommendations under 6.5A and 6.5B regarding intensified English exposure, a related recommendation is that these prescribed attainment levels be cautiously raised, particularly for the higher *vmbo* tracks and in the domain of oral proficiency. A pilot program implementing such measures could explore how increased target language use in the later *vmbo* years might elevate outcomes while safeguarding the inclusive and differentiated nature of *vmbo* CLIL.

Finally, sustained funding and increased visibility for *vmbo* CLIL within national educational discourse would affirm its legitimacy alongside the more established bilingual tracks, signaling that inclusivity and excellence are not mutually exclusive aims.

6.5D Summary of key recommendations, grouped thematically

In addition to the synthesized recommendations presented in this section, Attachment 6.1 (Appendix 6) presents a consolidated overview of teacher and pupil recommendations derived directly from the qualitative findings reported in Chapter 5. These practitioner and pupil perspectives are included to complement and contextualize the broader set of implications discussed here in Chapter 6.

Classroom and pedagogical practice

- **Maintain comprehensible challenge.** Ensure that English input in CLIL lessons remains accessible yet sufficiently demanding; avoid excessive simplification that limits exposure (see Chapter 5; Coyle et al., 2010).
- **Foster meaningful, enjoyable communication.** Integrate interactive, purposeful activities that promote both linguistic growth and pupil motivation (see Chapter 5; Dale & Tanner, 2012).
- **Encourage multilingual empathy.** Use translanguaging strategically to support understanding while affirming pupils' linguistic identities (see Chapter 5; García & Wei, 2014; Nikula, 2017b).
- **Strengthen cognitive engagement and global relevance.** Connect English learning more explicitly to international orientation (EIO) projects and pupils' vocational pathways, highlighting the personal and practical usefulness of English to sustain motivation and purpose (see Chapters 2 and 5; Goris et al., 2019; Pinner, 2021).

Pupil support and inclusion

- **Introduce an English immersion period before CLIL begins in year 1** to help pupils and teachers acclimate to hearing and using English and ease the transition into bilingual subject lessons (see Chapter 5; Rumlich, 2016).
- **Take steps to listen to 'learner voice'.** Systematically incorporate pupils' perspectives through questionnaires, focus groups, or reflective activities to inform CLIL program design and implementation, particularly regarding affective feasibility, perceived challenge, and inclusion (see Chapter 5; Coyle, 2013).
- **Appoint teachers strategically.** Consider carefully the specific pupil population's linguistic and educational needs when appointing a non-Dutch-speaking CLIL teacher, particularly in the lower *vmbo* levels (see Chapter 5).

School-level organization and professional development

- **Institutionalize collaboration.** Embed joint planning, cross-curricular projects, and reflective teacher communities in program design (see Chapter 5; Dale & Mearns, 2023).
- **Prioritize collective professional development.** Facilitate group participation in CLIL methodology courses and create in-school mentoring structures (see Chapter 5; Pérez Cañado, 2016).
- **Subsidize CLIL training and resources.** Provide financial support for teacher participation in accredited CLIL methodology programs and regional professional networks to ensure equal access across schools (see Chapter 5; Nuffic, n.d.-b).

Policy and system-level development

- **Establish *vmbo*-specific CLIL frameworks.** Develop national guidance for teacher training, assessment, and pedagogical adaptation tailored to pre-vocational contexts (see Chapter 5 and 6; Nuffic, n.d.-b).
- **Strengthen guidance and quality assurance.** Offer clearer preparatory guidance, mentoring, and additional interim evaluation during the Nuffic TTO accreditation process to help schools progress confidently toward accreditation and maintain consistent quality across programs (see Chapters 5 and 6; Nuffic, n.d.-b).
- **Recalibrate *vmbo* CLIL language-attainment levels.** Pilot modest upward adjustments, particularly for the highest *vmbo* track *(theoretische leerweg*) and oral proficiency, where increased exposure and improved teacher English use can support higher outcomes without compromising inclusivity (see Chapters 4 and 5).
- **Promote equitable international participation.** Support school exchanges and intercultural projects through public funding or regional coordination so that *all* pupils can benefit from authentic international exposure (see Chapter 5; Nuffic, n.d.-a).
- **Reinforce policy recognition and funding.** Sustain *vmbo* CLIL's visibility and long-term viability through explicit inclusion in national bilingual education policy and funding mechanisms (see Chapters 5 and 6; Nuffic, n.d.-a).

Together, these recommendations support the continued development of *vmbo* CLIL from early experimentation toward a more consolidated and sustainable phase—a development further traced in the following section.

6.6: Current situation and future outlook

The following sections situate the study's *vmbo* CLIL findings within subsequent developments in Dutch TTO policy and practice. They draw on Nuffic documentation, public information about *vmbo* TTO programs, and experience gained in TTO visitation procedures, rather than on the original data set itself.

6.6A The six participating CLIL schools: continuity and transformation

In the decade following data collection (2012–2014), the six schools that formed the basis of this research have taken markedly different paths. Three have since discontinued their CLIL programs, either entirely or as separate streams, and two are in the process of phasing it out. One (School A) achieved official TTO certification in 2021 and continues to offer a thriving *vmbo* CLIL stream. The discontinuation of the other programs stemmed from diverse, context-specific factors: financial pressures, declining pupil enrollment in the CLIL stream, a merger with a larger non-CLIL institution, or a strategic decision to adopt a different curricular profile.

These changes do not necessarily represent failure. Rather, they reflect the fragile and experimental status of early *vmbo* CLIL during the period in which this research was conducted. At that time, schools operated with minimal external guidance, a *vmbo*-specific framework geared towards future certification, and inconsistent access to teacher professionalization or shared best practices. The subsequent attrition from the research group should therefore be read as part of a broader process of institutional learning and policy maturation. The pioneering role of these schools—particularly School A, which benefitted from proximity to established bilingual *havo* and *vwo* tracks—helped lay the groundwork for more sustainable and better-supported implementations. Between 2012 and 2024, the number of CLIL programs at *vmbo* level more than doubled, from 15 to 34. This most recent number has since decreased slightly due to a combination of factors, but the rapid growth of pre-vocational CLIL in the Netherlands remains noteworthy.

Today, the legacy of those early pilots endures not in the continuity of programs but in the evolution of practice. The knowledge gained through their challenges and partial successes contributed to a clearer understanding of what *vmbo* CLIL requires to thrive: stable funding, administrative support, trained bilingual teachers, and integration into a coherent school-wide vision of internationalization and inclusion.

6.6B National developments (2012–2025)

Following the experimental phase described in this dissertation, Dutch *vmbo* CLIL gradually entered a period of consolidation. The early 'TTO Standard' (Europees Platform, 2009) provided only broad quantitative targets—most notably the requirement that at least 30 percent of total instructional time be delivered in English—but left methodology, assessment, and professional support largely to individual schools. This flexibility allowed innovation but also produced inconsistency, as seen in the six programs investigated here.

In 2018 the revised TTO 2.0 Standard redefined bilingual education at pre-vocational level as a qualitatively specified practice rather than a numerical target. The new framework raised the *aspirational* curricular goal to 40 percent, permitted compensation across year groups, and, crucially, formalized expectations for target language use, CLIL pedagogy, and teacher competence. English was now explicitly required as the language of instruction (*'de voertaal'),* and CLIL methodology was operationalized through four clear didactic principles: active learning, disciplinary language awareness, communicative interaction, and feedback on pupil language production. These refinements marked the transition from early improvisation to a shared professional vocabulary of practice.

The 2018 update also established systematic teacher development requirements. Schools were expected to maintain CLIL expertise collectively through regular training, peer observation, and coaching, supported by collaboration between subject and English teachers. Language-proficiency expectations were aligned with the CEFR (typically B2 minimum for teachers), and quality control moved from school self-description to documented evidence of practice. In addition, the framework introduced explicit language-outcome benchmarks for pupils and made participation in at least one international collaboration or exchange project a universal requirement, thus responding to earlier concerns about unequal access to international experiences.

Overall, the 2018 reform signaled the professionalization and normalization of *vmbo* CLIL. What had begun as an isolated, experimental innovation evolved into a system with clear pedagogical parameters, defined teacher standards, and structural accountability. These developments laid the groundwork for renewed growth in *vmbo* bilingual programs after 2020 and provide the policy context for interpreting the forthcoming *Grow your world: Ontwikkelagenda TTO 2025–2030* [Development Agenda for Bilingual Education 2025-2030] (Nuffic, 2025).

The TTO 2.0 framework thus evolved into a broader educational movement. Whereas early *vmbo* CLIL programs focused primarily on language exposure and methodological experimentation, the *Ontwikkelagenda* situates bilingual education within a vision of global citizenship and lifelong learning. English has become a vehicle for international

cooperation, empathy, and agency, echoing the inclusive ideals that first motivated *vmbo* CLIL.

6.6C Future outlook: TTO 2025–2030

As the bilingual education movement matures, the *Ontwikkelagenda TTO 2025–2030* (Nuffic, 2025) marks a new phase in its evolution. The agenda explicitly redefines the mission of bilingual education in the Netherlands: English is no longer the primary objective but a means for developing global citizenship, inclusion, and sustainability. This reframing reflects the ongoing normalization of English competence in Dutch society and the recognition that meaningful bilingualism now entails intercultural engagement and awareness as much as linguistic skill.

The accompanying *Global Citizen Competencies* (Nuffic, n.d.-a) framework translates this vision into nine developmental dimensions—curiosity, empathy, critical thinking, communication, collaboration, cultural awareness, connectedness, linguistic competence, and lifelong learning. These traits signal a deliberate move away from a narrow conception of language proficiency toward a holistic profile of the plurilingual, reflective learner. For *vmbo* contexts, this emphasis carries particular relevance: it validates practical, interpersonal, and experiential forms of learning that have long been strengths of pre-vocational education.

Within this vision, inclusive CLIL is positioned as both an equity measure and a catalyst for personal development. The agenda's call for *TTO for all learners* implicitly extends legitimacy to *vmbo* CLIL and recognizes its potential to nurture international orientation among pupils who might not otherwise access such experiences. The participatory and practically-oriented nature of *vmbo* learning aligns naturally with the competencies outlined in the framework, suggesting that the track once viewed as least compatible with bilingual education may, in fact, embody many of its emerging priorities.

The professionalization of bilingual teaching is also expected to deepen in this period. Building on the first endowed chair in *World Citizenship and Bilingual Education*, established in 2020 and held since 2024 by its second professor, the agenda emphasizes the growing partnership between research and practice. Strengthening collaboration among schools, universities, and teacher-education institutes is seen as essential for creating a sustainable and evidence-informed bilingual education ecosystem. Including the teacher-education programs that award the Dutch second-degree (*tweedegraads*) qualification is particularly critical, given that these programs prepare the majority of *vmbo* teachers and integrate *vmbo*-specific pedagogical approaches and diversity-focused training into their curriculum. For *vmbo* CLIL, such collaboration could provide precisely the continuity and mentoring that early initiatives lacked, and add an extra impetus to the stalled growth in *vmbo* CLIL programs.

In sum, the *Ontwikkelagenda 2025–2030* envisions a future in which bilingual education functions less as a linguistic enrichment program and more as a comprehensive educational philosophy. For the *vmbo* sector, this trajectory affirms that CLIL can—and should—serve as a bridge between language learning, vocational preparation, and global citizenship, ensuring that internationalization becomes not a privilege for the few but a shared educational experience for all.

While the *Ontwikkelagenda 2025–2030* charts an ambitious and inclusive path forward, the findings of this dissertation underscore how fragile the foundations of *vmbo* CLIL can be without sustained structural, financial, and pedagogical support. The early programs captured here function as a kind of baseline reference point, revealing both the promise and the precarity of bilingual innovation in the pre-vocational sector. As such, they invite longitudinal follow-up: research that traces not only linguistic growth, but also teacher professionalization, curriculum continuity, and institutional resilience in the years since. The following section therefore turns from current practice to the limitations of this study and the directions for future research that emerge from it.

6.7 Limitations and future research

To contextualize the scope and reliability of the findings, this section discusses methodological and contextual limitations and indicates directions for further research. The points are organized from design and measurement through analysis and interpretation, highlighting where future studies could extend or refine the present work.

Design and sampling

The participating schools were early adopters of *vmbo* CLIL and therefore atypical in their degree of innovation and institutional motivation. Because the sample was purposive rather than random, the findings may not fully represent *vmbo* schools with different resources, leadership, or demographics. Broader, multi-site longitudinal research incorporating both CLIL and non-CLIL schools would make it possible to examine how contextual factors such as leadership, resource allocation, or regional diversity influence program outcomes.

Measurement and data collection

Although classroom observations were originally planned for all fifteen CLIL classes, scheduling problems and some teachers' concerns about researcher presence reduced the total to only three lessons at two schools. Moreover, pupils in one observed class indicated that the teacher had used more English than usual, suggesting an observer effect. Future research should therefore include systematic yet unobtrusive observation

protocols, such as video recordings, peer observations, or structured CLIL checklists, to further triangulate with teacher self-reports and pupil perceptions.

Another limitation concerns rater reliability in the writing assessment. The use of student teachers as raters may have introduced inconsistencies, as inter-rater agreement proved weaker than anticipated. Employing better-trained and calibrated raters, or supplementing human ratings with objective linguistic indices such as lexical diversity or syntactic complexity, would strengthen future analyses.

A further contextual limitation concerns test motivation and familiarity. During data collection, many classes asked whether the language tests would count toward their grades. When told they were for research only, pupils expressed relief but may not have applied themselves fully (which may partly reflect the Dutch secondary school *zesjes-cultuur*, in which effort is often calibrated to securing a minimum passing grade rather than exceeding it). In later rounds, some also showed markedly less enthusiasm for repeating the same test battery, which may have slightly lowered overall performance. These effects likely affected both CLIL and non-CLIL groups equally but could have introduced a small downward bias in scores. Subsequent studies could address this by offering low-stakes incentives or using alternate test versions to maintain engagement across measurement points.

A fourth measurement gap involved the attempted Dutch-proficiency test, which had to be discarded due to irregular administration and suspect response patterns. Without a valid L1 measure, this study could not assess whether bilingual instruction affected maintenance of Dutch. Subsequent research should therefore implement strictly controlled testing conditions and include both L1 and content-subject measures to capture the full linguistic balance intended in bilingual education.

Analytical considerations

Because the study relied on intact classes and non-random assignment, the design is observational, and strictly causal claims about the effects of CLIL cannot be made. Statistical control for covariates mitigated but did not eliminate unobserved influences related to teachers, classrooms, and schools. In addition, the number of participating pupils per class and per school, together with the absence of some combinations of school type and CLIL/non-CLIL provision, precluded reliable class- or school-level comparisons and the use of these levels as moderating variables, limiting the generalizability of between-school differences. Cohort-based or introduction-year designs, tracking schools before and after CLIL implementation, would offer stronger grounds for identifying program effects.

Much CLIL research (e.g., Admiraal et al., 2006; Lasagabaster, 2008; Mearns, 2015) relies solely on ANCOVA. In Chapter 4, however, small but meaningful discrepancies emerged between ANCOVA and change-score analyses when evaluating gains in speaking proficiency and fluency. While van Breukelen (2013) argues that change-score methods may be less biased than ANCOVA in non-randomized settings with baseline differences, other methodological work (Tennant et al., 2022) cautions that change scores in observational data may themselves yield unstable or biased estimates when group assignment is not controlled. Taken together, these findings suggest that neither method is uniformly superior in observational designs; reporting both would enhance transparency, allow readers to assess the robustness of results, and facilitate comparability across CLIL contexts.

Finally, the longitudinal component covered only two school years. Longer tracking through all four *vmbo* years, and ideally into post-secondary *mbo* programs, would clarify whether linguistic and attitudinal gains persist when English exposure diminishes.

Context and program maturity

The research captured a formative stage of *vmbo* CLIL (2012–2014), when teacher language proficiency, CLIL materials, and program structures were still emerging. The results therefore portray feasibility rather than consolidation. Replicating similar analyses under the revised *TTO Standard* (Nuffic, n.d.-b) and within the framework of the *Ontwikkelagenda 2025–2030* would allow researchers to document how program maturity, professionalization, and institutional stability affect outcomes.

Although the participating schools initially agreed to share final-examination results for longitudinal comparison, only two ultimately provided them, preventing analysis of pupil outcomes against national norms. Future collaborations should formalize such data-sharing agreements to enable stronger benchmarking.

Theoretical scope

The dissertation focused primarily on linguistic and affective dimensions, leaving cognitive and intercultural learning outcomes outside its scope. Similarly, the research did not examine comparative gains in content subject knowledge or Dutch language proficiency between CLIL and non-CLIL pupils, thus excluding two of the pillars of Dutch CLIL evaluation. Future studies could integrate these missing elements by combining classroom discourse analysis, cognitive task data, and learner-reflection instruments. Such designs would extend the evidence base for the 4Cs model (Coyle, 2006b) and illuminate how CLIL promotes cognitive engagement and identity development among less academically advantaged learners.

6.7A Concluding reflection

The limitations outlined here delineate the boundaries of this pioneering phase of *vmbo* CLIL research. Despite its constraints, the study provides an essential 'zero measurement' of inclusive bilingual education in the Dutch pre-vocational context. Addressing these methodological and contextual gaps through triangulated classroom data, extended longitudinal designs, calibrated assessment, and broader theoretical integration will enable future research to move from proof of concept toward sustainable, evidence-based practice. In doing so, researchers can determine not merely *whether vmbo* CLIL works, but *how, for whom, and under what conditions it* achieves its most equitable and lasting results.

At the same time, the four studies together offer a coherent picture of what makes CLIL viable in this context. They show that CLIL in the pre-vocational stream can achieve tangible linguistic gains without academic selection, provided its pedagogical foundations remain strong. Growth in English proficiency across most skills affirms *vmbo* pupils' capacity to succeed when learning through another language, while the central role of *cognition*—the perceived usefulness of English—highlights the attitudinal power of relevance and purpose. Equally, progress in *vmbo* CLIL depends not only on linguistic or cognitive factors, but on the relationship between teachers and pupils that sustains engagement in challenging conditions, supported by a school culture that provides a stable backbone. Yet reluctance among some subject teachers to teach extensively in English, together with uneven attention to language, variety, and communicative practice, indicates that inclusive CLIL still requires clearer guidance and professional support. Even so, the *vmbo* experience demonstrates that when pedagogical integrity and mutual confidence are sustained, bilingual education can extend its benefits beyond the academic elite and contribute meaningfully to a more inclusive educational landscape.

References

Ackerl, C. (2007). Lexico-grammar in the essays of CLIL and non-CLIL students: Error analysis of written production. *Vienna English Working Papers*, *16*(3), 6–11.

ADIBE Project. (n.d.). *Attention to Diversity in Bilingual Education*. https://adibeproject.com

Admiraal, W., Westhoff, G., & de Bot, K. (2006). Evaluation of bilingual secondary education in the Netherlands: Students' language proficiency in English. *Educational Research and Evaluation*, *12*(1), 75–93. https://doi.org/10.1080/13803610500392160

Ajzen, I. (n.d.). *Constructing a Theory of Planned Behavior Questionnaire*. http://people.umass.edu/aizen/pdf/tpb.measurement.pdf

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology*, *32*(4), 665–683. https://doi.org/10.1111/j.1559-1816.2002.tb00236.x

Ajzen, I. (2005). *Attitudes, personality, and behavior* (2nd ed.). Open University Press / McGraw-Hill.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Prentice-Hall.

Albers, M. J. (2017). *Introduction to quantitative data analysis in the behavioral and social sciences*. John Wiley & Sons.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93. https://doi.org/10.2307/271083

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

Apsel, C. (2012). Coping with CLIL: Dropouts from CLIL streams in Germany. *International CLIL Research Journal*, *1*(4), 47–56.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology*, *40*(4), 471–499. https://doi.org/10.1348/014466601164939

Arribas, M. (2016). Analysing a whole CLIL school: Students' attitudes, motivation, and receptive vocabulary outcomes. *Latin American Journal of Content & Language Integrated Learning*, *9*(2), 267–292. https://doi.org/10.5294/laclil.2016.9.2.2

ATLAS.ti Scientific Software Development GmbH. (2022). *ATLAS.ti* [Computer software]. https://atlasti.com

Bagozzi, R. P., & Burnkrant, R. E. (1979). Attitude organization and the attitude–behavior relationship. *Journal of Personality and Social Psychology*, *37*(6), 913–929. https://doi.org/10.1037/0022-3514.37.6.913

Bagozzi, R. P., & Burnkrant, R. E. (1980). Single component versus multicomponent models of attitude: Some cautions and contingencies for their use. In J. C. Olson (Ed.), *Advances in Consumer Research* (Vol. 7, pp. 339–344).

Bagozzi, R. P., & Burnkrant, R. E. (1985). Attitude organization and the attitude-behavior relation: A reply to Dillon and Kumar. *Journal of Personality and Social Psychology*, *49*(1), 47–57. https://doi.org/10.1037/0022-3514.49.1.47

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*(1), 74–94. https://doi.org/10.1007/BF02723327

Baker, S. C., & MacIntyre, P. D. (2000). The role of gender and immersion in communication and second language orientations. *Language Learning*, *50*(2), 311–341. https://doi.org/10.1111/0023-8333.00119

Bakken, J., & Brevik, L. M. (2023). Challenging the notion of CLIL elitism: A study of secondary school students' motivation for choosing CLIL in Norway. *TESOL Quarterly*, *57*(4), 1091–1114. https://doi.org/10.1002/tesq.3173

Ball, P., Clegg, J., & Kelly, K. (2015). *Putting CLIL into practice*. Oxford University Press.

Banegas, D. L. (2012). Integrating content and language in English language teaching in secondary education: Models, benefits, and challenges. *Studies in Second Language Learning and Teaching*, *2*(1), 111–136. https://doi.org/10.14746/ssllt.2012.2.1.6

Banegas, D. L., & Cerrato Will, G. (2022). "Challenging, but not impossible": Student-teachers' views of CLIL. *Argentinian Journal of Applied Linguistics - ISSN 2314-3576*, *10*(2), 51–60.

Barrios, E. (2022). The effect of parental education level on perceptions about CLIL: A study in Andalusia. *International Journal of Bilingual Education and Bilingualism*, *25*(1), 183–195. https://doi.org/10.1080/13670050.2019.1646702

Barrios, E., & Acosta-Manzano, I. (2022). Primary students' satisfaction with CLIL and perceived CLIL linguistic difficulty. *Journal of Multilingual and Multicultural Development*, *43*(7), 665–678. https://doi.org/10.1080/01434632.2020.1759610

Barrios, E., & Milla Lara, M. D. (2020). CLIL methodology, materials and resources, and assessment in a monolingual context: An analysis of stakeholders' perceptions in Andalusia. *The Language Learning Journal*, *48*(1), 60–80. https://doi.org/10.1080/09571736.2018.1544269

Bauer-Marschallinger, S., Dalton-Puffer, C., Heaney, H., Katzinger, L., & Smit, U. (2023). CLIL for all? An exploratory study of reported pedagogical practices in Austrian secondary schools. *International Journal of Bilingual Education and Bilingualism*, *26*(9), 1050–1065. https://doi.org/10.1080/13670050.2021.1996533

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin*, *112*(3), 400–404. https://doi.org/10.1037/0033-2909.112.3.400

Berthoud, R. (2000). A measure of changing health. In R. Berthoud & J. Gershuny (Eds.), *Seven years in the lives of British families: Evidence on the dynamics of social change from the British household panel survey* (pp. 161–192). Policy Press.

Bonnet, A. (2012). Towards an evidence base for CLIL: How to integrate qualitative and quantitative as well as

process, product and participant perspectives in CLIL research. *International CLIL Research Journal, 1*(4), 66–78.

Bonnet, A., & Dalton-Puffer, C. (2013). Great expectations? Competence- and standard-related questions concerning CLIL moving into the mainstream. In D. Marsh & O. Meyer (Eds.), *Content and Language Integrated Learning (CLIL) in Europe: Practice, policy and perspectives* (pp. 269–284). Peter Lang.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3–15. https://doi.org/10.3102/0013189X033008003

Botes, E., Dewaele, J.-M., & Greiff, S. (2022). Taking stock: A meta-analysis of the effects of foreign language enjoyment. *Studies in Second Language Learning and Teaching*, *12*(2), 205–232. https://doi.org/10.14746/ssllt.2022.12.2.3

Boudreau, C., MacIntyre, P. D., & Dewaele, J.-M. (2018). Enjoyment and anxiety in second language communication: An idiodynamic approach. *Studies in Second Language Learning and Teaching*, *8*(1), 149–170. https://doi.org/10.14746/ssllt.2018.8.1.7

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. Sage.

Breckler, S. J., & Wiggins, E. C. (1989). Affect versus evaluation in the structure of attitudes. *Journal of Experimental Social Psychology*, *25*(3), 253–271. https://doi.org/10.1016/0022-1031(89)90022-X

Bredenbröker, W. (2002). Förderung der fremdsprachlichen Kompetenz durch bilingualen Sachfachunterricht: Empirische Untersuchungen. In S. Breidbach, G. . Bach, & D. Wolff (Eds.), *Mehrsprachigkeit in Schule und Unterricht: Bilingualer Sachfachunterricht. Didaktik, Lehrer-/Lernerforschung und Bildungspolitik zwischen Theorie und Empirie* (pp. 141–150). Peter Lang.

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

Brevik, L. M., & Moe, E. (2012). Effects of CLIL teaching on language outcomes. In D. Tsagari & I. Csépes (Eds.), *Collaboration in language testing and assessment* (Vol. 26, pp. 213–227). Peter Lang.

Brinberg, D. (1981). A comparison of two behavioral intention models. In K. B. Monroe (Ed.), *Advances in consumer research* (Vol. 8, pp. 48–52). Association for Consumer Research.

Broca, Á. (2016). CLIL and non-CLIL: Differences from the outset. *ELT Journal*, *70*(3), 320–331. https://doi.org/10.1093/elt/ccw011

Broekhof, K., Vaessen, K., Maarse, J., Aarssen, J., & van Velzen, T. (2019). *Cijfers en meningen over lezen in Nederland*. Sardes.

Brown, H., & Bradford, A. (2017). EMI, CLIL, & CBI: Differing approaches and goals. In P. Clements, A. Krause, & H. Brown (Eds.), *Transformation in language education* (pp. 328–334). JALT.

Bruton, A. (2011a). Are the differences between CLIL and non-CLIL groups in Andalusia due to CLIL? A reply to Lorenzo, Casal and Moore (2010). *Applied Linguistics*, *32*(2), 236–241. https://doi.org/10.1093/applin/amr007

Bruton, A. (2011b). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, *39*(4), 523–532. https://doi.org/10.1016/j.system.2011.08.002

Bruton, A. (2013). CLIL: Some of the reasons why … and why not. *System*, *41*(3), 587–597. https://doi.org/10.1016/j.system.2013.07.001

Bruton, A. (2015). CLIL: Detail matters in the whole picture. More than a reply to J. Hüttner and U. Smit (2014). *System*, *53*, 119–128. https://doi.org/10.1016/j.system.2015.07.005

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford University Press.

Buchner, A., Erdfelder, E., Faul, F., & Lang, A. (2019). *G*Power* (Version 3.1.9.4) [Computer software]. Universität Kiel. Retrieved March 13, 2026 from https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.

Buckingham, L. R., Fernández Álvarez, M., & Halbach, A. (2023). Differences between CLIL and non-CLIL students: Motivation, autonomy and identity. *Journal of Multilingual and Multicultural Development*, *44*(7), 626–640. https://doi.org/10.1080/01434632.2022.2102641

Canga Alonso, A. (2015). Receptive vocabulary of CLIL and non-CLIL primary and secondary school learners. *Complutense Journal of English Studies*, *23*, 59–77. https://doi.org/10.5209/rev_CJES.2015.v23.51301

Cenoz, J. (2001). Three languages in contact: Language attitudes in the Basque Country. In D. Lasagabaster & J. M. Sierra (Eds.), *Language awareness in the foreign language classroom* (pp. 37–60). University of the Basque Country Press.

Cenoz, J. (2015). Content-based instruction and Content and Language Integrated Learning: The same or different? *Language, Culture and Curriculum*, *28*(1), 8–24. https://doi.org/10.1080/07908318.2014.1000922

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

Centraal Bureau voor de Statistiek. (2014). *Jaarrapport Integratie 2014*. Retrieved March 13, 2026, from https://www.cbs.nl/nl-nl/publicatie/2014/47/jaarrapport-integratie-2014.

Centraal Bureau voor de Statistiek. (2016). *VO; leerlingen, onderwijssoort, leerjaar, herkomstgroepering, generatie* [Dataset]. Retrieved March 13, 2026, from https://opendata.cbs.nl/#/CBS/nl/dataset/37220/table?ts=1773377556038.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Clément, R., Baker, S. C., & MacIntyre, P. D. (2003). Willingness to communicate in a second language: The effects of context, norms, and vitality. *Journal of Language and Social Psychology*, *22*(2), 190–209. https://doi.org/10.1177/0261927X03022002003

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly*, *23*(3), 509. https://doi.org/10.2307/3586923

Conner, M., & Armitage, C. J. (1998). Extending the theory of planned behavior: A review and avenues for further research. *Journal of Applied Social Psychology*, *28*(15), 1429–1464. https://doi.org/10.1111/j.1559-1816.1998.tb01685.x

Conner, M., & Sparks, P. (2005). The theory of planned behaviour. In M. Conner & P. Norman (Eds.), *Predicting health behaviour: Research and practice with social cognition models* (2nd ed., pp. 121–162). Open University Press.

Coonan, C. M. (2007). Insider views of the CLIL class through teacher self-observation–introspection. *International Journal of Bilingual Education and Bilingualism*, *10*(5), 625–646. https://doi.org/10.2167/beb463.0

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

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

Coyle, D. (2006b). Developing CLIL: Towards a theory of practice. In *Monograph 6*. APAC Barcelona. https://www.research.ed.ac.uk/en/publications/developing-clil-towards-a-theory-of-practice/

Coyle, D. (2007). Content and Language Integrated Learning: Towards a connected research agenda for CLIL pedagogies. *International Journal of Bilingual Education and Bilingualism*, *10*(5), 543–562. https://doi.org/10.2167/beb459.0

Coyle, D. (2013). Listening to learners: An investigation into 'successful learning' across CLIL contexts. *International Journal of Bilingual Education and Bilingualism*, *16*(3), 244–266. https://doi.org/10.1080/13670050.2013.777384

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and Language Integrated Learning*. Cambridge University Press.

Coyle, D., & Meyer, O. (2021). *Beyond CLIL: Pluriliteracies teaching for deeper learning*. Cambridge University Press.

Cummins, J. (1979a). *Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters* (Working Papers on Bilingualism (No. 19)). Ontario Institute for Studies in Education.

Cummins, J. (1979b). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251. https://doi.org/10.3102/00346543049002222

Cummins, J. (1984a). *Bilingualism and special education: Issues in assessment and pedagogy*. Multilingual Matters.

Cummins, J. (1984b). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language proficiency and academic achievement* (pp. 2–19). Multilingual Matters.

Dale, L., & Mearns, T. (2023). CLIL challenges in designing learning experiences. In D. L. Banegas & S. Zappa-Hollman, *The Routledge handbook of Content and Language Integrated Learning* (1st ed., pp. 313–327). Routledge. https://doi.org/10.4324/9781003173151-26

Dale, L., & Tanner, R. (2012). *CLIL activities: A resource for subject and language teachers*. Cambridge University Press.

Dale, L., van der Es, W., & Tanner, R. (2011). *CLIL skills*. European Platform / Expertisecentrum MVT, ICLON, Leiden University.

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. (2007). *Discourse in Content and Language Integrated Learning (CLIL) classrooms*. John Benjamins.

Dalton-Puffer, C. (2008a). Communicative competence in ELT and CLIL classrooms: Same or different? *Vienna English Working Papers*, *17*(3), 14–21.

Dalton-Puffer, C. (2008b). Outcomes and processes in Content and Language Integrated Learning (CLIL): Current research from Europe. In W. Delanoy & L. Volkmann (Eds.), *Future Perspectives for English Language Teaching* (pp. 139–157). Universitätsverlag Winter.

Dalton-Puffer, C. (2011). Content-and-Language Integrated Learning: From practice to principles? *Annual Review of Applied Linguistics*, *31*, 182–204. https://doi.org/10.1017/S0267190511000092

Dalton-Puffer, C. (2017). *CLIL in practice: What does the research tell us?* Goethe-Institut. Retrieved March 11, 2026, from https://www.goethe.de/prj/dlp/dlapi/v1/index.cfm?endpoint=/tlm/download&file_ID=1768&tlm_ID=358.

Dalton-Puffer, C., Hüttner, J., Jexenflicker, S., Schindlegger, V., & Smit, U. (2008). *CLIL an Österreichischen HTLs*. University of Vienna/BMUKK.

Dalton-Puffer, C., Hüttner, J., & Llinares, A. (2022). CLIL in the 21st century: Retrospective and prospective challenges and opportunities. *Journal of Immersion and Content-Based Language Education*, *10*(2), 182–206. https://doi.org/10.1075/jicb.21021.dal

Dalton-Puffer, C., Nikula, T., & Smit, U. (2010a). Charting policies, premises and research on Content and Language Integrated Learning. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 1–19). John Benjamins.

Dalton-Puffer, C., Nikula, T., & Smit, U. (2010b). Language use and language learning in CLIL: Current findings and contentious issues. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 279–291). John Benjamins.

Dalton-Puffer, C., & Smit, U. (2013). Content and Language Integrated Learning: A research agenda. *Language Teaching*, *46*(4), 545–559. https://doi.org/10.1017/S0261444813000256

Davis, L. E., Ajzen, I., Saunders, J., & Williams, T. (2002). The decision of African American students to complete high school: An application of the theory of planned behavior. *Journal of Educational Psychology*, *94*(4), 810–819. https://doi.org/10.1037/0022-0663.94.4.810

de Bot, K., & Maljers, A. (2009). De enige echte vernieuwing: Tweetalig onderwijs. In R. de Graaff & D. Tuin (Eds.), *De toekomst van het talenonderwijs: Nodig? Anders? Beter?* (pp. 131–146). IVLOS, Universiteit Utrecht.

de Graaff, R., Koopman, G. J., Anikina, Y., & Westhoff, G. (2007). An observation tool for effective L2 pedagogy in Content and Language Integrated Learning (CLIL). *International Journal of Bilingual Education and Bilingualism*, *10*(5), 603–624. https://doi.org/10.2167/beb462.0

de Graaff, R., & van Wilgenburg, O. (2015). The Netherlands: Quality control as a driving force in bilingual education. In P. Mehisto & F. Genesee (Eds.), *Building bilingual education systems: Forces, mechanisms and counterweights* (pp. 167–179). Cambridge University Press.

de Kraay, A. P. (2016). *Differentiation to improve the articulation between levels: In the teaching of English in primary and secondary education in the Netherlands.* [Doctoral dissertation, Universiteit Groningen]. https://hdl.handle.net/11370/45938466-b49e-4882-aef1-3435f2016a20

Denessen, E. J. P. G. (2017). *Verantwoord omgaan met verschillen: Sociale-culturele achtergronden en differentiatie in het onderwijs.* Universiteit Leiden. https://hdl.handle.net/1887/51574

Denman, J., Tanner, R., & de Graaff, R. (2013). CLIL in junior vocational secondary education: Challenges and opportunities for teaching and learning. *International Journal of Bilingual Education and Bilingualism*, *16*(3), 285–300. https://doi.org/10.1080/13670050.2013.777386

Denman, J., van Schooten, E., & de Graaff, R. (2018). Attitudinal factors and the intention to learn English in pre-vocational secondary bilingual and mainstream education. *Dutch Journal of Applied Linguistics*, *7*(2), 203–226. https://doi.org/10.1075/dujal.18005.den

Dentler, S. (2002). Case 17: Mixed ability—Secondary. In D. Marsh (Ed.), *CLIL/EMILE - the European dimension. Actions, trends and foresight potential* (pp. 167–171). University of Jyväskylä.

Dewaele, J.-M. (2019). The effect of classroom emotions, attitudes toward English, and teacher behavior on willingness to communicate among English foreign language learners. *Journal of Language and Social Psychology*, *38*(4), 523–535. https://doi.org/10.1177/0261927X19864996

Dewaele, J.-M., & MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*(2), 237–274. https://doi.org/10.14746/ssllt.2014.4.2.5

Dewaele, J.-M., Saito, K., & Halimi, F. (2025). How teacher behaviour shapes foreign language learners' enjoyment, anxiety and attitudes/motivation: A mixed modelling longitudinal investigation. *Language Teaching Research*, *29*(4), 1580–1602. https://doi.org/10.1177/13621688221089601

Dillon, W. R., & Kumar, A. (1985). Attitude organization and the attitude–behavior relation: A critique of Bagozzi and Burnkrant's reanalysis of Fishbein and Ajzen. *Journal of Personality and Social Psychology*, *49*(1), 33–46. https://doi.org/10.1037/0022-3514.49.1.33

Doiz, A., Lasagabaster, D., & Sierra, J. M. (2014a). CLIL and motivation: The effect of individual and contextual variables. *The Language Learning Journal*, *42*(2), 209–224. https://doi.org/10.1080/09571736.2014.889508

Doiz, A., Lasagabaster, D., & Sierra, J. M. (2014b). Giving voice to the students: What (de)motivates them in CLIL classes? In D. Lasagabaster, A. Doiz, & J. M. Sierra (Eds.), *Language learning & language teaching* (Vol. 40, pp. 117–138). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.40.06doi

Dönszelmann, S. (2020). Doeltaal-leertaal: Taal om van te leren. In S. Dönszelmann, C. van Beuningen, A. Kaal, & R. de Graaff (Eds.), *Handboek vreemdetalendidactiek: Vertrekpunten, vaardigheden, vakinhoud* (pp. 71–86). Uitgeverij Coutinho.

Dönszelmann, S. (2024). *Handbook doeltaaldidactiek: Taalverwerving en doeltaal-leertaal*. Uitgeverij Coutinho.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Lawrence Erlbaum.

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9–42). Multilingual Matters.

Dörnyei, Z. (2019). Towards a better understanding of the L2 learning experience, the Cinderella of the L2 motivational self system. *Studies in Second Language Learning and Teaching*, *9*(1), 19–30. https://doi.org/10.14746/ssllt.2019.9.1.2

Dörnyei, Z., Csizér, K., & Németh, N. (2006). *Motivation, language attitudes and globalisation: A Hungarian perspective*. Multilingual Matters.

Du Bois-Reymond, M., & Altinyelken, H. K. (2011). *Complexities and challenges of learning in Dutch vocational schools* (No. 14) [National report for WP6]. Governance of educational trajectories in Europe (GOETE). Retrieved March 13, 2026, from https://pure.uva.nl/ws/files/1157982/162386_356296.pdf.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Pearson Assessments.

Escobar Urmeneta, C. (2004). Content and Language Integrated Learning: Do they learn content? Do they learn language? In J. D. Anderson, J. M. Oro, & J. Varela (Eds.), *Linguistic perspectives from the classroom: Language teaching in a multicultural Europe* (pp. 27–38). Universidade de Santiago de Compostela.

Escobar Urmeneta, C. (2013). Learning to become a CLIL teacher: Teaching, reflection and professional development. *International Journal of Bilingual Education and Bilingualism*, *16*(3), 334–353. https://doi.org/10.1080/13670050.2013.777389

Escobar Urmeneta, C. (2019). An introduction to Content and Language Integrated Learning (CLIL) for teachers and teacher educators. *CLIL. Journal of Innovation and Research in Plurilingual and Pluricultural Education*, *2*(1), 7–19. https://doi.org/10.5565/rev/clil.21

European Commission. (1995). *Teaching and learning. Towards the learning society. White paper on education and training*. Office for Official Publications of the European Communities. Retrieved March 13, 2026, from https://op.europa.eu/en/publication-detail/-/publication/8722c9fb-71da-435f-94b6-33aab67eb081/language-en.

European Commission/EACEA/Eurydice. (2012). *Key data on teaching languages at school in Europe – 2012 edition*. Publications Office of the European Union. https://eurydice.eacea.ec.europa.eu/publications/key-data-teaching-languages-school-europe-2012

European Education and Culture Executive Agency. (2017). *Key data on teaching languages at school in Europe: 2017 edition*. Publications Office of the European Union. https://doi.org/10.2797/839825

Europees Platform. (n.d.). *Standaard tweetalig onderwijs Engels havo/vwo* [Policy document]. (Later incorporated into Nuffic; document no longer publicly available.)

Europees Platform. (2009). *Toelichting bij de Standaard Europees Platform tweetalig vmbo* [Policy document]. (Later incorporated into Nuffic; document no longer publicly available.)

Fasoglio, D., de Jong, K., Pennewaard, L., Trimbos, B., & Tuin, D. (2015). *Moderne vreemde talen: Vakspecifieke trendanalyse 2015*. SLO.

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

Fen, Y. S., & Sabaruddin, N. A. (2009). An extended model of theory of planned behaviour in predicting exercise intention. *International Business Research*, *1*(4), p108. https://doi.org/10.5539/ibr.v1n4p108

Fernández-Agüero, M., & Hidalgo-McCabe, E. (2022). CLIL students' affectivity in the transition between education levels: The effect of streaming at the beginning of secondary education. *Journal of Language, Identity & Education*, *21*(6), 363–377. https://doi.org/10.1080/15348458.2020.1795864

Feskens, R. C. W., Kuhlemeier, H., & Limpens, G. (2016). *Resultaten PISA-2015 in vogelvlucht: Praktische kennis en vaardigheden van 15-jarigen*. Cito. https://cito.nl/media/qp3ban3a/resultaten-pisa-2015-in-vogelvlucht-praktische-kennis-en-vaardigheden-van-15-jarigen.pdf

Florian, L., & Black-Hawkins, K. (2011). Exploring inclusive pedagogy. In J. Dillon & M. Maguire (Eds.), *Becoming a teacher: Issues in secondary education* (4th ed., pp. 182–193). Open University Press.

Gallagher, F., & Colohan, G. (2017). T(w)o and fro: Using the L1 as a language teaching tool in the CLIL classroom. *The Language Learning Journal*, *45*(4), 485–498. https://doi.org/10.1080/09571736.2014.947382

Gálvez Gómez, M. D. M. (2021). The effects of CLIL on FL learning: A longitudinal study. In M. L. Pérez Cañado (Ed.), *Content and Language Integrated Learning in monolingual settings* (Vol. 38, pp. 141–165). Springer. https://doi.org/10.1007/978-3-030-68329-0_8

Garcia López, M., & Bruton, A. (2013). Potential drawbacks and actual benefits of CLIL initiatives in public secondary schools. In C. Abello-Contesse, P. M. Chandler, R. Chacón-Beltrán, & M. D. López-Jiménez (Eds.), *Bilingual and multilingual education in the 21st century: Building on experience* (pp. 256–274). Multilingual Matters.

García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism, and education*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137385765

Genesee, F. (1986). The baby and the bathwater or what immersion has to say about bilingual education: Teaching and learning in bilingual education—significant immersion instructional features. *NABE Journal*, *10*(3), 227–254. https://doi.org/10.1080/08881685.1986.10668514

Genesee, F. (1987). *Learning through two languages: Studies of immersion and bilingual education*. Newbury House.

Genesee, F. (2004). What do we know about bilingual education for majority-language students? In T. K. Bhatia & W. C. Richie (Eds.), *The handbook of bilingualism* (pp. 547–576). Blackwell.

Genesee, F. (2007). French immersion and at-risk students: A review of research evidence. *The Canadian Modern Language Review*, *63*(5), 655–687. https://doi.org/10.3138/cmlr.63.5.655

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

Gierlinger, E. (2015). 'You can speak German, sir': On the complexity of teachers' L1 use in CLIL. *Language and Education*, *29*(4), 347–368. https://doi.org/10.1080/09500782.2015.1023733

Gierlinger, E. (2017). I feel traumatized: Teachers' beliefs on the roles of languages and learning in CLIL. In K. Valcke & R. Wilkinson (Eds.), *Integrating content and language in higher education: Perspectives on professional practice* (pp. 97–116). Peter Lang.

Gierlinger, E., & Wagner, T. A. (2016). The more the merrier – Revisiting CLIL-based vocabulary growth in secondary education. *Latin American Journal of Content & Language Integrated Learning*, *9*(1), 37–63. https://doi.org/10.5294/laclil.2016.9.1.3

Godin, G., & Kok, G. (1996). The theory of planned behavior: A review of its applications to health-related behaviors. *American Journal of Health Promotion*, *11*(2), 87–98. https://doi.org/10.4278/0890-1171-11.2.87

Gordani, Y., Ezhdehakosh, M., Arabani, A. S., & Bazargani, D. T. (2025). The impact of emotional contagion on second/foreign language learning environments in terms of student motivation and achievement. *Asian-Pacific Journal of Second and Foreign Language Education*, *10*(1), 31. https://doi.org/10.1186/s40862-025-00336-x

Goriot, C., van Hout, R., Broersma, M., Lobo, V., McQueen, J. M., & Unsworth, S. (2021). Using the Peabody Picture Vocabulary Test in L2 children and adolescents: Effects of L1. *International Journal of Bilingual Education and Bilingualism*, *24*(4), 546–568. https://doi.org/10.1080/13670050.2018.1494131

Goris, J., Denessen, E., & Verhoeven, L. (2013). Effects of the Content and Language Integrated Learning approach to EFL teaching: A comparative study. *Written Language & Literacy*, *16*(2), 186–207. https://doi.org/10.1075/wll.16.2.03gor

Goris, J., Denessen, E., & Verhoeven, L. (2019). The contribution of CLIL to learners' international orientation and EFL confidence. *The Language Learning Journal*, *47*(2), 246–256. https://doi.org/10.1080/09571736.2016.1275034

Goris, J., Denessen, E., & Verhoeven, L. (2022). Determinants of EFL learning success in Content and Language Integrated Learning. *The Language Learning Journal*, *50*(1), 103–118. https://doi.org/10.1080/09571736.2019.1709886

Graus, J., & Coppen, P. (2017). The interface between student teacher grammar cognitions and learner-oriented cognitions. *The Modern Language Journal*, *101*(4), 643–668. https://doi.org/10.1111/modl.12427

Greene, B. A, Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology*, *29*(4), 462-482. https://doi.org/10.1016/j.cedpsych.2004.01.006

Hacquebord, H. (2007). De leesvaardigheid van vmbo-leerlingen. In D. Schram (Ed.), *Lezen in het vmbo. Onderzoek – interventie—praktijk* (pp. 55–76). Stichting Lezen.

Hammond, J., & Gibbons, P. (2005). Putting scaffolding to work: The contribution of scaffolding in articulating ESL education. *Prospect*, *20*(1), 6–30.

Hamstra, D. G., & van den Ende, J. (2006). *De vmbo-leerling. Onderwijspedagogische- en ontwikkelingspsychologische theorieën*. CPS.

Harley, B. (1991). Directions in immersion research. *Journal of Multilingual and Multicultural Development*, *12*(1–2), 9–19. https://doi.org/10.1080/01434632.1991.9994441

Harrop, E. (2012). Content and Language Integrated Learning (CLIL): Limitations and possibilities. *Encuentro*, *21*, 57–70.

Hartgers, M. (2007). Scholen in de randstad sterk gekleurd. *Sociaaleconomische Trends*, (3e kwartaal 2007), 12–18.

Heras, A., & Lasagabaster, D. (2015). The impact of CLIL on affective factors and vocabulary learning. *Language Teaching Research*, *19*(1), 70–88. https://doi.org/10.1177/1362168814541736

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 461–473). John Benjamins.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Quantitative Methodology Series. Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huibregtse, I. (2001). *Effecten en didactiek van tweetalig voortgezet onderwijs in Nederland.* [Doctoral dissertation, Universiteit Utrecht].

Huibregtse, I., van der Poel, M., & Hoornweg, J. (1994). *De vreemde taal als instructietaal: Een literatuurstudie naar vormen, voorwaarden en effecten van inhoudgeoriënteerd vreemde talenonderwijs*. KUN / IVLOS.

Hunt, M. (2011). Learners' perceptions of their experiences of learning subject content through a foreign language. *Educational Review*, *63*(3), 365–378. https://doi.org/10.1080/00131911.2011.571765

Hüttner, J., Dalton-Puffer, C., & Smit, U. (2013). The power of beliefs: Lay theories and their influence on the implementation of CLIL programmes. *International Journal of Bilingual Education and Bilingualism*, *16*(3), 267–284. https://doi.org/10.1080/13670050.2013.777385

Hüttner, J., & Smit, U. (2014). CLIL (Content and Language Integrated Learning): The bigger picture. A response to: A. Bruton. 2013. CLIL: Some of the reasons why … and why not. System 41 (2013): 587–597. *System*, *44*, 160–167. https://doi.org/10.1016/j.system.2014.03.001

Iglesias Diéguez, K., & Martínez-Adrián, M. (2017). The influence of CLIL on receptive vocabulary: A preliminary study. *Journal of English Studies*, *15*, 107–134. https://doi.org/10.18172/jes.3210

Inspectie van het Onderwijs. (2008). *Basisvaardigheden taal in het voortgezet onderwijs: Resultaten van een inspectieonderzoek naar taalvaardigheid in de onderbouw van het vmbo en praktijkonderwijs.* https://www.fi.uu.nl/publicaties/literatuur/inspectierapport_basisvaardigheden_vo_2009.pdf

Inspectie van het Onderwijs. (2016). *De staat van het onderwijs 2014/2015*. https://archief26.sitearchief.nl/archives/sitearchief/20250726204644/www.onderwijsinspectie.nl/onderwerpen/staat-van-het-onderwijs/documenten/publicaties/2016/04/13/staat-van-het-onderwijs-2014-2015.

Inspectie van het Onderwijs. (2019). *De staat van het onderwijs 2019*. https://www.onderwijsinspectie.nl/documenten/2019/04/10/rapport-de-staat-van-het-onderwijs-2019.

Iwaniec, J., & Halbach, A. (2024). Teachers' views on their methodology and their profiles: In search of the possible reasons for the levelling effect of CLIL. *Journal of Multilingual and Multicultural Development*, *45*(4), 1112–1126. https://doi.org/10.1080/01434632.2021.1946548

Jäppinen, A.-K. (2005). Thinking and content learning of mathematics and science as cognitional development in Content and Language Integrated Learning (CLIL): Teaching through a foreign language in Finland. *Language and Education*, *19*(2), 147–168. https://doi.org/10.1080/09500780508668671

Jeffery, J. V., & van Beuningen, C. (2020). Language education in the EU and the US: Paradoxes and parallels. *PROSPECTS*, *48*(3–4), 175–191. https://doi.org/10.1007/s11125-019-09449-x

Jexenflicker, S., & Dalton-Puffer, C. (2010). The CLIL differential: Comparing the writing of CLIL and non-CLIL students in higher colleges of technology. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 169–190). John Benjamins.

Jiménez Catalán, R. M., & Pilar Agustín Llach, M. (2017). CLIL or time? Lexical profiles of CLIL and non-CLIL EFL learners. *System*, *66*, 87–99. https://doi.org/10.1016/j.system.2017.03.016

Jiménez Catalán, R. M., & Ruiz de Zarobe, Y. (2009). The receptive vocabulary of EFL learners in two instructional contexts: CLIL versus non-CLIL instruction. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (Eds.), *Content and Language Integrated Learning: Evidence from research in Europe* (pp. 81–92). Multilingual Matters.

Jiménez Catalán, R. M., Ruiz de Zarobe, Y., & Cenoz, J. (2006). Vocabulary profiles of English foreign language learners in English as a subject and as a vehicular language. *Vienna English Working Papers*, *16*(3), 23–27.

Johnson, R. K., & Swain, M. (1994). From core to content: Bridging the L2 proficiency gap in late immersion. *Language and Education*, *8*(4), 211–229. https://doi.org/10.1080/09500789409541392

Juan, M. (2010). Oral fluency development in secondary education CLIL learners. *Vienna English Working Papers*, *19*(3), 42–48.

Kang, S.-J. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System*, *33*(2), 277–292. https://doi.org/10.1016/j.system.2004.10.004

Kashiwagi, K., & Tomecsek, J. (2015). How CLIL classes exert a positive influence on teaching style in student centered language learning through overseas teacher training in Sweden and

Finland. *Procedia - Social and Behavioral Sciences*, *173*, 79–84. https://doi.org/10.1016/j.sbspro.2015.02.034

Keil, L. J. (1991). Attitude development. In T. Husén & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., Vol. 1, pp. 353–358). Pergamon Press.

Kim, H., & Graham, K. M. (2022). Teachers' needs and professional development: A systematic review. *Latin American Journal of Content & Language Integrated Learning*, *15*(1), e1515. https://doi.org/10.5294/laclil.2022.15.1.5

Kneyber, R. (2009). *Orde houden in het vmbo*. Handelingsbrutaal.

Kontio, J., & Sylvén, L. K. (2015). Language alternation and language norm in vocational Content and Language Integrated Learning. *The Language Learning Journal*, *43*(3), 271–285. https://doi.org/10.1080/09571736.2015.1053279

Koomen, H. M. Y., Verschueren, K., van Schooten, E., Jak, S., & Pianta, R. C. (2012). Validating the student-teacher relationship scale: Testing factor structure and measurement invariance across child gender and age in a Dutch sample. *Journal of School Psychology*, *50*(2), 215–234. https://doi.org/10.1016/j.jsp.2011.09.001

Kordes, J., & Gille, E. (2012). *Vaardigheden Engels en Duits van Nederlandse leerlingen in Europees perspectief: Resultaten ESLC 2011*. Cito.

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*(3), 643–680. https://doi.org/10.3102/0034654315626799

Krashen, S. D. (1981). Bilingual education and second language acquisition theory. In C. F. Leyba (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 51–79). California State Department of Education.

Küppers, A., & Trautmann, M. (2013). It is not CLIL that is a success—CLIL students are! Some critical remarks on the current CLIL boom. In S. Breidbach & B. Viebrock (Eds.), *Content and Language Integrated Learning (CLIL) in Europe. Research perspectives on policy and practice* (pp. 285–296). Peter Lang.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania.

Lahuerta Martínez, A. C. (2015). The written competence of Spanish secondary education students in bilingual and non-bilingual programs. *Porta Linguarum*, *24*, 47–61. https://doi.org/10.30827/Digibug.53796

Lahuerta Martínez, A. C. (2020). Analysis of accuracy in the writing of EFL students enrolled on CLIL and non-CLIL programmes: The impact of grade and gender. *The Language Learning Journal*, *48*(2), 121–132. https://doi.org/10.1080/09571736.2017.1303745

Lancaster, N. K. (2016). Stakeholder perspectives on CLIL in a monolingual context. *English Language Teaching*, *9*(2), 148. https://doi.org/10.5539/elt.v9n2p148

Land, J., van den Bergh, H., & Sanders, T. (2007). Leesplezier en leesfrequentie op het vmbo. Een kwantitatief inventarisatie-onderzoek naar leesgedrag en waardering voor studieteksten. In D. Schram (Ed.), *Lezen in het vmbo. Onderzoek – Interventie—Praktijk* (pp. 151–182). Stichting Lezen.

Lasagabaster, D. (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal*, *1*(1), 30–41. https://doi.org/10.2174/1874913500801010030

Lasagabaster, D. (2011). English achievement and student motivation in CLIL and EFL settings. *Innovation in Language Learning and Teaching*, *5*(1), 3–18. https://doi.org/10.1080/17501229.2010.519030

Lasagabaster, D. (2017). Integrating content and foreign language learning: What do CLIL students believe? *Journal of Immersion and Content-Based Language Education*, *5*(1), 4–29. https://doi.org/10.1075/jicb.5.1.01las

Lasagabaster, D., & Sierra, J. M. (2009). Language attitudes in CLIL and traditional EFL classes. *International CLIL Research Journal*, *1*(2), *4*–17.

Lasagabaster, D., & Sierra, J. M. (2010). Immersion and CLIL in English: More differences than similarities. *ELT Journal*, *64*(4), 367–375. https://doi.org/10.1093/elt/ccp082

Ledoux, G., Meijer, J., & van der Veen, I. (2010). *Opvattingen van docenten en leerlingen over leer- en gedragsproblemen in het voortgezet onderwijs*. Kohnstamm Instituut.

Lialikhova, D. (2018). Triggers and constraints of lower secondary students' willingness to communicate orally in English in a CLIL setting in the Norwegian context. *Journal of Immersion and Content-Based Language Education*, *6*(1), 27–56. https://doi.org/10.1075/jicb.16013.lia

Llinares, A. (2015). Integration in CLIL: A proposal to inform research and successful pedagogy. *Language, Culture and Curriculum*, *28*(1), 58–73. https://doi.org/10.1080/07908318.2014.1000925

Lo, Y. Y., & Macaro, E. (2015). Getting used to Content and Language Integrated Learning: What can classroom interaction reveal? *The Language Learning Journal*, *43*(3), 239–255. https://doi.org/10.1080/09571736.2015.1053281

Lorenzo, F., Casal, S., & Moore, P. (2010). The effects of Content and Language Integrated Learning in European education: Key findings from the Andalusian bilingual sections evaluation project. *Applied Linguistics*, *31*(3), 418–442. https://doi.org/10.1093/applin/amp041

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

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. John Benjamins.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

MacDonald, J. R., Clément, R., & MacIntyre, P. D. (2003). *Willingness to communicate in a L2 in a bilingual context: A qualitative investigation of Anglophone and Francophone students* [Unpublished manuscript]. Cape Breton University.

MacIntyre, P. D., Baker, S. C., Clément, R., & Conrod, S. (2001). Willingness to communicate, social support, and language-learning orientations of immersion students. *Studies in Second Language Acquisition*, *23*(3), 369–388. https://doi.org/10.1017/S0272263101003035

MacIntyre, P. D., Baker, S. C., Clément, R., & Donovan, L. A. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Language Learning*, *52*(3), 537–564. https://doi.org/10.1111/1467-9922.00194

MacIntyre, P. D., Baker, S., Clément, R., & Donovan, L. A. (2003). Talking in order to learn: Willingness to communicate and intensive language programs. *The Canadian Modern Language Review*, *59*(4), 589–608. https://doi.org/10.3138/cmlr.59.4.589

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*(4), 545–562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

Madrid, D., & Pérez Cañado, M. L. (2018). Innovations and challenges in attending to diversity through CLIL. *Theory Into Practice*, *57*(3), 241–249. https://doi.org/10.1080/00405841.2018.1492237

Madrid, D., & Barrios, E. (2018). A comparison of students' educational achievement across programmes and school types with and without CLIL provision. *Porta Linguarum*, *29*, 29–50. https://doi.org/10.30827/Digibug.54021

Mady, C., & Arnett, K. (2009). Inclusion in French immersion in Canada: One parent's perspective. *Exceptionality Education International*, *19*(2). https://doi.org/10.5206/eei.v19i2.7642

Mahan, K. R. (2022). The comprehending teacher: Scaffolding in Content and Language Integrated Learning (CLIL). *The Language Learning Journal*, *50*(1), 74–88. https://doi.org/10.1080/09571736.2019.1705879

Mahan, K. R., & Norheim, H. (2021). 'Something new and different': Student perceptions of Content and Language Integrated Learning. *ELT Journal*, *75*(1), 77–86. https://doi.org/10.1093/elt/ccaa057

Marsh, D. (Ed.). (2002). *CLIL/EMILE – The European dimension: Actions, trends, and foresight potential.* European Commission.

Marsh, D. (2012). *Content and Language Integrated Learning (CLIL). A development trajectory*. Servicio de Publicaciones de la Universidad de Córdoba.

Marsh, D., Maljers, A., & Hartiala, A.-K. (2001). *Profiling European CLIL classrooms*. University of Jyväskylä.

Marsh, D., Mehisto, P., Wolff, D., & Frigols Martín, M. J. (2009a). *CLIL: Content and Language Integrated Learning*. Cambridge University Press.

Marsh, D., Mehisto, P., Wolff, D., & Frigols Martín, M. J. (2009b). *European framework for CLIL teacher education: A framework for the professional development of CLIL teachers*. European Centre for Modern Languages.

Martínez Adrián, M., & Gutiérrez Mangado, M. J. (2015). Is CLIL instruction beneficial in terms of general proficiency and specific areas of grammar? *Journal of Immersion and Content-Based Language Education*, *3*(1), 51–76. https://doi.org/10.1075/jicb.3.1.03adr

Martínez Agudo, J. de D. (2019). Which instructional programme (EFL or CLIL) results in better oral communicative competence? Updated empirical evidence from a monolingual context. *Linguistics and Education*, *51*, 69–78. https://doi.org/10.1016/j.linged.2019.04.008

McClintic, G. H. (2022). Attention to diversity in a Spanish CLIL classroom: Teachers' perceptions. *Journal of Language Teaching and Research*, *13*(5), 897–907. https://doi.org/10.17507/jltr.1305.01

McDougald, J. S. (2015). Teachers' attitudes, perceptions and experiences in CLIL: A look at content and language. *Colombian Applied Linguistics Journal*, *17*(1), 25. https://doi.org/10.14483/udistrital.jour.calj.2015.1.a02

Mearns, T. L. (2015). *Chicken, egg, or a bit of both? Motivation in bilingual education (TTO) in the Netherlands.* [Doctoral dissertation, University of Aberdeen]. EThOS.

Mearns, T. L., & de Graaff, R. (2018a). Bilingual education and CLIL in the Netherlands: The paradigm and the pedagogy. *Dutch Journal of Applied Linguistics*, *7*(2), 122–128. https://doi.org/10.1075/dujal.00002.int

Mearns, T. L., & de Graaff, R. (2018b). Bucking the trend? Motivational differences between boys and girls who opt in or out of bilingual education. *Journal of Immersion and Content-Based Language Education*, *6*(1), 1–26. https://doi.org/10.1075/jicb.17003.mea

Mearns, T. L., de Graaff, R., & Coyle, D. (2020). Motivation *for* or *from* bilingual education? A comparative study of learner views in the Netherlands. *International Journal of Bilingual Education and Bilingualism*, *23*(6), 724–737. https://doi.org/10.1080/13670050.2017.1405906

Mearns, T. L., van Kampen, E., & Admiraal, W. (2023). CLIL in the Netherlands: Three decades of innovation and development. In D. L. Banegas & S. Zappa-Hollman (Eds.), *The Routledge handbook of Content and Language Integrated Learning* (1st ed., pp. 403–418). Routledge.

Menezes, E., & Juan-Garau, M. (2015). English learners' willingness to communicate and achievement in CLIL and formal instruction contexts. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 221–236). Springer.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance: *Medical Care*, *44*(Suppl 3), S69–S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Merikivi, R., & Pietilä, P. (2014). Vocabulary in CLIL and in mainstream education. *Journal of Language Teaching and Research*, *5*(3), 487–497. https://doi.org/10.4304/jltr.5.3.487-497

Merino, J. A., & Lasagabaster, D. (2018). The effect of Content and Language Integrated Learning programmes' intensity on English proficiency: A longitudinal study. *International Journal of Applied Linguistics*, *28*(1), 18–30. https://doi.org/10.1111/ijal.12177

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

Merisuo-Storm, T. (2011). *Literacy skills and learning attitudes in bilingual education. Education in Finland*. Taiwan National Institute of Educational Resource and Research.

Mewald, C. (2007). A comparison of oral language performance of learners in CLIL and mainstream classes at lower secondary level in Lower Austria. In C. Dalton-Puffer & U. Smit (Eds.), *Empirical perspectives on CLIL classroom discourse* (pp. 139–178). Peter Lang.

Milla Lara, M. D., & Casas Pedrosa, A. V. (2018). Teacher perspectives on CLIL implementation: A within-group comparison of key variables. *Porta Linguarum*, *29*, 159–180. https://doi.org/10.30827/Digibug.54032

Mitchell, R., Myles, F., & Marsden, E. (2019). *Second language learning theories* (4th ed.). Routledge.

Möller, V. (2016). *A question of cause and effect: How educational settings at the primary and secondary level interact with intelligence, motivation and anxiety* [Paper presentation]. Focus on the Learner 2016, Konin, Poland.

Möller, V. (2017). *Language acquisition in CLIL and non-CLIL settings: Learner corpus and experimental evidence on passive constructions*. John Benjamins.

Morinaj, J., Hadjar, A., & Hascher, T. (2020). School alienation and academic achievement in Switzerland and Luxembourg: A longitudinal perspective. *Social Psychology of Education*, *23*(2), 279–314. https://doi.org/10.1007/s11218-019-09540-3

Morton, T. (2016). Content and Language Integrated Learning. In G. Hall (Ed.), *The Routledge handbook of English language teaching*. Routledge.

Muthén, L. K., & Muthén, B. O. (2006). *MPlus user's guide* (4th ed.). Muthén & Muthén.

Naber, R., & Lowie, W. (2012). Hoe vroeger, hoe beter? Een onderzoek naar de effectiviteit van vroeg vreemdetalenonderwijs. *Levende Talen Tijdschrift*, *13*(1), 13–21.

Nation, I. S. P. (2017). *The BNC/COCA headword lists* (Version 1.0.0.) [Dataset]. Victoria University of Wellington, School of Linguistics and Applied Language Studies. Retrieved March 6, 2026, from https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

Navés, T. (2011). How promising are the results of integrating content and language for EFL writing and overall EFL proficiency? In Y. Ruiz de Zarobe, J. M. Sierra, & F. Gallardo del Puerto (Eds.), *Content and foreign language integrated learning: Contributions to multilingualism in European contexts* (pp. 155–186). Peter Lang.

Navés, T., & Victori, M. (2010). CLIL in Catalonia: An overview of research studies. In D. Lasagabaster & Y. Ruiz de Zarobe (Eds.), *CLIL in Spain: Implementation, results and teacher training* (pp. 30–54). Cambridge Scholars.

Nederlands Jeugdinstituut. (2021). *Cijfers over jeugd met een migratieachtergrond*.

Nieto Moreno de Diezmas, E. (2016). The impact of CLIL on the acquisition of L2 competences and skills in primary education. *International Journal of English Studies*, *16*(2), 81–101. https://doi.org/10.6018/ijes/2016/2/239611

Nikula, T. (2017a). CLIL: A European approach to bilingual education. In N. Van Deusen-Scholl & S. May (Eds.), *Second and foreign language education* (pp. 111–124). Springer. https://doi.org/10.1007/978-3-319-02246-8_10

Nikula, T. (2017b). Translanguaging in CLIL: Tensions between theories and practices. In B. Paulsrud, J. Rosén, B. Straszer, & Å. Wedin (Eds.), *New perspectives on translanguaging and education* (pp. 233–254). Multilingual Matters.

Nuffic. (n.d.-a). *Global citizen competencies framework [Concept version]*. Retrieved November 19, 2025, from https://www.nuffic.nl/sites/default/files/2025-10/Global%20Citizen%20Competencies.pdf.

Nuffic. (n.d.-b). *Standaard voor tweetalig vmbo*. Nuffic. Retrieved November 19, 2025, from https://www.nuffic.nl/standaard-voor-t-vmbo.

Nuffic. (2025). *Grow your world: Ontwikkelagenda TTO 2025–2030 [Development Agenda for Bilingual Education 2025–2030]*. Retrieved November 19, 2025, from https://www.nuffic.nl/en/grow-your-world-ontwikkelagenda-tto-2025-2030.

OECD. (2023). *PISA 2022 Results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Otwinowska, A., & Foryś, M. (2017). They learn the CLIL way, but do they like it? Affectivity and cognition in upper-primary CLIL classes. *International Journal of Bilingual Education and Bilingualism*, *20*(5), 457–480. https://doi.org/10.1080/13670050.2015.1051944

Oxbrow, G. (2018). Students' perspectives on CLIL programme development: A quantitative analysis. *Porta Linguarum*, *29*, 137–158. https://doi.org/10.30827/Digibug.54026

Paran, A. (2013). Content and Language Integrated Learning: Panacea or policy borrowing myth? *Applied Linguistics Review*, *4*(2), 317–342. https://doi.org/10.1515/applirev-2013-0014

Paulsrud, B. (2018). Just a little plus: The CLIL student perspective. In L. K. Sylvén (Ed.), *Investigating Content and Language Integrated Learning: Insights from Swedish high schools* (pp. 282–297). Multilingual Matters.

Pels, T., Jonkman, H., & Drost, L. (2011). *Socialiseren, leren, presteren: Over motivatie van leerlingen in het voortgezet onderwijs*. Verwij-Jonker Instituut. https://www.verwey-jonker.nl/doc/jeugd/Socialiseren_Leren_Presteren_2845_web.pdf

Pérez Cañado, M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, *15*(3), 315–341. https://doi.org/10.1080/13670050.2011.630064

Pérez Cañado, M. L. (2016). Are teachers ready for CLIL? Evidence from a European study. *European Journal of Teacher Education*, *39*(2), 202–221. https://doi.org/10.1080/02619768.2016.1138104

Pérez Cañado, M. L. (2018). CLIL and educational level: A longitudinal study on the impact of CLIL on language outcomes. *Porta Linguarum*, *29*, 51–70. https://doi.org/10.30827/Digibug.54022

Pérez Cañado, M. L. (2020). CLIL and elitism: Myth or reality? *The Language Learning Journal*, *48*(1), 4–17. https://doi.org/10.1080/09571736.2019.1645872

Pérez Cañado, M. L. (2023). Inclusion and diversity in bilingual education: A European comparative study. *International Journal of Bilingual Education and Bilingualism*, *26*(9), 1129–1145. https://doi.org/10.1080/13670050.2021.2013770

Pérez Cañado, M. L., & Lancaster, N. K. (2017). The effects of CLIL on oral comprehension and production: A longitudinal case study. *Language, Culture and Curriculum*, *30*(3), 300-316. https://doi.org/10.1080/07908318.2017.1338717

Pérez Vidal, C. (2009). The integration of content and language in the classroom: A European approach to education (the second time around). In E. Dafouz & M. C. Guerrini (Eds.), *CLIL across educational levels* (pp. 3–16). Richmond.

Pérez Vidal, C., & Roquet, H. (2015). CLIL in context: Profiling language abilities. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 237–255). Springer.

Piesche, N., Jonkmann, K., Fiege, C., & Keßler, J.-U. (2016). CLIL for all? A randomised controlled field experiment with sixth-grade students on the effects of content and language integrated science learning. *Learning and Instruction*, *44*, 108–116. https://doi.org/10.1016/j.learninstruc.2016.04.001

Pihko, M. K. (2007). Foreign language anxiety in Content and Language Integrated Learning (CLIL) and in traditional foreign language classes. In S. Tella (Ed.), *From brawn to brain: Strong signals in foreign language education* (pp. 129–142). University of Helsinki.

Pinner, R. (2021). Authenticity and motivation in CLIL: Creating a meaningful purpose by experiencing the language in use. In C. Hemmi & D. L. Banegas (Eds.), *International perspectives on CLIL* (pp. 23–41). Springer. https://doi.org/10.1007/978-3-030-70095-9_2

Pittas, E., & Tompkins, L. (2024). A systematic review of student learning outcomes in CLIL in LOTE. *Frontiers in Education*, *9*, 1447270. https://doi.org/10.3389/feduc.2024.1447270

Pladevall-Ballester, E. (2015). Exploring primary school CLIL perceptions in Catalonia: Students', teachers' and parents' opinions and expectations. *International Journal of Bilingual Education and Bilingualism*, *18*(1), 45–59. https://doi.org/10.1080/13670050.2013.874972

Pokrivčáková, S. (2013). Latest trends in CLIL research. In M. Bilá & K. Zákutná (Eds.), *Language, literature and culture in a changing transatlantic world II (Part II: literature and methodology of English language teaching)* (pp. 215–223). University of Prešov.

Prieto-Arranz, J. I., Rallo Fabra, L., Calafat-Ripoll, C., & Catrain-González, M. (2015). Testing progress on receptive skills in CLIL and non-CLIL contexts. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 123–137). Springer.

Rallo Fabra, L., & Jacob, K. (2015). Does CLIL enhance oral skills? Fluency and pronunciation errors by Spanish-Catalan learners of English. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 163–177). Springer.

Rasbash, J., Steele, F., Browne, W., & Goldstein, H. (2009). *A user's guide to MLwiN, Version 2.10*. Centre for Multilevel Modelling, University of Bristol.

Roiha, A. S. (2014). Teachers' views on differentiation in Content and Language Integrated Learning (CLIL): Perceptions, practices and challenges. *Language and Education*, *28*(1), 1–18. https://doi.org/10.1080/09500782.2012.748061

Roiha, A. S. (2019). Investigating former pupils' experiences and perceptions of CLIL in Finland: A retrospective analysis. *Nordic Journal of Studies in Educational Policy*, *5*(2), 92–103. https://doi.org/10.1080/20020317.2019.1586514

Roquet, H., Llopis, J., & Pérez-Vidal, C. (2016). Does gender have an impact on the potential benefits learners may achieve in two contexts compared: Formal instruction and formal instruction + Content and Language Integrated Learning? *International Journal of Bilingual Education and Bilingualism*, *19*(4), 370–386. https://doi.org/10.1080/13670050.2014.992389

Roquet, H., & Pérez-Vidal, C. (2015). Do productive skills improve in Content and Language Integrated Learning contexts? The case of writing. *Applied Linguistics*, *38*(4), 489–511. https://doi.org/10.1093/applin/amv050

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart & Winston.

Ruiz de Zarobe, Y. (2008). CLIL and foreign language learning: A longitudinal study in the Basque Country. *International CLIL Research Journal*, *1*(1), 60–73.

Ruiz de Zarobe, Y. (2010). Written production and CLIL: An empirical study. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 191–212). John Benjamins.

Ruiz de Zarobe, Y. (2015). The effects of implementing CLIL in education. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 51–68). Springer.

Ruiz de Zarobe, Y. (2023). CLIL and educational policy. In D. L. Banegas & S. Zappa-Hollman (Eds.), *The Routledge handbook of Content and Language Integrated Learning* (pp. 99–111). Routledge.

Rumlich, D. (2013). Students' general English proficiency prior to CLIL: Empirical evidence for substantial differences between prospective CLIL and non-CLIL students in Germany. In S. Breidbach & B. Viebrock (Eds.), *Content and Language Integrated Learning (CLIL) in Europe. Research perspectives on policy and practice* (pp. 181–201). Peter Lang.

Rumlich, D. (2014). Prospective CLIL and non-CLIL students' interest in English (classes): A quasi-experimental study on German sixth-graders. In R. Breeze, C. Llamas Saíz, C. Martínez Pasamar, & C. Tabernero Sala (Eds.), *The integration of theory and practice in CLIL* (pp. 75–96). Rodopi.

Rumlich, D. (2016). *Evaluating bilingual education in Germany: CLIL students' general English proficiency, EFL self-concept and interest*. Peter Lang.

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

Saladrigues, G., & Llanes, À. (2010). Examining the impact of exposure on L2 development with CLIL and non-CLIL teenage students. *Sintagma: Revista de Lingüística*, *26*, 133–147.

San Isidro, X. (2010). An insight into Galician CLIL: Provision and results. In Y. Ruiz de Zarobe & D. Lasagabaster (Eds.), *CLIL in Spain: Implementation, results and teacher training* (pp. 55–78). Cambridge Scholars.

San Isidro, X., & Lasagabaster, D. (2019). Monitoring of teachers' views on both CLIL and the development of pluriliteracies: A longitudinal qualitative study. *English Language Teaching*, *12*(2), 1–16. https://doi.org/10.5539/elt.v12n2p1

San Isidro, X., & Pérez Cañado, M. L. (2023). The learner's perspective on CLIL: Attitudes, motivations, and perceptions. In D. L. Banegas & S. Zappa-Hollman (Eds.), *The Routledge handbook of Content and Language Integrated Learning* (pp. 268–283).

Sandberg, Y. (2017). CLIL in English in Sweden: Why, how and what? Student perspectives. In C. M. Coonan, L. Favaro, & M. Menegale (Eds.), *A journey through the Content and Language Integrated Learning landscape: Problems and prospects* (pp. 29–44). Cambridge Scholars.

Scheltens, F., Hollenberg-Vos, J., Limpens, G., & Stolwijk, R. (2020). Testing in mathematics education in the Netherlands. In M. van den Heuvel-Panhuizen (Ed.), *National reflections on the Netherlands didactics of mathematics* (pp. 303–330). Springer. https://doi.org/10.1007/978-3-030-33824-4_16

Schijf, G. M. (2009). *Lees- en spellingvaardigheden van brugklassers*. Kohnstamm Instituut.

Schwab, G. (2011). From dialogue to multilogue: A different view on participation in the English foreign-language classroom. *Classroom Discourse*, *2*(1), 3–19. https://doi.org/10.1080/19463014.2011.562654

Schwab, G. (2013). Bili für alle? Ergebnisse und Perspektiven eines Forschungsprojektes zur Einführung bilingualer Module in einer Hauptschule. In S. Breidbach & B. Viebrock (Eds.), *Content*

and Language Integrated Learning (CLIL) in Europe. Research perspectives on policy and practice (pp. 297–314). Peter Lang.

Schwab, G. (2014). CLIL and non-CLIL students in German secondary schools: Differences in motivation, beliefs, and attitudes. *International CLIL Research Journal*, *3*(1), 35–46.

Schwab, G. (2015). Teaching methods and approaches: Looking into a unique CLIL classroom in Germany. In C. J. Jenks & P. Seedhouse (Eds.), *International perspectives on ELT classroom interaction* (pp. 10–27). Palgrave Macmillan.

Schwab, G., Keßler, J.-U., & Hollm, J. (2014). CLIL goes Hauptschule. Chancen und Herausforderungen bilingualen Unterrichts an einer Hauptschule. Zentrale Ergebnisse einer Longitudinalstudie. *Zeitschrift für Fremdsprachenforschung*, *25*(1), 3–37.

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

Sideridis, G. D., & Padeliadu, S. (2001). The motivational determinants of students at risk of having reading difficulties: Planned behavior theory and goal importance. *Remedial and Special Education*, *22*(5), 268–279. https://doi.org/10.1177/074193250102200502

Sieben, I., & van Ginderen, N. (2014). De keuze voor tweetalig onderwijs. *Mens & Maatschappij*, *89*(3), 233–255. https://doi.org/10.5117/MEM2014.3.SIEB

Siepmann, P., Rumlich, D., Matz, F., & Römhild, R. (2023). Attention to diversity in German CLIL classrooms: Multi-perspective research on students' and teachers' perceptions. *International Journal of Bilingual Education and Bilingualism*, *26*(9), 1080–1096. https://doi.org/10.1080/13670050.2021.1981821

Sierra, J. M., & Lasagabaster, D. (2008). Los programas AICLE en aulas diversas: ¿Una alternativa para todos? [CLIL programs in different classrooms: An alternative for all?]. *Lenguaje y Textos*, *28*, 131–142.

Simons, M., Vanhees, C., Smits, T., & Van De Putte, K. (2019). Remedying Foreign Language Anxiety through CLIL? A mixed-methods study with pupils, teachers and parents. *Revista de Lingüística y Lenguas Aplicadas*, *14*(1), 153–172. https://doi.org/10.4995/rlyla.2019.10527

Skinnari, K., & Bovellan, E. (2016). CLIL teachers' and students' beliefs: Reflections on their relationship to classroom practices. In T. Nikula, E. Dafouz, P. Moore, & U. Smit (Eds.), *Conceptualising integration in CLIL and multilingual education* (pp. 145–167). Multilingual Matters.

Smala, S. (2022). Situated emergence of CLIL. In E. Codó, *Global CLIL: Critical, ethnographic and language policy perspectives* (1st ed., pp. 52–73). Routledge. https://doi.org/10.4324/9781003147374-4

Smiskova, H., Verspoor, M., & Lowie, W. (2012). Conventionalized ways of saying things (CWOSTs) and L2 development. *Dutch Journal of Applied Linguistics*, *1*(1), 125–142. https://doi.org/10.1075/dujal.1.1.09smi

Somers, T. (2017). Content and Language Integrated Learning and the inclusion of immigrant minority language students: A research review. *International Review of Education*, *63*(4), 495–520. https://doi.org/10.1007/s11159-017-9651-4

Somers, T. (2018). Multilingualism for Europeans, monolingualism for immigrants? Towards policy-based inclusion of immigrant minority language students in Content and Language Integrated Learning (CLIL). *European Journal of Language Policy*, *10*(2), 203–228. https://doi.org/10.3828/ejlp.2018.12

Somers, T., & Llinares, A. (2021). Students' motivation *for* Content and Language Integrated Learning and the role of programme intensity. *International Journal of Bilingual Education and Bilingualism*, *24*(6), 839–854. https://doi.org/10.1080/13670050.2018.1517722

Sultan, S., Borland, H., & Eckersley, B. (2012, July). *English as medium of instruction (EMI) in Indonesian public junior secondary schools: Students' language use, attitudes/motivation, and foreign*

language outcomes. Paper presented at the ACTA International TESOL Conference, Victoria University, Cairns, Australia.

Surmont, J., Van de Craen, P., Struys, E., & Somers, T. (2014). Evaluating a CLIL student: Where to find the CLIL advantage. In R. Breeze, C. Llamas Saíz, C. Martínez Pasamar, & C. Tabernero Sala (Eds.), *Integration of theory and practice in CLIL* (pp. 55–74). Brill.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–256). Newbury House.

Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review*, *50*(1), 158–164. https://doi.org/10.3138/cmlr.50.1.158

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation* (Vol. 55, pp. 37–76). Academic Press.

Sweller, J. (2017). Cognitive load theory and teaching English as a second language to adult learners. *TESL Ontario CONTACT Magazine*, *43*(3), 1–8.

Sylvén, L. K. (2004). *Teaching in English or English teaching? On the effects of Content and Language Integrated Learning on Swedish learners' incidental vocabulary acquisition.* [Doctoral dissertation, Göteborgs Universitet]. http://hdl.handle.net/2077/16228

Sylvén, L. K., & Ohlander, S. (2014). The CLISS project: Receptive vocabulary in CLIL versus non-CLIL groups. *Moderna Språk*, *108*(2), 80–114. https://doi.org/10.58221/mosp.v108i2.8002

Sylvén, L. K., & Thompson, A. S. (2015). Language learning motivation and CLIL: Is there a connection? *Journal of Immersion and Content-Based Language Education*, *3*(1), 28–50. https://doi.org/10.1075/jicb.3.1.02syl

Tammenga-Helmantel, M., & Mossing Holsteijn, L. (2016). *Doeltaal in het moderne vreemde talen onderwijs: Gebruik en opvattingen van leraren in opleiding*. Rijksuniversiteit Groningen & Meesterschapsteam Vakdidactiek MVT.

Tanner, R., & de Graaff, R. (2011). *Proud to be tvmbo: Teachers and students' opinions about good practice in bilingual junior secondary vocational education (tweetalig vmbo)*. Utrecht Centre for Teaching and Learning, Utrecht University.

Tedick, D. J. (2022). Looking back and looking forward: An afterword. *Journal of Immersion and Content-Based Language Education*, *10*(2), 400–415. https://doi.org/10.1075/jicb.22011.ted

Tennant, P. W. G., Arnold, K. F., Ellison, G. T. H., & Gilthorpe, M. S. (2022). Analyses of 'change scores' do not estimate causal effects in observational data. *International Journal of Epidemiology*, *51*(5), 1604–1615. https://doi.org/10.1093/ije/dyab050

Thijs, A. M., Trimbos, B., Bodde, M., & de Graaff, R. (2011). *Engels in het basisonderwijs*. SLO.

Thompson, L., Kenyon, D. M., & Rhodes, N. C. (2002). *A validation study of the Student Oral Proficiency Assessment (SOPA)*. Center for Applied Linguistics. https://eric.ed.gov/?id=ED465287

Ting, Y. L. T. (2010). CLIL appeals to how the brain likes its information: Examples from CLIL-(neuro) science. *International CLIL Research Journal*, *1*(3), 1–18.

Tompkins, F. L. (2022). Socioeconomic status, English exposure and CLIL motivation in high and low exposure CLIL groups. *CLIL Journal of Innovation and Research in Plurilingual and Pluricultural Education*, *5*(1), 41–52. https://doi.org/10.5565/rev/clil.67

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, *19*(6), 349–357. https://doi.org/10.1093/intqhc/mzm042

Trafimow, D., & Sheeran, P. (1998). Some tests of the distinction between cognitive and affective beliefs. *Journal of Experimental Social Psychology*, *34*(4), 378–397. https://doi.org/10.1006/jesp.1998.1356

Trapman, M., van Gelderen, A., van Steensel, R., van Schooten, E., & Hulstijn, J. (2014). Linguistic knowledge, fluency and meta-cognitive knowledge as components of reading comprehension in adolescent low achievers: Differences between monolinguals and bilinguals. *Journal of Research in Reading*, *37*(S1), S3–S21. https://doi.org/10.1111/j.1467-9817.2012.01539.x

Triandis, H. C. (1977). *Interpersonal behavior*. Brooks/Cole.

University of Cambridge ESOL Examinations. (2012). *Cambridge English: Key for schools – Teacher's handbook*. Cambridge University Press.

Valois, P., Desharnais, R., & Godin, G. (1988). A comparison of the Fishbein and Ajzen and the Triandis attitudinal models for the prediction of exercise intention and behavior. *Journal of Behavioral Medicine*, *11*(5), 459–472. https://doi.org/10.1007/BF00844839

van Batenburg, E. S. L., Oostdam, R. J., van Gelderen, A. J. S., Fukkink, R. G., & de Jong, N. H. (2019). Oral interaction in the EFL classroom: The effects of instructional focus and task type on learner affect. *The Modern Language Journal*, *103*(1), 308–326. https://doi.org/10.1111/modl.12545

van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895–922. https://doi.org/10.1080/00273171.2013.831743

Van de Craen, P., Surmont, J., Ceuleers, E., & Allain, L. (2013). How policies influence multilingual education and the impact of multilingual education on practices. In F. Grin & G. Lüdi (Eds.), *Exploring the dynamics of multilingualism: The DYLAN project* (pp. 343–361). John Benjamins.

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, *47*(2), 497–527. https://doi.org/10.3102/0002831209353594

van Dongen, A. H. (2017). *Issues in the implementation of CLIL in pre-vocational education in the Netherlands.* [Doctoral dissertation, University of Exeter]. https://ore.exeter.ac.uk/articles/thesis/Issues_in_the_Implementation_of_CLIL_in_Pre-Vocational_Education_in_The_Netherlands/29739326/1

van Kampen, E. (2019). *What's CLIL about bilingual education? A window on Content and Language Integrated Learning pedagogies.* [Doctoral dissertation, Universiteit Leiden]. https://hdl.handle.net/1887/77223

van Kampen, E., Meirink, J., Admiraal, W., & Berry, A. (2020). Do we all share the same goals for Content and Language Integrated Learning (CLIL)? Specialist and practitioner perceptions of 'ideal' CLIL pedagogies in the Netherlands. *International Journal of Bilingual Education and Bilingualism*, *23*(8), 855–871. https://doi.org/10.1080/13670050.2017.1411332

Van Mensel, L., Hiligsmann, P., Mettewie, L., & Galand, B. (2020). CLIL, an elitist language learning approach? A background analysis of English and Dutch CLIL pupils in French-speaking Belgium. *Language, Culture and Curriculum*, *33*(1), 1–14. https://doi.org/10.1080/07908318.2019.1571078

van Nuland, H. J. C. (2011). *Eliciting classroom motivation: Not a piece of cake.* [Doctoral dissertation, Universiteit Leiden]. https://hdl.handle.net/1887/16693

van Schooten, E., de Glopper, K., & Stoel, R. D. (2004). Development of attitude toward reading adolescent literature and literary reading behavior. *Poetics*, *32*(5), 343–386. https://doi.org/10.1016/j.poetic.2004.07.001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Várkuti, A. (2010). Linguistic benefits of the CLIL approach: Measuring linguistic competences. *International CLIL Research Journal*, *1*(3), 67–79.

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

Verspoor, M., & Edelenbos, P. (2011). Tweetalig onderwijs zorgt voor een duurzame voorsprong. *Levende Talen Tijdschrift*, *12*(4), 3–13.

Verspoor, M. H., de Bot, K., & van Rein, E. (2011). English as a foreign language: The role of out-of-school language input. In A. De Houwer & A. Wilton (Eds.), *English in Europe today: Sociocultural and educational perspectives* (pp. 147–166). John Benjamins.

Verspoor, M. H., Schuitemaker-King, J., van Rein, E. M. J., de Bot, K., & Edelenbos, P. (2010). *Tweetalig onderwijs: Vormgeving en prestaties*. Rijksuniversiteit Groningen.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239–263. https://doi.org/10.1016/j.jslw.2012.03.007

Verspoor, M., Xu, X., & de Bot, C. J. L. (2013). *Verslag OTTO-2 aan Europees Platform*. Rijksuniversiteit Groningen.

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

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Weenink, D. (2005). *Upper middle class resources of power in the education arena: Dutch elite schools in an age of globalization.* [Doctoral dissertation, Universiteit van Amsterdam]. https://edepot.wur.nl/117531

Wegner, A. (2012). Seeing the bigger picture: What students and teachers think about CLIL. *International CLIL Research Journal*, *1*(4), 29–35.

Westhoff, G. (1998). *Doeltaal–voertaal: Naar meer vreemde talen in het voortgezet onderwijs*. SLO.

Westhoff, G. (2005). Engels is niet genoeg. *Het Onderwijsblad*, *1*, 27–28.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. University of Hawaii at Manoa.

Wouters, G. (2012). Tweetalig onderwijs: Een voorsprong voor het leven. *Levende Talen Magazine*, *99*(3), 12–16.

Zanoni, F. (2021). Students' view on CLIL: Perceived benefits and limitations. *Educazione Linguistica—Language Education*, *10*(2), 261–278. https://doi.org/10.30687/ELLE/2280-6792/2021/02/005

Zhang, F. (2022). A theoretical review on the impact of EFL/ESL students' self-sabotaging behaviors on their self-esteem and academic engagement. *Frontiers in Psychology*, *13*, 873734. https://doi.org/10.3389/fpsyg.2022.873734

Zink, C. (2016). *Bilingual classes for all or for a selected few? CLIL with less privileged students* [Unpublished manuscript]. Freie Universität Berlin, Institute of English Didactics.

Zydatiß, W. (2007). *Deutsch-Englische Züge im Berliner Sekundarbereich I: Empirische Untersuchung zur Bilingualen Sprachkompetenzentwicklung*. Peter Lang.

Zydatiß, W. (2012). Linguistic thresholds in the CLIL classroom? The threshold hypothesis revisited. *International CLIL Research Journal*, *1*(4), 17–28.

Appendix 2

Attachment 2.1: **Biodata questionnaire for CLIL pupils**
(non-CLIL pupils completed the same questionnaire, minus the final five CLIL-related questions).

Naam _____ Leeftijd_____ Meisje / Jongen *(onderstrepen)*

School_____ Klas_____

Ik ben geboren in: ☐ Nederland ☐ Anders, namelijk _____

Wat was je CITO score op de basisschool? _____ ☐ weet ik niet

Heb je ooit in een Engelstalig land gewoond? ☐ ja, (land) _____ ☐ nee

Hoeveel jaar Engels had je op de basisschool? ☐ 1 ☐ 2 ☐ 3 ☐ 4 of meer

Hoeveel lessen per week had je Engels op de basisschool? ☐ 1 ☐ 2 ☐ 3 ☐ 4 of meer

Sprak de juf of meester in het Engels of in het Nederlands in die lessen Engels?
☐ meestal Engels ☐ meestal Nederlands ☐ weet ik niet meer

Spraken de leerlingen Engels of Nederlands in die lessen Engels?
☐ meestal Engels ☐ meestal Nederlands ☐ weet ik niet meer

Zijn je ouders in Nederland geboren? ☐ ja, vader ☐ ja, moeder
☐ nee, mijn vader is geboren in _____
☐ nee, mijn moeder is geboren in _____

Wat is de moedertaal (eerste taal) van je vader? _____
Wat is de moedertaal (eerste taal) van je moeder? _____

Welke taal spreek je meestal met je vader? _____
Welke taal spreek je meestal met je moeder? _____

Kan je vader Engels? ☐ ja ☐ nee ☐ weet ik niet
Kan je moeder Engels? ☐ ja ☐ nee ☐ weet ik niet

Heb je oudere broers of zussen die Engels hebben geleerd? ☐ ja ☐ nee

Hebben jullie thuis een Nederlandse krant? ☐ ja ☐ nee ☐ weet ik niet
Zo ja, welke? _____ ☐ weet ik niet

Wat is het beroep van je vader? _____ ☐ weet ik niet
Wat is het beroep van je moeder? _____ ☐ weet ik niet

Heeft je moeder of vader gestudeerd aan de HBO of universiteit?
☐ ja, mijn vader ☐ ja, mijn moeder ☐ nee, geen van beide ☐ weet ik niet

Ongeveer hoeveel boeken hebben jullie thuis (van het hele gezin, geen stripboeken)?
☐ meer dan 500 ☐ 100 – 500 ☐ 50 – 100 ☐ 10 – 50 ☐ minder dan 10

Helpen je ouders met je huiswerk?
☐ ja, bijna iedere dag ☐ ja, wel eens ☐ bijna nooit ☐ nooit

Komen je ouders (of één van je ouders) op school voor een ouderavond of gesprek?
☐ ja, altijd ☐ ja, wel eens ☐ nee ☐ weet ik niet

Hoeveel uur per dag (gemiddeld) kijk je naar TV programma's **in het Engels**?
☐ meer dan 2 uur ☐ 1 – 2 uur ☐ een beetje, maar minder dan 1 uur ☐ niet

Als je TV programma's **in het Engels** kijkt, kijk je **met** of **zonder** Nederlandse ondertiteling?
☐ altijd met ☐ soms met, soms zonder ☐ altijd zonder ☐ ik kijk geen TV in het Engels

Hoeveel uur per dag (gemiddeld) luister je naar muziek **in het Engels** (ook YouTube & video clips)?
☐ meer dan 2 uur ☐ 1 – 2 uur ☐ een beetje, maar minder dan 1 uur ☐ niet

Hoeveel uur per dag (gemiddeld) ben je op Internet sites **in het Engels**?
☐ meer dan 2 uur ☐ 1 – 2 uur ☐ een beetje, maar minder dan 1 uur ☐ niet

Hoeveel uur per dag (gemiddeld) speel je games online of op de computer **in het Engels**?
☐ meer dan 2 uur ☐ 1 – 2 uur ☐ een beetje, maar minder dan 1 uur ☐ niet

Als je online games (of games op de computer) **in het Engels** speel, welke games zijn dat?

_____

Hoe vaak **per week** spreek je Engels buiten school (met vrienden, familie, kennissen, toeristen, enz.)?
☐ iedere dag ☐ een paar keer per week ☐ 1 keer per week of minder ☐ nooit

Hoe vaak lees je in het Engels buiten school (**niet je huiswerk!)**?
☐ iedere dag ☐ een paar keer per week ☐ 1 keer per week of minder ☐ nooit

Gebruik je Engels als je met vakantie gaat? ☐ ja, altijd ☐ ja, wel eens ☐ nee
Zo ja, in welke landen?_____

Wat vind je het leukst bij **het vak Engels** op school?

_____

Wat vind je het leukst aan **TTO (tweetalig onderwijs)?**

_____

Wat vind je het meest vervelend bij **het vak Engels** op school?

_____

Wat vind je het meest vervelend aan **TTO (tweetalig onderwijs)?**

_____

Jij zit in een TTO klas. Als je nog een keer zou kiezen, zou je weer TTO doen?
☐ ja, zeker! ☐ ja, maar het is moeilijker dan ik dacht ☐ misschien niet ☐ nee, ik zou het niet doen

Note. *In addition to variables reported in the analyses, this biodata questionnaire included items designed to capture socioeconomic background, English-language help or exposure via family, and extracurricular exposure to English. These items were piloted but excluded from analysis due to low response quality, ambiguity in interpretation, or overlap between categories that precluded reliable estimation.*

Attachment 2.2: **MPB Questionnaire** (translated from Dutch): Original 40 items. Italicized items indicate later deletions.

Item	***Cognition***
1	English is one of the most important school subjects.
2	*Learning English at school is a waste of time.*
3	If I learn English well, it will help me everywhere in the world.
4	*English is important for my future job.*
5	If I am good at English, people will respect me.
6	It's important to get good marks for English.
7	English is important for my future training or studies.
8	If I learn English better, I can speak with more and different kinds of people.
9	By learning English, you also learn Dutch better.
10	When you learn English, you also learn to appreciate the culture of English-speaking countries.
11	People who learn English well have a better future.
	Affect
12	It's fun to learn English.
13	I like it when my teachers speak English.
14	I enjoy understanding song lyrics and video clips in English.
15	I like learning English together with my classmates.
16	*English at school is boring.*
17	*Getting a bad mark for English makes me feel terrible.*
18	I enjoy learning English at school.
	Subjective norm
19	I think learning English well is something I should be expected to do.
20	All junior vocational pupils should learn English well.
21	Everyone in the world should be able to use English to communicate with each other.
22	I would be ashamed later if I couldn't speak English.
23	Pupils should work harder to learn English.
	Perceived behavioral control
24	I'm afraid to speak English in class.
25	I have trouble concentrating in the English lessons.
26	I just think the English language is too hard for me to learn.
27	*If I needed help with English, my parents would try to get me some tutoring.*
28	*At school we have good books and materials for learning English.*
29	I know how to easily look up English words on the Internet or in a dictionary if I don't understand them.
30	I don't have enough time to study hard to learn English.
31	*At school we have useful activities and projects for learning English.*
32	I don't know how to use the Internet to help me learn English.
	Intention
33	In the future I'm going to pay very close attention in English class.
34	I intend to work on my English every single day.
35	*It's enough if I can get by with a bare minimum of English.*
36	I'm going to really pay attention when someone corrects my English mistakes.
37	If I have trouble with English, I will ask the teacher for extra help.
38	I'm going to try to really focus when I read or hear English online.
39	I'm going to make a huge effort to learn English as well as I possibly can.
40	*I'm just too lazy to make an effort to learn English.*

Table 2.6: Results of multilevel analyses for verifying the interactions between bilingual education (BE/CLIL) and year: Does the difference between CLIL and non-CLIL pupils on the scores for MPB constructs become larger in higher years (gm =grand mean centered)?

Different numbers of pupils from 25 classes and 7 schools		***Cognition* model 1**	***Cognition* model 2**	***Affect* model 1**	***Affect* model 2**	***Affect* model 2b**	***Sub. Norm* model 1**	***Sub. Norm* model 2**	***PBC* mod. 1**	***PBC* mod. 2**	***Intention* mod. 1**	***Intention* mod. 2**
N		488	488	488	488	488	486	486	486	486	487	487
Variance	class	.013 (.008)	.012 (.008)	.100 (.040)	.063 (.030)	.063 (.030)	.017 (.014)	.013 (.013)	.013 (.011)	.006 (.009)	.008 (.010)	.006 (.009)
	pupil	.298 (.020)	.298 (.020)	.775 (.051)	.775 (.051)	.775 (.051)	.632 (.042)	.633 (.042)	.472 (.031)	.472 (.031)	.537 (.035)	.537 (.035)
	total	.311	.310	.875	.838	.838	.649	.646	.485	.478	.545	.543
Proportion of explained variance	class				37.0%	37.0%				53.8%		
	pupil				.0%	.0%				.0%		
	total				4.2%	4.2%				1.4%		
	intercept	***3.738*** (.058)	***3.726*** (.065)	***3.887*** (.127)	***3.745*** (.125)	***3.677*** (.196)	***3.638*** (.076)	***3.673*** (.084)	***4.084*** (.067)	***4.010*** (.069)	***3.889*** (.065)	***3.845*** (.072)
Year	1					.068 (.232)						
	2	.044 (.078)	.088 (.101)	-.023 (.172)	.325 (.193)	.393 (.245)	**.289** (.103)	.234 (.130)	-.079 (.089)	.060 (.106)	**-.217** (.087)	-.117 (.111)
	3	-.032 (.090)	-.044 (.122)	-.160 (.201)	-.068 (.232)		.021 (.119)	-.121 (.157)	-.043 (.104)	.096 (.128)	*-.285* (.100)	-.238 (.134)
CLIL stream	Non-CLIL	*-.215* (.069)	-.187 (.098)	*-.497* (.153)	-.164 (.189)	-.409 (.292)	-.089 (.091)	-.169 (.126)	-.147 (.079)	.025 (.103)	*-.276* (.077)	-.179 (.107)
Year*CLIL stream	Non-CLIL* year1					.245 (.348)						
	Non-CLIL* year2		-.107 (.156)		*-.870* (.301)	-.626 (.375)		.055 (.200)		-.**337** (.163)		-.236 (.170)
	Non-CLIL* year3		.025 (.180)		-.245 (.348)			.302 (.229)		-.**311** (.186)		-.105 (.194)
Fit (-2loglikelihood)		809.342	808.722	1291.219	1283.923	1283.923	1166.762	1165.073	1025.095	1020.364	1085.911	1084.045

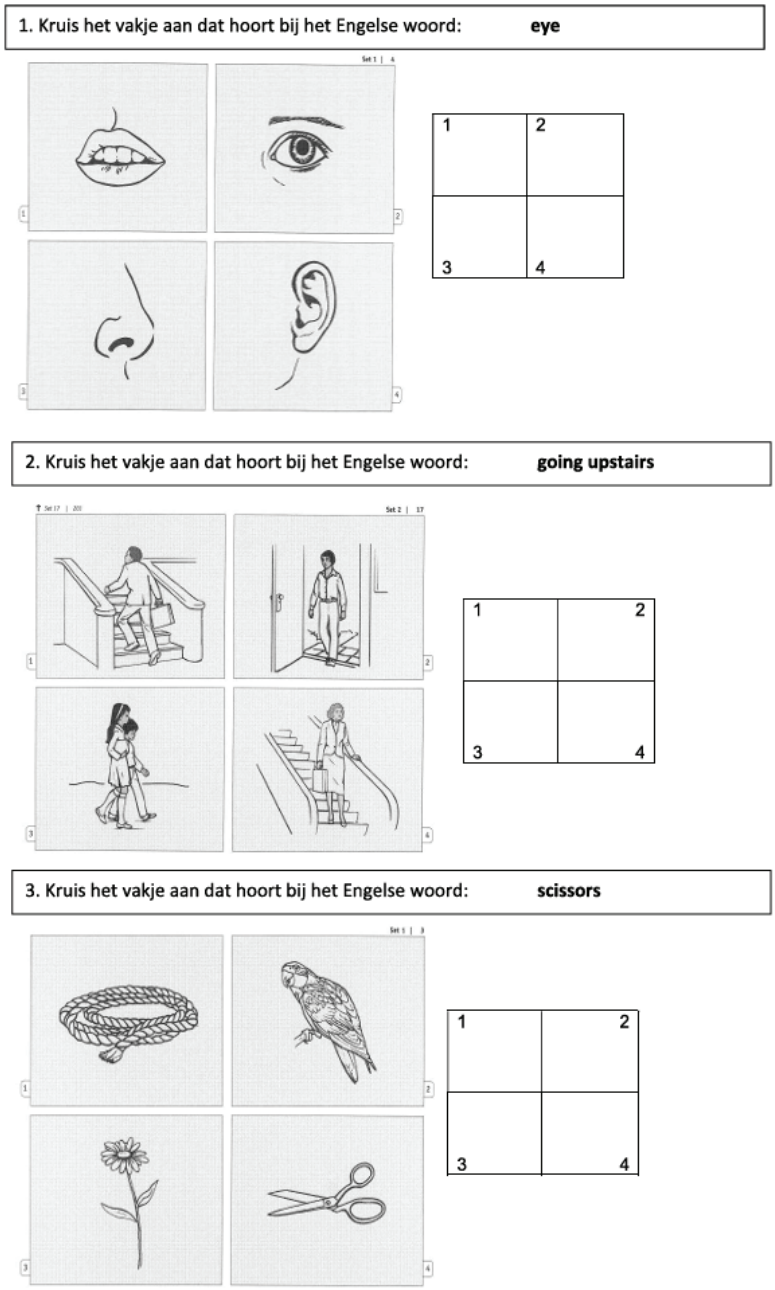
Table 2.6: Results of multilevel analyses for verifying the interactions between bilingual education (BE/CLIL) and year: Does the difference between CLIL and non-CLIL pupils on the scores for MPB constructs become larger in higher years (gm =grand mean centered)? (*continued*)

Different numbers of pupils from 25 classes and 7 schools	***Cognition* model 1**	***Cognition* model 2**	***Affect* model 1**	***Affect* model 2**	***Affect* model 2b**	***Sub. Norm* model 1**	***Sub. Norm* model 2**	***PBC* mod. 1**	***PBC* mod. 2**	***Intention* mod. 1**	***Intention* mod. 2**
Fit improvement compared to model …		*Cognition* model 1		*Affect* model 1	*Affect* model 1		*Sub. Norm* model 1		*PBC* model 1		*Intention* model 1
Difference in -2logl.		.620		7.296	7.296		1.689		4.731		1.866
Difference df		2		2	2		2		2		2
p value		n.s.		p<.05	p<.05		n.s.		p<.10		n.s.

ns = not significant, bold = $p < .05$, italic = $p < .01$, bold and italic = $p < .001$

Appendix 3

Attachment 3.1: **Vocabulary test**. Sample page from the PPVT used to assess receptive vocabulary. Reproduced for illustrative purposes only.



Attachment 3.2: **Lexico-grammar test** (open cloze format) developed for this study.

NAAM_____KLAS_____

Instructies: In dit verhaal over twee oude vrouwen zijn er woorden weggehaald. Schrijf iedere keer één woord dat het beste in de zin past, en goed in het verhaal.



Two sisters and the cat

Sally Wilson and Jane Wilson are old ladies. They ..... sisters.
They live ..... England. They visit ..... other every week.
Sally lives in a house in Beckenham. Her house ..... a big garden.
Jane lives ..... a cat named Sammy in ..... apartment in London.



One day Sally went .....London to visit her sister.
When Jane answered the door, Sally saw tears in her eyes. Jane ..... crying.
"What's ..... matter?" asked Sally. "..... you have a bad trip?"
Jane said, "My cat Sammy died ..... night and I have no place to bury him. What can I do?" She started ..... cry again.

Sally was sad because she knew her sister loved the cat very ..... .
Then Sally said, "I ..... bury your cat in ..... own garden in Danson Park and you can come and put flowers on his grave sometimes."
Then Jane was happy. "..... is a good idea!" she said.
She ..... crying.
"..... you like a cup ..... tea?" asked Sally.
"Then we can ..... a nice chat until it's ..... for you to go home."

At five o'clock Jane thanked Sally and she ..... ready to leave.
She put ..... her hat ..... her coat.
Sally put the dead Sammy into a shopping bag and ..... Jane a big hug.

Jane took the shopping bag and walked ..... it to the bus stop.
She had to wait a ..... time for the bus so she bought a newspaper.
When the bus ..... , she got on. She sat ..... and put the shopping bag on the floor next to ..... feet. She then began to ..... the newspaper.
..... were interesting stories.
She forgot all ..... the shopping bag.



When the bus arrived in Beckenham, Jane got ..... the bus and walked quickly ..... about two minutes.
Suddenly she realized she ..... left the shopping bag on the bus!

Table 3.3: Results of multilevel models for effect CLIL on growth in vocabulary, unstandardized regression coefficients. Standard errors in parentheses. See also complete Table 3.5. (Coeff.= coefficient; SE=standard error; df=degrees of freedom; p=probability; χ^2= Chi-square; Nrepmeas=number of repeated measurements; Npupil;s=number of pupils; Nclasses=number of classes; Nschools=number of schools)

Model	**6b**	**7**
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	23.756 (.977)	23.490 (.985)
Time (in weeks)	.083*** (.004)	.091*** (.005)
non-CLIL (0=clil; 1=non-clil)	-3.984** (1.121)	-3.379* (1.150)
Cohort2 (cohort1=ref.cat.)	4.326** (1.204)	4.459** (1.200)
Cohort3	8.791*** (1.420)	8.783*** (1.416)
pre-vocational level (gm) (1=higher: 5=lower)	-1.625* (.547)	-1.541* (.547)
2-way interactions		
non-CLIL*Time		-.016* (.007)
Time*pre-voc level	-.008** (.003)	-.011*** (.003)
pre-voc*cohort2	.407 (1.182)	.372 (1.179)
pre-voc*cohort3	-1.814 (1.385)	-1.769 (1.381)
non-CLIL*pre-voc	.342 (.789)	.375 (.789)
Time*cohort2	-.021** (.008)	-.023** (.008)
Time*cohort3	-.017* (.008)	-.017* (.008)
non-CLIL*cohort2	-1.231 (1.563)	-1.370 (1.558)
non-CLIL*cohort3	-3.665 (2.263)	-3.680 (2.255)
Random part		
School variance	3.360 (2.124)	3.397 (2.140)
Class variance	.277 (.642)	.258 (.636)
Pupil variance	32.764 (2.305)	32.851 (2.308)
Repeated measures variance	12.001 (.605)	11.908 (.600)
Deviance	8498.430	8492.859
model of reference and fit improvement		model 6b χ^2=5.571 df=1 p<.01
prop. expl. var. school level		-

Table 3.3: Results of multilevel models for effect CLIL on growth in vocabulary, unstandardized regression coefficients. Standard errors in parentheses. See also complete Table 3.5. (Coeff.= coefficient; SE=standard error; df=degrees of freedom; p=probability; χ^2= Chi-square; Nrepmeas=number of repeated measurements; Npupil;s=number of pupils; Nclasses=number of classes; Nschools=number of schools) (*continued*)

Model	**6b**	**7**
prop. expl. var. class level		.069
prop. expl. var. Pupil level		-
prop. expl. var. Rep. meas. level		.008
Sample sizes	Nrepmeas=1377; Npupils=592; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%.

Table 3.4: Results of multilevel models for effect of CLIL on growth in grammar, unstandardized regression coefficients. Standard errors in parentheses. (See also complete Table 3.7.) (Coeff.= coefficient; SE=standard error; df=degrees of freedom; p=probability; χ^2= Chi-square; Nrepmeas=number of repeated measurements; Npupil;s=number of pupils; Nclasses=number of classes; Nschools=number of schools)

Model	**6b**	**7**	**7b**	**8**
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	13.441 (1.101)	13.242 (1.107)	13.407 (1.114)	13.076 (1.135)
Time (in weeks)	.111*** (.004)	.116*** (.005)	.112*** (.004)	.121*** (.006)
non-CLIL (0=clil; 1=non-clil)	-4.011* (1.425)	-3.551* (1.448)	-4.092* (1.408)	-3.028 (1.455)
Cohort2 (cohort1=ref.cat.	5.221* (1.736)	5.316** (1.736)	4.368* (1.729)	4.742* (1.744)
Cohort3	12.367*** (1.937)	12.363*** (1.937)	12.279*** (1.878)	13.582*** (1.919)
pre-vocational level (gm) (1=highest: 5=lowest)	-2.557** (.689)	-2.497** (.690)	-2.297** (.689)	-2.401** (.701)
2-way interactions				
non-CLIL*Time		-.011# (.006)		.008 (.018)
Time*pre-voc level	-.014*** (.003)	-.016*** (.003)	-.019*** (.003)	-.015*** (.004)
pre-voc*cohort2	-1.519 (1.634)	-1.542 (1.634)	-3.227 (1.887)	-3.469# (1.898)
pre-voc*cohort3	-1.264 (1.946)	-1.239 (1.946)	-2.282 (1.901)	-3.199 (1.941)
non-CLIL*pre-voc	1.105 (.956)	1.134 (.956)	.834 (.986)	1.362 (1.017)
Time*cohort2	-.042*** (.007)	-.044*** (.007)	-.084*** (.020)	-.116*** (.025)
Time*cohort3	-.052*** (.008)	-.052*** (.008)	-.139*** (.025)	-.238*** (.043)
non-CLIL*cohort2	-3.911 (2.325)	-4.019 (2.326)	-.609 (3.458)	-1.824 (3.502)
non-CLIL*cohort3	-3.992 (3.237)	-4.029 (3.238)	-4.158 (3.105)	-6.906 (3.211)
3-way interaction				
pre-voc*time*cohort2			.020* (.010)	.028** (.010)
pre-voc*time*cohort3			.034*** (.009)	.058*** (.013)
pre-voc*time*non-CLIL				-.012* (.005)
pre-voc*cohort2*non-CLIL			3.857 (3.298)	3.850 (3.310)
pre-voc*cohort3*non-CLIL			.000 (.000)	.000 (.000)
time *cohort2*non-CLIL				.035* (.015)
time *cohort3*non-CLIL				.072*** (.021)
Random part				
School variance	2.088 (1.870)	2.092 (1.871)	2.726 (2.130)	2.855 (2.203)
Class variance	2.163 (1.485)	2.161 (1.485)	1.711 (1.332)	1.725 (1.339)
Pupil variance	46.893 (3.087)	46.957 (3.089)	46.984 (3.086)	47.312 (3.094)
Repeated measures variance	10.107 (.507)	10.060 (.504)	9.895 (.496)	9.531 (.478)

Table 3.4: Results of multilevel models for effect of CLIL on growth in grammar, unstandardized regression coefficients. Standard errors in parentheses. (See also complete Table 3.7.) (Coeff.= coefficient; SE=standard error; df=degrees of freedom; p=probability; χ^2= Chi-square; Nrepmeas=number of repeated measurements; Npupil;s=number of pupils; Nclasses=number of classes; Nschools=number of schools) (*continued*)

Model	**6b**	**7**	**7b**	**8**
Deviance	8593.094	8589.777	8575.157	8547.269
model of reference and fit improvement		model 6b X^2=3.317 df=1 p<.10		model 7b X^2=27.888 df=4 p<.001
prop. expl. var. school level		-		-
prop. expl. var. class level		.001		-
prop. expl. var. Pupil level		-		-
prop. expl. var. Rep. meas. level		.005		.037
Sample sizes	Nrepmeas=1386; Npupils=591; Nclasses=25; Nschools=7			

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 3.5: Results of longitudinal multilevel models for vocabulary, effects of CLIL. Standard errors in parentheses. All three- and four-way interactions are non-significant and thus not included in the table.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**6a**	**6b**	**7**
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	24.248 (.335)	24.845 (.985)	24.083 (1.406)	25.574 (1.637)	25.461 (1.637)	22.236 (.947)	24.588 (.950)	23.756 (.977)	23.490 (.985)
Time (in weeks)	.073*** (.003)	.075*** (.003)	.075*** (.003)	.075*** (.003)	.079*** (.004)	.075*** (.003)	.075*** (.003)	.083*** (.004)	.091*** (.005)
non-CLIL (0=clil; 1=non-clil)				-4.215* (1.510)	-3.945* (1.525)		-5.177*** (.770)	-3.984** (1.121)	-3.379* (1.150)
Cohort2 (cohort1=ref. cat.)						3.145* (1.477)	2.754** (.754)	4.326** (1.204)	4.459** (1.200)
Cohort3						7.146*** (1.618)	6.718*** (.893)	8.791*** (1.420)	8.783*** (1.416)
pre-vocational level (gm) (1=highest: 5=lowest)						-1.733** (.569)	-1.969*** (.462)	-1.625* (.547)	-1.541* (.547)
2-way interactions									
non-CLIL*Time					-.008 (.006)				-.016* (.007)
Time*pre-voc level								-.008** (.003)	-.011*** (.003)
pre-voc*cohort2								.407 (1.182)	.372 (1.179)
pre-voc*cohort3								-1.814 (1.385)	-1.769 (1.381)
non-CLIL*pre-voc								.342 (.789)	.375 (.789)
Time*cohort2								-.021** (.008)	-.023** (.008)
Time*cohort3								-.017* (.008)	-.017* (.008)
non-CLIL*cohort2								-1.231 (1.563)	-1.370 (1.558)
non-CLIL*cohort3								-3.665 (2.263)	-3.680 (2.255)
Random part									

Table 3.5: Results of longitudinal multilevel models for vocabulary, effects of CLIL. Standard errors in parentheses. All three- and four-way interactions are non-significant and thus not included in the table. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**6a**	**6b**	**7**
School variance			9.211 (7.391)	13.320 (8.609)	13.252 (8.572)	.933 (1.882)	3.980 (2.470)	3.360 (2.124)	3.397 (2.140)
Class variance		22.123 (6.728)	13.032 (4.906)	7.050 (2.965)	7.051 (2.966)	6.522 (2.738)	.337 (.666)	.277 (.642)	.258 (.636)
Pupil variance	52.871 (3.426)	32.865 (2.316)	32.867 (2.316)	33.062 (2.331)	33.096 (2.332)	32.690 (2.309)	32.895 (2.319)	32.764 (2.305)	32.851 (2.308)
Repeated measures variance	12.162 (.614)	12.182 (.615)	12.183 (.615)	12.188 (.614)	12.160 (.613)	12.181 (.614)	12.184 (.614)	12.001 (.605)	11.908 (.600)
Deviance	8740.148	8561.306	8557.426	8552.418	8550.891	8534.619	8515.192	8498.430	8492.859
model of reference and fit improvement		model 1 X^2=178.842 df=1 p<.001	model 2 X^2=3.88 df=1 p<.05	model 3 X^2=5.008 df=1 p<.05	model 4 X^2=1.527 df=1 p=n.s.	model 3 X^2=22.807 df=3 p<.001	model 6 X^2=19.427 df=1 p<.001	model 6a X^2=16.762 df=8 p<.05	model 6b X^2=5.571 df=1 p<.01
prop. expl. var. school level				-		.899	-	.156	-
prop. expl. var. class level				.459		.500	.958	.178	.069
prop. expl. var. Pupil level				-		.005	-	.004	-
prop. expl. var. Rep. meas. level				-		.000	-	.015	.008
Sample sizes			Nrepmeas=1377; Npupils=592; Nclasses=25; Nschools=7						

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 3.6: Results of longitudinal multilevel models for vocabulary, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses.

Model	1	2	3	4	5	6
Fixed part	**Gender zero model**	**Gender with moderator effect**	**Cito (general pre-secondary academic achievement test) zero model**	**Cito with moderator effect**	**Home language zero model**	**Home language with moderator effect**
Intercept	23.248 (1.046)	23.222 (1.049)	23.654 (1.045)	23.650 (1.045)	24.261 (1.042)	24.255 (1.042)
Time (in weeks)	.086*** (.006)	.087*** (.006)	.081*** (.007)	.082*** (.007)	.089*** (.009)	.093*** (.010)
non-CLIL (0=clil; 1=non-clil)	-4.853*** (.989)	-4.791*** (1.009)	-4.220** (1.039)	-4.218** (1.039)	-5.083*** (.823)	-5.061*** (.824)
Cohort2 (cohort1=ref.cat.	3.334** (.861)	3.339** (.860)	3.851** (1.069)	3.858** (1.070)	3.457*** (.795)	3.466*** (.795)
Cohort3	7.208*** (1.006)	7.205*** (1.005)	6.563*** (1.217)	6.548*** (1.220)	7.041*** (.936)	7.067***(.937)
pre-vocational level (gm) (1=highest: 5=lowest)	-1.702** (.500)	-1.701** (.499)	-.271 (.620)	-.264 (.620)	-1.519** (.490)	-1.519** (.490)
Gender (boy=1; girl=0)	1.880* (.735)	1.937* (.760)				
Cito=gm			.224** (.077)	.221** (.078)		
Home language (gm)					-.019 (.270)	.019 (.275)
2-way interactions						
non-CLIL*Time	-.016* (.007)	-.018* (.009)	-.014# (.008)	.160 (.655)	-.016* (.007)	-.016* (.007)
Time*pre-vocational level	-.012*** (.003)	-.012*** (.003)	-.014** (.005)	-.014** (.005)	-.009** (.003)	-.009** (.003)
Time*cohort2	-.025** (.008)	-.025** (.008)	-.008 (.010)	-.008 (.010)	-.021** (.008)	-.021** (.008)
Time*cohort3	-.017* (.008)	-.017* (.008)	-.014 (.010)	-.014 (.010)	-.016* (.008)	-.017* (.008)
Gender*non-CLIL	.199 (1.023)	.066 (1.115)				
Gender*Time	.012* (.006)	.010 (.009)				
Cito*Time			-.001 (.001)	-.001 (.001)		
Cito*non-CLIL			.100 (.111)	.111 (.119)		
Home language*Time					-.000 (.002)	-.001 (.003)
Home language*non-CLIL					.683# (.401)	.562 (.436)

Table 3.6: Results of longitudinal multilevel models for vocabulary, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses. (*continued*)

Model	1	2	3	4	5	6
3-way interaction						
Time*gender*non-CLIL		.004 (.013)				
Time*cito*non-CLIL				-.000 (.001)		
Time*home language*non-CLIL						.003 (.005)
Random part						
School variance	4.107 (2.600)	4.101 (2.596)	3.211 (2.359)	3.207 (2.357)	5.081 (3.060)	5.070 (3.055)
Class variance	.765 (.794)	.762 (.792)	1.158 (1.232)	1.164 (1.234)	.287 (.667)	.293 (.668)
Pupil variance	31.321 (2.217)	31.300 (2.215)	27.252 (2.623)	27.254 (2.623)	32.665 (2.333)	32.624 (2.330)
Repeated measures variance	11.871 (.598)	11.874 (.598)	11.261 (.734)	11.258 (.734)	11.563 (.585)	11.564 (.585)
Deviance	8472.173	8472.083	4863.832	4863.762	8239.895	8239.398
model of reference and fit improvement		model 1 X^2=.090 df=1 p=n.s.		model 3 X^2=.070 df=1 p=n.s.		model 5 X^2=.479 df=1 p=n.s.
N	Nrepmeas=1377; Npupils=592; Nclasses=25; Nschools=7		Nrepmeas=802; Npupils=333; Nclasses=25; Nschools=7		Nrepmeas=1342; Npupils=563; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 3.6 (Continuation): Results of longitudinal multilevel models for vocabulary, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses.

Model	7	8	9	10
Fixed part	**Years Primary English zero model**	**Years Primary English with moderator effect**	**Frequency Primary English zero model**	**Frequency Primary English with moderator effect**
Intercept	23.810 (.973)	23.804 (9.971)	24.017 (.975)	24.016 (.975)
Time (in weeks)	.090*** (.006)	.090*** (.006)	.097*** (.008)	.097*** (.010)
non-CLIL (0=clil; 1=non-clil)	-4.698*** (.805)	-4.687*** (.804)	-4.506*** (.796)	-4.507*** (.797)
Cohort2 (cohort1=ref.cat.	3.965*** (.799)	3.985*** (.799)	3.589*** (.782)	3.589*** (.782)
Cohort3	7.690*** (.939)	7.670*** (.938)	6.931*** (.918)	6.933*** (.918)
pre-vocational level (gm) (1=highest: 5=lowest)	-1.415* (.475)	-1.417* (.474)	-1.651** (.472)	-1.651** (.472)
Years Pri.English (gm)	1.038** (.392)	.947* (.400)		
Frequency Pri.English (gm)			1.261*** (.364)	1.268*** (.372)
2-way interactions				
non-CLIL*Time	-.015* (.007)	-.015* (.007)	-.016* (.007)	-.017* (.014)
Time*pre-vocational level	-.010*** (.003)	-.010*** (.003)	-.010*** (.003)	-.010*** (.003)
Time*cohort2	-.023** (.008)	-.024** (.008)	-.024** (.008)	-.024** (.008)
Time*cohort3	-.018* (.008)	-.018* (.008)	-.018* (.008)	-.018* (.008)
Years Pri.English*Time	-.000 (.004)	.003 (.005)		
Years Pri.English*non-CLIL	.436 (.576)	.722 (.631)		
Frequency Pri.English*Time			-.004 (.003)	-.004 (.004)
Frequency Pri.English*non-CLIL			.595 (.566)	.572 (.619)
3-way interaction				
Time*Years Pri.English*non-CLIL		-.008 (.007)		
Time*Freq.Pri.English*non-CLIL				.001 (.007)
Random part				

Table 3.6 (Continuation): Results of longitudinal multilevel models for vocabulary, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses. (*continued*)

Model	**7**	**8**	**9**	**10**
School variance	4.096 (2.526)	4.070 (2.512)	4.232 (2.592)	4.229 (.975)
Class variance	.334 (.667)	.330 (.666)	.281 (.643)	.282 (.644)
Pupil variance	31.539 (2.271)	31.581 (2.272)	31.141 (2.248)	31.138 (2.248)
Repeated measures variance	11.862 (.599)	11.838 (.598)	11.867 (.600)	11.867 (.600)
Deviance	8262.528	8261.293	8244.099	8244.091
model of reference and fit improvement		model 7 X^2=1.235 df=1 p=n.s.		model 9 X^2=.008 df=1 p=n.s.
N	Nrepmeas=1345; Npupils=563; Nclasses=25; Nschools=7		Nrepmeaus=1343; Npupils=563; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 3.7: Results of longitudinal multilevel models for **grammar**, effects of CLIL and moderator effects of pre-vocational level and grade. Standard errors in parentheses. All four-way interactions are non-significant and thus not included in the table.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**6a**	**6b**	**7**	**7b**	**8**
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	14.750 (.396)	15.544 (1.296)	14.536 (1.815)	16.614 (2.146)	16.611 (2.148)	12.171 (1.182)	15.318 (1.169)	13.441 (1.101)	13.242 (1.107)	13.407 (1.114)	13.076 (1.135)
Time (in weeks)	.089 *** (.003)	.091*** (.003)	.091*** (.003)	.091*** (.003)	.091*** (.004)	.091*** (.003)	.091*** (.003)	.111*** (.004)	.116*** (.005)	.112*** (.004)	.121*** (.006)
non-CLIL (0=clil; 1=non-clil)				-5.983** (2.003)	-5.977** (2.014)		-6.821*** (1.107)	-4.011* (1.425)	-3.551* (1.448)	-4.092* (1.408)	-3.028 (1.455)
Cohort2 (cohort1=ref.cat.)						3.911# (2.002)	3.235* (1.141)	5.221* (1.736)	5.316** (1.736)	4.368* (1.729)	4.742* (1.744)
Cohort3						9.558*** (2.148)	8.697*** (1.320)	12.367*** (1.937)	12.363*** (1.937)	12.279*** (1.878)	13.582*** (1.919)
pre-vocational level (gm) (1=highest: 5=lowest)						-2.452*** (.690)	-2.921*** (.611)	-2.557** (.689)	-2.497** (.690)	-2.297** (.689)	-2.401** (.701)
2-way interactions											
non-CLIL*Time					-.000 (.006)				-.011# (.006)		.008 (.018)
Time*pre-voc level								-.014*** (.003)	-.016*** (.003)	-.019*** (.003)	-.015*** (.004)
pre-voc*cohort2								-1.519 (1.634)	-1.542 (1.634)	-3.227 (1.887)	-3.469# (1.898)
pre-voc*cohort3								-1.264 (1.946)	-1.239 (1.946)	-2.282 (1.901)	-3.199 (1.941)

Table 3.7: Results of longitudinal multilevel models for **grammar**, effects of CLIL and moderator effects of pre-vocational level and grade. Standard errors in parentheses. All four-way interactions are non-significant and thus not included in the table. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**6a**	**6b**	**7**	**7b**	**8**
non-CLIL*pre-voc								1.105 (.956)	1.134 (.956)	.834 (.986)	1.362 (1.017)
Time*cohort2								-.042*** (.007)	-.044*** (.007)	-.084*** (.020)	-.116*** (.025)
Time*cohort3								-.052*** (.008)	-.052*** (.008)	-.139*** (.025)	-.238*** (.043)
non-CLIL*cohort2								-3.911 (2.325)	-4.019 (2.326)	-.609 (3.458)	-1.824 (3.502)
non-CLIL*cohort3								-3.992 (3.237)	-4.029 (3.238)	-4.158 (3.105)	-6.906 (3.211)
3-way interaction											
pre-voc*time*cohort2										.020* (.010)	.028** (.010)
pre-voc*time*cohort3										.034*** (.009)	.058*** (.013)
pre-voc*time*non-CLIL											-.012* (.005)
pre-voc*cohort2*non-CLIL										3.857 (3.298)	3.850 (3.310)
pre-voc*cohort3*non-CLIL										.000 (.000)	.000 (.000)
time *cohort2*non-CLIL											.035* (.015)

Table 3.7: Results of longitudinal multilevel models for **grammar**, effects of CLIL and moderator effects of pre-vocational level and grade. Standard errors in parentheses. All four-way interactions are non-significant and thus not included in the table. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**6a**	**6b**	**7**	**7b**	**8**
time *cohort3*non-CLIL											.072*** (.021)
Random part											
School variance			14.667 (12.460)	22.662 (14.802)	22.660 (14.801)	.000 (.000)	4.565 (3.207)	2.088 (1.870)	2.092 (1.871)	2.726 (2.130)	2.855 (2.203)
Class variance		39.241 (11.799)	24.866 (9.093)	13.596 (5.367)	13.597 (5.368)	13.310 (4.420)	2.332 (1.555)	2.163 (1.485)	2.161 (1.485)	1.711 (1.332)	1.725 (1.339)
Pupil variance	80.619 (5.009)	46.843 (3.113)	46.830 (3.112)	47.003 (3.124)	47.033 (3.124)	46.529 (3.090)	46.662 (3.101)	46.893 (3.087)	46.957 (3.089)	46.984 (3.086)	47.312 (3.094)
Repeated measures variance	10.991 (.551)	11.024 (.0552)	11.025 (.553)	11.028 (.553)	11.028 (.553)	11.023 (.553)	11.026 (.553)	10.107 (.507)	10.060 (.504)	9.895 (.496)	9.531 (.478)
Deviance	8935.292	8715.062	8711.765	8705.895	8705.894	8687.259	8669.219	8593.094	8589.777	8575.157	8547.269
model of reference and fit improvement		model 1 X^2=220.230 df=1 p<.001	model 2 X^2=3.297 df=1 p<.10	model 3 X^2=5.870 df=1 p<.05	model 4 X^2=.001 df=1 p=n.s.	model 3 X^2=24.506 df=3 p<.001	model 6 X^2=18.040 df=1 p<.001	model 6a X^2=76.125 df=8 p<.001	model 6b X^2=3.317 df=1 p<.10	model 6b X^2=17.937df=4 p<.01	model 7b X^2=27.888 df=4 p<.001

Table 3.7: Results of longitudinal multilevel models for **grammar**, effects of CLIL and moderator effects of pre-vocational level and grade. Standard errors in parentheses. All four-way interactions are non-significant and thus not included in the table. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**6a**	**6b**	**7**	**7b**	**8**
prop. expl. var. school level				-		1.000	-	.543	-	-	-
prop. expl. var. class level				.453		.465	.825	.072	.001	.208	-
prop. expl. var. Pupil level				-		.006	-	-	-	-	-
prop. expl. var. Rep. meas. level				-		.000	-	.083	.005	.016	.037
			Nrepmeas=1386; Npupils=591; Nclasses=25; Nschools=7								

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 3.8: Results of longitudinal multilevel models for grammar, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses. All four-way interactions are non-significant and thus not included in the table.

Model	1	2	3	4
Fixed part	**Gender zero model**	**Gender with moderator effect**	**Cito (general pre-secondary academic achievement test) zero model**	**Cito with moderator effect**
Intercept	13.329 (1.186)	13.248 (1.187)	12.621 (1.408)	12.582 (1.410)
Time (in weeks)	.164*** (.014)	.165*** (.014)	.257 (.389)	.342 (.494)
non-CLIL (0=clil; 1=non-clil)	-3.128# (1.588)	-3.011 (1.595)	-3.523 (1.908)	-3.390 (1.911)
Cohort2 (cohort1=ref.cat.	5.244* (1.754)	5.228* (1.752)	6.071* (2.487)	6.124# (2.490)
Cohort3	13.661*** (1.970)	13.665*** (1.968)	13.469** (2.653)	13.502** (2.656)
pre-vocational level (gm) (1=highest: 5=lowest)	-2.519** (.701)	-2.523* (.700)	-1.147 (1.019)	-1.296 (1.022)
Gender (boy=1; girl=0)	-.498 (.866)	-.373 (.884)		
Cito=gm			.176# (.097)	.143 (.098)
2-way interactions				
non-CLIL*Time	.007 (.018)	.005 (.018)	-.034 (.025)	-.009 (.028)
Time*pre-vocational level	-.015*** (.004)	-.015*** (.004)	-.023** (.007)	-.018* (.007)
Time*cohort2	-.116*** (.025)	-.117*** (.025)	-.123*** (.034)	-.120*** (.034)
Time*cohort3	-.241*** (.043)	-.243*** (.043)	-.237*** (.059)	-.235*** (.059)
pre-voc*cohort2	-2.330 (1.669)	-2.344 (1.667)	-2.075 (2.278)	-2.041 (2.280)
pre-voc*cohort3	-3.185 (2.005)	-3.203 (2.003)	-3.973 (2.814)	-3.947 (2.816)
pre-voc*non-CLIL	1.668 (.986)	1.686 (.986)	1.401 (1.365)	1.754 (1.378)
non-CLIL*cohort2	-5.073# (2.372)	-5.031# (2.369)	-2.998 (3.267)	-3.013 (3.274)
non-CLIL*cohort3	-6.900# (3.324)	-6.914# (3.320)	-6.952 (4.593)	-7.150 (4.601)
Gender*non-CLIL	.460 (1.220)	.173 (1.286)		
Gender*Time	-.008 (.006)	-.012 (.008)		

Table 3.8: Results of longitudinal multilevel models for grammar, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses. All four-way interactions are non-significant and thus not included in the table. (*continued*)

Model	**1**	**2**	**3**	**4**
Cito*Time			-.000 (.001)	.001 (.001)
Cito*non-CLIL			.148 (.149)	.245 (.157)
3-way interaction				
Time*pre-voc*cohort2	.029** (.010)	.029** (.010)	.022 (.015)	.021 (.015)
Time* pre-voc*cohort3	.059*** (.013)	.060*** (.013)	.053** (.019)	.052** (.019)
Time*pre-voc*non-CLIL	-.012* (.005)	-.012* (.005)	-.003 (.007)	-.013 (.009)
Time*non-CLIL*cohort2	.036* (.015)	.035* (.015)	.059** (.020)	.059** (.020)
Time*non-CLIL*cohort3	.073*** (.021)	.073*** (.021)	.092** (.028)	.095*** (.028)
Time*gender*non-CLIL		.008 (.011)		
Time*cito*non-CLIL				-.003** (.001)
Random part				
School variance	2.266 (1.958)	2.271 (1.958)	1.268 (2.150)	1.248 (2.147)
Class variance	2.118 (1.473)	2.106 (1.468)	4.747 (2.890)	4.787 (2.902)
Pupil variance	47.340 (3.095)	47.302 (3.093)	44.390 (3.940)	44.394 (3.937)
Repeated measures variance	9.498 (.476)	9.498 (.476)	9.632 (.623)	9.557 (.619)
Deviance	8545.487	8544.997	4968.342	4964.519
model of reference and fit improvement		model 1 X^2=.490 df=1 p=n.s.		model 3 X^2=3.823 df=1 p=n.s.
N	Nrepmeas=1386; Npupils=591; Nclasses=25; Nschools=7	Nrepmeas=1386; Npupils=591; Nclasses=25; Nschools=7	Nrepmeas=809; Npupils=332; Nclasses=25; Nschools=7	Nrepmeas=809; Npupils=332; Nclasses=25; Nschools=7

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant

Table 3.8 (Continuation): Results longitudinal multilevel models for **Grammar**, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses.

Model	**9**	**10**
Fixed part	**Frequency Primary English zero model**	**Frequency Primary English with moderator effect**
Intercept	13.173 (1.131)	13.161 (1.130)
Time (in weeks)	.162*** (.015)	.168*** (.015)
non-CLIL (0=clil; 1=non-clil)	-2.996# (1.487)	-3.007# (1.487)
Cohort2 (cohort1=ref.cat.	4.783* (1.765)	4.793* (1.767)
Cohort3	13.223*** (1.983)	13.216*** (1.985)
pre-vocational level (gm) (1=highest: 5=lowest)	-2.666** (.702)	-2.672* (.702)
Frequency Pri.English (gm)	1.634*** (.421)	1.718*** (.427)
2-way interactions		
non-CLIL*Time	-.000 (.018)	-.013 (.020)
Time*pre-vocational level	-.015*** (.004)	-.015*** (.004)
Time*cohort2	-.116 *** (.025)	-.117*** (.025)
Time*cohort3	-.243*** (.043)	-.240*** (.043)
pre-voc*cohort2	-2.637 (1.664)	-2.643 (1.665)
pre-voc*cohort3	-2.815 (2.001)	-2.797 (2.003)
pre-voc*non-CLIL	1.574 (.985)	1.607 (.985)
non-CLIL*cohort2	-4.651# (2.375)	-4.679 (2.378)
non-CLIL*cohort3	-6.441 (3.350)	-6.370 (3.353)
Frequency Pri.English*Time	.000 (.003)	-.003 (.004)
Frequency Pri.English*non-CLIL	.635 (.664)	.325 (.703)
3-way interaction		
Time*pre-voc*cohort2	.028** (.010)	.028** (.010)
Time*pre-voc*cohort3	.059*** (.013)	.058*** (.013)

Table 3.8 (Continuation): Results longitudinal multilevel models for **Grammar**, moderator effects of pupil characteristics on the effect of CLIL. Standard errors in parentheses. (*continued*)

Model	**9**	**10**
Time*pre-voc*non-CLIL	-.010* (.005)	-.011* (.005)
Time*non-CLIL*cohort2	.039* (.015)	.040** (.015)
Time*non-CLIL*cohort3	.078*** (.021)	.076*** (.021)
Time*Freq.Pri.English*non-CLIL		.008 (.006)
Random part		
School variance	2.269 (1.957)	2.250 (1.945)
Class variance	2.081 (1.464)	2.096 (1.460)
Pupil variance	45.247 (3.036)	45.205 (3.024)
Repeated measures variance	9.434 (.475)	9.419 (.475)
Deviance	8283.171	8281.384
model of reference and fit improvement		model 9 X^2=1.787 df=1 p=n.s.
N	Nrepmeas=1351; Npupils=562; Nclasses=25; Nschools=7	Nrepmeas=1351; Npupils=562; Nclasses=25; Nschools=7

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Appendix 4

Attachment 4.1: **Writing test** *(in Dutch so all pupils understand the prompts, and not copy the English words)*

NAAM _____ KLAS _____

Schrijftaak: Je klas heeft contact met een school in Engeland. De Engelse leerlingen komen binnenkort bij jullie op bezoek. Je schrijft een brief aan een van de Engelse leerlingen, je mag de naam zelf verzinnen. Je schrijft natuurlijk in het Engels!

- zeg hoe je heet en hoe oud je bent (met woorden, geen cijfer!)
- zeg dat je het leuk vindt dat de klas naar Nederland komt
- zeg dat je een beetje Engels spreekt
- zeg dat jullie op school geen schooluniformen dragen
- zeg dat je veel vragen hebt
- vraag of hij (of zij) kan fietsen
- vraag wat zijn (of haar) lievelings eten is
- vraag of hij (of zij) Facebook heeft
- (je mag daarna ook andere dingen vragen of vertellen als je wilt)

Dear _____ ,

NAME _____ ***CLASS*** _____

[Writing task: Your class has contact with a school in England. The English pupils are coming to visit you soon. You write a letter to one of the English pupils; you may make up a name. Of course you will write in English!

- *tell your name and how old you are (in words, not a number!)*
- *say that you think it's fun that the class is coming to the Netherlands*
- *say that you can speak a bit of English*
- *say that you don't wear uniforms at your school*
- *say that you have a lot of questions*
- *ask if he (or she) can ride a bicycle*
- *ask what his or her favorite food is*
- *ask if he (or she) is on Facebook*
- *(after this you can ask or tell other things if you want to)]*

Attachment 4.2: Sample Speaking test (recorded and transcribed)

Speaking Test (3 min. total) (Intro in Dutch: This is not for a school grade, and it's OK to make mistakes. Try to speak in complete sentences and talk as much as you can. All right?)

Part 1: General conversation (ca. 2 min.) – one question from each prompt group, or a further question if communication stalls.

(Present) *What's your name? What class are you in? How old are you?*

Tell me a little bit about yourself.

Prompts:
Tell me about where you live.
Tell me about your family.
Tell me about your school.
Tell me about your hobbies.
Tell me about your friends.
What you like to do with your friends?
What do you like to do on the weekend?

(Past) *What did you do last weekend?*
What did you do last summer?
What did you do yesterday evening?
What did you do on your last vacation?

(Future) *Tell me something you would like to do or try in the future/ What would you like to do in the future?*

Part 2: Picture task (ca. 1 minute)
Here are pictures of two different places. I'd like you to describe the pictures and tell me which place you would like to visit, and why. (Repeat in Dutch if necessary.)





Thank you. That's the end of your speaking test.

Attachment 4.3: Writing assessment rubric

Categories	0	1	2	3	4	5	6
Sentences/ Syntax	(Almost) no sentences.	A few sentences. Fragments. Syntax errors impede comprehension.	Simple sentence structure, (S – V or S-V-DO) Errors distract from meaning. Word order problems.	Some variety in sentence structure. Basic structural patterns are clear. Errors may distract from meaning. Some word order problems.	Control of most basic structures and most word order. Some correct use of adverbials, some complex or compound structures attempted.	Attempts a variety of structural patterns. May make some errors in complex structures and tenses.	Syntactic variety. Few or no syntactical errors despite complex constructions (dependent clauses, linking words).
Grammar	(Almost) no grammar.	Riddled with grammar errors which impede comprehension. Use of different tenses is absent or inaccurate.	Simple grammar, with frequent errors which sometimes distract from meaning. Tense use may be attempted but not accurate.	Some control of basic grammar and tense (simple present, simple past), though with many errors.	More ambitious grammar and tense, though with some errors. Attempt to use present, past and future tenses.	Variety of grammar and tense, with few errors. Generally accurate use of present, past and future tenses.	Grammatical variety and high degree of control; (almost) no errors. Accurate tense use (present, past, future)
Spelling	Most words misspelled.	Many words misspelled and meaning is hard to discern.	Many words misspelled, but meaning is mostly clear.	Some misspellings but meaning is almost always clear.	A few misspellings and meaning is not affected at all.	Very few misspellings, or only for challenging words.	Almost no misspellings, even with more challenging words.
Task (length and completeness)	Clearly insufficient. Hardly any task parts addressed.	Very few task parts addressed.	Some task parts addressed successfully.	Completes quite a few of the task parts.	Completes many of the task parts with some elaboration.	Completes all or nearly all of the task parts and goes beyond the task.	Goes far beyond the limitations of the task, with some extra contributions.
Dutch interference	Consistent Dutch interference; would hardly be understood by a non-Dutch speaker.	Frequent Dutch interference, with some words or phrases not comprehensible to a non-Dutch speaker.	Some Dutch interference, mostly comprehensible. Interference errors are typical for low-level learners.	Some Dutch interference, and some low-level interference errors may be avoided.	A bit of Dutch interference, but completely comprehensible.	Almost no Dutch interference, or only interference with very challenging words or structures.	No Dutch interference, and typical interference errors are avoided.

Attachment 4.3: Writing assessment rubric (*continued*)

Categories	0	1	2	3	4	5	6
Vocabulary	Very simple vocabulary. Depends highly on Dutch cognates or Dutch words.	Very simple vocabulary, lots of repetition, but some words that are not related to Dutch.	Enough English vocabulary not to be consistently dependent on Dutch. A few words or phrases that go beyond absolute beginner level.	Vocabulary meets the task requirements in general, perhaps missing vocab. for a few prompts. A few good lexical items, phrases and/or collocations.	Mostly accurate vocabulary for the task, with nearly all task prompts addressed. Some phrasal verbs, fixed expressions, collocations and/or more ambitious vocabulary attempted.	Accurate vocabulary for the task, with nearly all task prompts addressed. Phrasal verbs, fixed expressions, collocations and/or more ambitious vocabulary attempted.	Very accurate vocabulary, with (occasional) skillful use of more ambitious lexical items.
Clarity	Frequently unclear what is meant.	Unclear what is meant in a few instances.	Generally clear, with perhaps one unclear item.	Generally clear and showing some small degree of organization (similar subjects together, punctuation, paragraphing).	Clear and with some organization.	Very clear and generally organized.	Completely clear and well-organized.

Adapted from *Cambridge English: Key for Schools – Teacher's handbook* (Cambridge ESOL, 2012) and Verspoor, (personal communication, August 27, 2012).

Attachment 4.4: Speaking assessment rubric

Categories	0	1	2	3	4	5	6
Sentences/ Syntax	(Almost) no sentences or control over word order.	A few sentences. Fragments. Syntax errors impede comprehension.	Simple sentence structure, (S – V or S-V-DO) Errors distract from meaning. Word order problems.	Some variety in sentence structure. Basic structural patterns are clear. Errors may distract from meaning. Some word order problems.	Control of most basic structures and most word order. Some correct use of adverbials, some complex or compound structures attempted.	Attempts a variety of structural patterns. Clauses are linked.	Syntactic variety. Few or no syntactical errors despite complex constructions.
Grammar	(Almost) no grammar.	Riddled with grammar errors which impede comprehension.	Simple grammar, with frequent errors which sometimes distract from meaning.	Some control of basic grammar and tense (simple present, some past or future tense), though with many errors.	More ambitious grammar and tense, though with some errors. Attempted use of present, past and future tenses.	Variety of grammar and tense, with (nearly) accurate use of present, past and future tenses.	Grammatical variety and high degree of control; (almost) no errors. Present, past and future tenses used correctly.
Task (length and completeness)	Clearly insufficient. Hardly any task parts addressed. Needs help understanding the questions.	Very few task parts addressed. May need help understanding the questions.	Some task parts addressed successfully. Able to answer the questions, at least with a minimum sufficient amount of information.	Completes quite a few of the task parts. Develops each answer at least a little bit (more than just one short answer)	Completes the task parts with some elaboration; no trouble adding more information than the minimum.	Completes all of the task parts and goes beyond the task by giving additional information.	Goes far beyond the limitations of the task, with some extra contributions; develops each idea.
Dutch interference	Consistent Dutch interference; would not, or hardly, be understood by a non-Dutch speaker.	Frequent Dutch interference, with some words or phrases not comprehensible to a non-Dutch speaker.	Some Dutch interference, mostly comprehensible. Interference errors are typical for low-level learners.	Some Dutch interference, but some low-level interference errors are avoided.	A bit of Dutch interference, but completely comprehensible.	Almost no Dutch interference, or only interference with very challenging words or structures.	No Dutch interference, and typical interference errors are avoided.

Attachment 4.4: Speaking assessment rubric (*continued*)

Categories	0	1	2	3	4	5	6
Vocabulary	Very simple vocabulary. Depends highly on Dutch cognates or Dutch words.	Very simple vocabulary, lots of repetition, but frequent use of Dutch or words related to Dutch.	Enough English vocabulary not to be consistently dependent on Dutch. A few words or phrases that go beyond absolute beginner level.	Vocabulary meets the task requirements in general, perhaps missing vocab. for a few prompts. A few good lexical items, phrases and/or collocations.	Mostly accurate vocabulary for the task, with nearly all task prompts addressed. Phrasal verbs, fixed expressions, collocations and/or more ambitious vocabulary attempted.	Accurate vocabulary for the task, with nearly all task prompts addressed. Phrasal verbs, fixed expressions, collocations and/or more ambitious vocabulary attempted.	Very accurate vocabulary, with (occasional) skillful use of more ambitious lexical items.
Clarity	Unclear what is meant in most of the utterances.	Unclear what is meant in more than a few utterances.	Unclear in a few utterances.	Generally clear and showing some control over thought-to-language.	Clear and with some control.	Very clear and controlled.	Completely clear and completely under control.

Adapted from *Cambridge English: Key for Schools – Teacher's handbook* (Cambridge ESOL, 2012) and the Student Oral Proficiency Assessment Rating Scale (SOPA) (Thompson et al., 2002).

Attachment 4.5: **WtC questionnaire items** (translated from Dutch)

Would you dare to do these things **in English**?
Choose from: Yes, definitely! / Yes, probably / Maybe / Probably not / Definitely not!

Willingness to Communicate at school (inside the classroom/school)

1. Ask to be excused to go to the toilet
2. Ask the teacher to repeat something
3. Read aloud from a book in class
4. Tell what you did last weekend
5. Ask questions about a film or a story
6. Lead a group discussion
7. Give a tour of your school to a new English teacher

Willingness to Communicate in London (outside the classroom/school)

1. Order a meal in a restaurant
2. Buy something in a shop
3. Ask for help when you get lost
4. Talk to strangers in the bus
5. Ask questions during a city tour
6. Talk on the phone to get information about a fun activity
7. Give a presentation at an English school

Attachment 4.6: Sample writing anchor texts for holistic proficiency assessment (scores in unanimous agreement by all assessors in the assessment group)

WRITING ANCHOR TEXTS (selection)

Score: 0 (cohort 1, non-CLIL, 43 English words)
hy X im twelf jers old and itst verry nace dat you hive camt iam larna thrie jer Engels O als you and yor klas hir komt an gaan wi naaar trird Fat (dats is a atracie park) ant mij clas ant i hebben veel vragen wer you dr
ben you ooit in nederland geweest
can you bike
wats your favrit fout
have you facbook
iam wil you ont moet

Score: 1 (cohort 1, CLIL, 86 English words)
Hello I am X.
I am tirteen years old.
I learn al tree years inglish.
We goo to the zoo.
I have a lot of questions.
Hav yuw live in the nederland.
Can yuw fietsen.
wat is yor faforite food.
Has youw facebook.
I like computer games.
Mij sport is jocking.
I have 1 little brother and 1 little sister. I have divorcet parents.
I go to vakantie on a camping in Epe.
I go in a afgelope weekend to my grandmother.
I will later work at the zoo.
I like to sy youw.

Score: 2 (cohort 1, CLIL, 114 English words)
I'm x and Im twelve years old.
I'm good in English and I learning that al three years. When you and the class are hear than we going to a musea. I have a lot of questions to asked but not know.
Have you al bien in Holland?
Can you cycling? What is your favorit food?

Have you Facebook because my class to that always. I love playing football and playing outside with my frends.
I do dancing and diving every week.
My family is verey nice.
I have one sister and one liddle brother.
There names are x and x.
x is sixteen years old and x is nine years old. I whant to are a Actrice.
I whant to see you.

Score: 3 (cohort 3, non-CLIL, 134 English words)
My name is X and I'm fiveteen years old. I'm learning now four years English. If you're coming with your class we are going to watch a movie in our classroom. I have a lot of questions for you and the other boys and girls in your class. Do you even come here in Holland? And can you ciceling on a bike? What is your favorite food? Do you have facebook? or twitter maybe?

I'm going every Saterday to my work. I work in SUS there I help in the administration. I love it. SUS is a soccer (vereniging). My dad is there a coach and I have a brother who plays soccer! And if I don't work than I'm going and the mall with my mother. I love shopping! Now I can't wait to see you and your class!

Score: 4 (cohort 2, CLIL, 210 English words)
I'm fourteen years old and my name is X. I started learning English 3 years ago. If you arrive then first we go to my house the next day we go to Den Haag there we eat a ice cream and then we goes back to school for some lessons. After the lessons we go to my home and then in the evening we eat pancakes. I have many questions for you about England. But first I ask something to you. Do you was ever in Holland. Do you now that are many bikes in our country? Can you cycle? I hope your favorite meal is pancakes because we gonna eat that. What is your favorite meal? Do you have Facebook? We can chat if you have that is much easier than writing a letter you know. Than you can read what all my hobbies are and things like that. Here is a preview from my hobbies and things: My hobby is my cockatiel his name is Elmo. And I have 2 sisters and 1 brother I'm the oldest child at home. Now I'm training for my work later because that is working by the korps mariniers and than you have to have a good condition and much muscle's so you have to be strong.
I realy wanna meat you

Attachment 4.7: Sample speaking anchor texts for holistic proficiency assessment (scores in unanimous agreement by all assessors in the assessment group)

SPEAKING ANCHOR TEXTS (selection)

Score: 0 (cohort 1, non-CLIL, 37 English words)

My name is X, een B.
I am twelve year old and I'm my hobbies zijn, football, badminton, tennis, mijn familie woont in marokko, My friends are X, X, X.
We ging altijd football, zwemming.
School is good, is ja ik weet niet, is gewoon een goeie school.
swimming, naar familie, naar vrienden, naar de straat, naar children disco.
Ook wat je wilt worden?
Ik wil techniek en ik wil I have weg, uitbergen.
Good footballer.
Two, avontuur en cool sport voor men ja, and I need some hobby ook.

Score: 1 (cohort 3, non-CLIL, 78 English words)

X. 3M2. I'm fourteen years old, I live in X, I play soccer, fourteen years. And that's it.
I have a mother, brother, sister, and lots of uncles and aunts and I have no pets.
I stay in Holland and go play with my friends.
As job, as job I want in the, ja hoe zeg je dat nou weer. Hoe zeg je dat in het Engels. In de haven werken. And transport, that will I do.
This, because this is a different culture of in Holland and this, it can be in Holland. This will be fun.

Score: 2 (cohort 2, CLIL, 149 English words)

My name is X. I'm in 2B. I have two little sisters, one big sister and one big brother. I don't have a father anymore and I have a mother, that's just all it. I really like to dance, ja, that was it. Ja, I like to be creative with, ja, clay and, ja, that stuff. Ja, that was it.
Wat deed ik? Toen was ik. I was at home I think. I don't know anymore. I think just sleep everywhere, by my aunt and nieces and I think that, ja.
I don't really know. Just, I think, ja. No, I don't know.
The city, because I don't really like the nature. It is really beautiful, but I don't really like it. Just on pictures, that's all. So I would choose the city, because it's, ja, I have been there all my life, so I know what's in there, and I know what, how to do and I know what to expect in there.

Score: 3 (cohort 2, non-CLIL, 107 English words)

My name is X. I am in TwoE. I am thirteen years old and I live in X. I have one brother. His name is X and my mother X is forty-four years old. And my dad is X and he is seventy-four years old [sic]. My sport is mountain bike and snowboarding and ja dat is het.

I like it very much here because it's, ik weet niet wat gezellig is. Fun. Just talk and chill and check the email.

Last night? I made homework and I watch tv and I go on the computer. I want to become a copper later. A copper, you know?

This one because I never seen that one in real life and this, yeah, I like this more.

Score: 4 (cohort 3, CLIL, 263 English words)

I'm fourteen years old, I lived on X for three years, on an American school so that's why I chose TTO. I like to sing, my mother is sing teacher so I'm singing like since my six, since I was six.

The girl that was here net is my best friend and I have another friend who was in the first class with me, this year but she had to go a level lower so she's not on this school any more. I have friends in Rotterdam and friends all over, actually.

I was outside with a girlfriends, with my sister, the girlfriend of my sister, my friends and another friend of me. We ate with each other and then we went to Cookers, like a restaurant to get some ice. I think I was home like nine o'clock or something.

I like to be, I like to do something with music because I like singing. I like to go to Codarts? A school in Rotterdam. Last year I did audition and I was, I was allowed to go there so I did audition and they accept me. But my level was too low so now this year I'm going to try the same thing. I hope that this time I can go.

Definitely this one. I like forests, I like to walk, I like to do activities. Every summer I go to, I do active things, in the mountains, climbing. Because my mother, she's also doing that. I just really like nature. This would also be fun, I think but I've never done it so I don't know.

Table 4.4: Inter Rater Reliability, writing proficiency 0-measurement: Correlations parametric (r) and non-parametric (r_s); r = Pearson correlation (below the diagonal); r_s = Spearman's rho (above the diagonal).

		Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**	**Rater 6**
Rater 1	r/r_s	1	.853	.792	.831	.829	.763
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	175	56	50	62	60	55
Rater 2	r/r_s	.828	1	.801	.910	.848	.809
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	56	174	47	65	58	58
Rater 3	r/r_s	.785	.818	1	.876	.803	.742
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	50	47	176	63	59	67
Rater 4	r/r_s	.810	.891	.882	1	.842	.760
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	62	65	63	198	60	67
Rater 5	r/r_s	.785	.808	.754	.826	1	.788
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	60	58	59	60	187	60
Rater 6	r/r_s	.760	.810	.754	.765	.768	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	55	58	67	67	60	195

Table 4.5: Inter Rater Reliability, writing proficiency 1-measurement: Correlations parametric (r) and non-parametric (r_s); r = Pearson correlation (below the diagonal); r_s = Spearman's rho (above the diagonal)

		Rater 7	Rater 8	Rater 9	Rater 10	Rater 11	Rater 12	Rater 13
Rater 7	r/r_s	1	.751	.779	.818	.521	.855	.776
	Sig. (2-tailed)		.000	.000	.000	.005	.000	.000
	N	161	34	30	25	27	30	161
Rater 8	r/r_s	.794	1	.743	.804	.816	.719	.824
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
	N	34	148	27	31	23	25	148
Rater 9	r/r_s	.818	.700	1	.812	.923	.664	.798
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
	N	30	27	149	23	23	30	149
Rater 10	r/r_s	.895	.823	.789	1	.843	.832	.825
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
	N	25	31	23	154	33	32	154
Rater 11	r/r_s	.531	.869	.930	.788	1	.902	.869
	Sig. (2-tailed)	.004	.000	.000	.000		.000	.000
	N	27	23	23	33	132	25	132
Rater 12	r/r_s	.840	.638	.678	.838	.919	1	.861
	Sig. (2-tailed)	.000	.001	.000	.000	.000		.000
	N	30	25	30	32	25	158	158
Rater 13	r/r_s	.818	.825	.813	.839	.889	.870	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	161	148	149	154	132	158	487

Table 4.6: Inter Rater Reliability, writing proficiency 2-measurement: Correlations parametric (r) and non-parametric (r_s); r = Pearson correlation (below the diagonal); r_s = Spearman's rho (above the diagonal)

		Rater 14	Rater 15	Rater 16	Rater 17	Rater 18	Rater 19	Rater 20	Rater 21
Rater 14	r/r_s	1	.931	.812	.880	.871	.797	.873	.879
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
	N	118	20	26	21	26	23	17	118
Rater 15	r/r_s	.897	1	.740	.885	.912	.904	.981	.901
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
	N	20	124	26	25	29	27	12	124
Rater 16	r/r_s	.832	.806	1	.915	.998	.771	.903	.911
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
	N	26	26	120	23	23	20	18	120
Rater 17	r/r_s	.908	.896	.906	1	.886	.875	.933	.937
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
	N	21	25	23	117	21	21	23	117
Rater 18	r/r_s	.871	.910	.992	.876	1	.850	.869	.936
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
	N	26	29	23	21	119	21	14	119
Rater 19	r/r_s	.848	.869	.826	.905	.905	1	.820	.869
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
	N	23	27	20	21	21	112	16	112
Rater 20	r/r_s	.907	.977	.942	.935	.881	.844	1	.924
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
	N	17	12	18	23	14	16	84	84
Rater 21	r/r_s	.891	.900	.933	.938	.937	.884	.939	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	
	N	118	124	120	117	119	112	84	389

Table 4.7: Inter Rater Reliability: speaking proficiency 0-measurement: Correlations parametric (r) and non-parametric (r_s); r = Pearson correlation (below the diagonal); r_s = Spearman's rho (above the diagonal)

		Rater 22	Rater 23	Rater 24	Rater 25	Rater 26	Rater 27
Rater 22	r/r_s	1	.745	.831	.820	.732	.715
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	149	58	65	74	69	55
Rater 23	r/r_s	.723	1	.735	.817	.750	.707
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	58	152	60	83	58	67
Rater 24	r/r_s	.842	.749	1	.898	.772	.864
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	65	60	160	81	68	70
Rater 25	r/r_s	.844	.831	.876	1	.769	.855
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	74	83	81	303	88	82
Rater 26	r/r_s	.752	.774	.783	.822	1	.745
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	69	58	68	88	165	66
Rater 27	r/r_s	.730	.701	.847	.844	.732	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	55	67	70	82	66	157

Table 4.8: Inter Rater Reliability: speaking proficiency 2-measurement: Correlations parametric (r) and non-parametric (r_s); r = Pearson correlation (below the diagonal); r_s = Spearman's rho (above the diagonal)

		Rater 28	Rater 29	Rater 30	Rater 31	Rater 32	Rater 33	Rater 34	Rater 35
Rater 28	r/r_s	1	.904	.878	.915	.863	.970	.938	.962
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
	N	91	22	23	17	19	18	15	19
Rater 29	r/r_s	.892	1	.915	.901	.720	.877	.826	.929
	Sig. (2-tailed)	.000		.000	.000	.002	,000	.000	.000
	N	22	93	20	20	16	19	19	21
Rater 30	r/r_s	.886	.905	1	.848	.934	.931	.899	.930
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
	N	23	20	98	20	16	19	22	22
Rater 31	r/r_s	.914	.897	.924	1	.966	.861	.950	.913
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
	N	17	20	20	86	14	18	19	21
Rater 32	r/r_s	.876	.796	.943	.963	1	.876	.899	.943
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
	N	19	16	16	14	70	15	15	19
Rater 33	r/r_s	.961	.883	.926	.926	.862	1	.901	.872
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
	N	18	19	19	18	15	91	24	22
Rater 34	r/r_s	.940	.860	.896	.951	.900	.894	1	.887
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
	N	15	19	22	19	15	24	89	19
Rater 35	r/r_s	.937	.911	.935	.898	.928	.872	.896	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	
	N	19	21	22	21	19	22	19	97

Table 4.9: Results longitudinal multilevel models for writing proficiency (z-scores), effects of CLIL. Standard errors in parentheses.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	-.104 (.039)	.043 (.131)	-.143 (.182)	.088 (.227)	.075 (.227)	-.045 (.125)	-.157 (.121)	-.209 (.123)	-.228 (.124)	-.228 (.124)
Time (in weeks)	.002***[1] (.000)	.002*** (.000)	.002*** (.000)	.002*** (.000)	.003*** (.001)	.002* (.000)	.004*** (.001)	.005*** (.001)	.006*** (.001)	.006*** (.001)
non-CLIL (0=clil; 1=non-clil)				-.677** (.197)	-.647** (.199)	-.558*** (.143)	-.383* (.169)	-.265 (.173)	-.212 (.176)	-.212 (.176)
Grade 8 (grade 7=ref.cat.)						.139 (.171)	.268 (.219)	.292 (.219)	.238 (.230)	.244 (.231)
Grade 9 (grade 7=ref.cat.)						.804*** (.184)	1.236*** (.232)	1.249*** (.232)	1.260*** (.176)	1.259* (.176)
Pre-voc level (gm) (1=high: 5=low)						-.309*** (.064)	-.279** (.078)	-.265** (.078)	-.256* (.080)	-.257 (.080)
2-way interactions										
non-CLIL*Time					-.001 (.001)			-.003*** (.001)	-.002 (.003)	-.002 (.003)
Time*pre-voc-level-gm							-.002*** (.000)	-.002*** (.000)	-.003*** (.001)	-.002* (.001)
pre-voc level*grade 8							-.164 (.194)	-.168 (.194)	-.356 (.234)	-.344 (.237)
pre-voc level*grade 9							-.458# (.233)	-.453# (.233)	-.506# (.243)	-.505 (.243)
non-CLIL*pre-voc level							.199# (.109)	.204# (.109)	.225 (.117)	.227 (.117)

Table 4.9: Results longitudinal multilevel models for writing proficiency (z-scores), effects of CLIL. Standard errors in parentheses. (*continued*)

Model	1	2	3	4	5	6	7	8	9	10
Time*grade 8							-.004*** (.001)	-.005*** (.001)	-.015*** (.004)	-.014*** (.004)
Time*grade 9							-.003** (.001)	-.004*** (.001)	-.008 (.006)	-.008 (.006)
non-CLIL*grade 8							-.224 (.306)	-.254 (.307)	-.276 (.470)	-.302 (.477)
non-CLIL*grade9							-.696 (.408)	-.728 (.409)	-.753 (.424)	-.751 (.424)
3-way interactions										
pre-voc level*time*grade 8									.004*** (.001)	.004* (.002)
pre-voc level*time*grade 9									.002 (.002)	.002 (.002)
pre-voc level*time*non-CLIL									-.001 (.001)	-.001 (.001)
pre-voc level*grade 8*non-CLIL									.167 (.426)	.133 (.438)
pre-voc level*grade 9*non-CLIL									.000 (.000)	.000 (.000)
time *grade 8*non-CLIL									.005* (.002)	.003 (.006)
time *grade 9*non-CLIL									-.001 (.003)	-.001 (.003)

Table 4.9: Results longitudinal multilevel models for writing proficiency (z-scores), effects of CLIL. Standard errors in parentheses. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
4-way interactions										
pre-voc level*time*grade 8*non-CLIL										.001 (.003)
pre-voc level*time*grade 9*non-CLIL										.000 (.000)
Random part										
School variance			.143 (.125)	.269 (.169)	.269 (.168)	.006 (.022)	.000 (.000)	.000 (.000)	.000 (.000)	.000 (.000)
Class variance		.406 (.121)	.266 (.095)	.132 (.051)	.132 (.051)	.091 (.036)	.062 (.023)	.062 (.023)	.062 (.023)	.062 (.023)
Pupil variance	.655 (.045)	.324 (.026)	.324 (.026)	.325 (.026)	.326 (.026)	.323 (.026)	.327 (.026	.331 (.026)	.333 (.026)	.333 (.026)
Repeated measures variance	.223 (.011)	.224 (.011)	.224 (.011)	.224 (.011)	.224 (.011)	.224 (.011)	.214 (.011)	.210 (.011)	.206 (.010)	.206 (.010)
Deviance	3016.219	2759.430	2756.472	2749.270	2748.195	2727.174	2678.909	2669.094	2653.101	2652.991
Model of reference and fit improvement		model 1 X^2=256.789 df=1 p<.001	model 2 X^2=2.958 df=1 p<.10	model 3 X^2=7.202 df=1 p<.01	model 4 X^2=1.075 df=1 p=n.s.	model 4 X^2=22.096 df=3 p<.001	model 6 X^2=48.265 df=8 p<.001	model 7 X^2=9.815 df=1 p<.01	model 8 X^2=15.993 df=7 p<.05	model 9 X^2=.110 df=2 p=n.s.
prop. expl. var. school level				-		.978	1.000	-	-	

Table 4.9: Results longitudinal multilevel models for writing proficiency (z-scores), effects of CLIL. Standard errors in parentheses. (*continued*)

Model	1	2	3	4	5	6	7	8	9	10
prop. expl. var. class level				.504		.311	.319	-	-	
prop. expl. var. Pupil level				-		.006	-	-	-	
prop. expl. var. Rep. meas. level				-		-	.045	.019	.019	
	Nrepmeas=1367; Npupils=594; Nclasses=25; Nschools=7									

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.10: Results of longitudinal multilevel models for writing fluency (number of words in English), effects of CLIL. Standard errors in parentheses.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	97.322 (2.128)	99.352 (5.713)	95.606 (7.607)	107.135 (6.791)	106.424 (6.837)	105.263 (5.989)	102.360 (6.359)	99.734 (6.457)	96.828 (6.567)	96.754 (6.565)
Time (in weeks)	.267*** (.029)	.294*** (.029)	.295*** (.029)	.294*** (.029)	.315*** (.039)	.296*** (.029)	.420*** (.038)	.491*** (.048)	.571*** (.056)	.574*** (.056)
non-CLIL (0=clil; 1=non-clil)				-19.174# (10.362)	-17.557 (10.542)	-28.579*** (6.855)	-28.111** (8.747)	-22.241* (9.093)	-17.759 (9.337)	-17.018 (9.332)
Grade 8 (grade 7=ref. cat.)						1.419 (8.384)	10.409 (11.404)	11.577 (11.421)	12.417 (12.162)	11.866 (12.212)
Grade 9						21.843* (8.894)	35.224* (12.264)	35.902* (12.276)	44.010* (12.759)	44.082 (12.754)
Pre-voc level (gm) (1=high: 5=low)						-14.430*** (3.013)	-15.678** (4.076)	-15.004** (4.088)	-12.953# (4.282)	-12.829 (4.288)
2-way interactions										
non-CLIL*Time					-.046 (.058)			-.140* (.059)	-.423* (.171)	-.442* (.176)
Time*pre-voc level							-.101*** (.024)	-.119*** (.025)	-.176*** (.038)	-.180*** (.039)
pre-voc level*grade 8							.732 (10.089)	.538 (10.098)	-7.964 (12.467)	-9.131 (12.706)
pre-voc level*grade 9							-9.590 (12.233)	-9.319 (12.247)	-18.754 (13.096)	-18.876 (13.092)
non-CLIL*pre-voc level							12.637* (5.666)	12.928# (5.671)	11.030 (6.241)	10.771 (6.262)
Time*grade 8							-.319*** (.070)	-.342*** (.070)	-1.030*** (.238)	-1.109*** (.292)

Table 4.10: Results of longitudinal multilevel models for writing fluency (number of words in English), effects of CLIL. Standard errors in parentheses. (*continued*)

Model	1	2	3	4	5	6	7	8	9	10
Time*grade 9							-.297*** (.082)	-.314*** (.082)	-1.232** (.423)	-1.244** (.423)
non-CLIL*grade 8							4.235 (15.877)	2.874 (15.899)	-9.898 (24.514)	-7.341 (25.100)
non-CLIL*grade9							-6.955 (21.374)	-8.402 (21.403)	-26.639 (22.796)	-26.838 (22.790)
3-way interactions										
pre-voc level*time*grade 8									.248* (.100)	.284* (.127)
pre-voc level*time*grade 9									.264* (.131)	.268* (.127)
pre-voc level*time*non-CLIL									.056 (.052)	.063 (.054)
pre-voc level*grade 8*non-CLIL									1.288 (22.157)	4.493 (23.176)
pre-voc level*grade 9*non-CLIL									.000 (.000)	.000 (.000)
time *grade 8*non-CLIL									.414** (.149)	.593 (.410)
time *grade 9*non-CLIL									.504* (.223)	.511* (.224)
4-way interactions										
pre-voc level*time*grade 8*non-CLIL										-.097 (.206)

Table 4.10: Results of longitudinal multilevel models for writing fluency (number of words in English), effects of CLIL. Standard errors in parentheses. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
pre-voc level*time*grade 9*non-CLIL										.000 (.000)
Random part										
School variance			223.549 (216.740)							
Class variance		722.320 (223.734)	510.823 (192.610)	611.580 (192.304)	608.758 (191.514)	200.419 (75.360)	149.794 (61.110)	149.809 (61.196)	150.894 (61.443)	150.613 (61.368)
Pupil variance	1605.055 (123.722)	1001.020 (90.870)	1000.486 (90.840)	1002.257 (90.949)	1005.173 (91.067)	1001.989 (90.906)	1015.578 (90.403)	1026.628 (90.759)	1029.208 (90.457)	1029.846 (90.482)
Repeated measures variance	1027.241 (52.088)	1027.982 (51.937)	1028.099 (51.942)	1028.145 (51.945)	1026.395 (51.859)	1028.965 (52.000)	987.536 (49.926)	977.386 (49.414)	962.358 (48.658)	961.920 (48.636)
Deviance	14241.214	14102.559	14100.612	14099.420	14098.792	14076.908	14037.485	14031.887	14018.242	14018.022
Model of reference and fit improvement		model 1 X^2=138.655 df=1 p<.001	model 2 X^2=1.947 df=1 p=n.s.	model 2 X^2=3.139 df=1 p<.10	model 4 X^2=.628 df=1 p=n.s.	model 4 X^2=22.512 df=3 p<.001	model 6 X^2=39.423 df=8 p<.001	model 7 X^2=5.598 df=1 p<.05	model 8 X^2=13.645 df=7 p<.10	model 9 X^2=.220 df=2 p=n.s.
prop. expl. var. class level				-		.672	.253	-		
prop. expl. var. Pupil level				-		.000	-	-		
prop. expl. var. Rep. meas. level				-		-	.040	.010		
	Nrepmeas=1367; Npupils=594; Nclasses=25; Nschools=7									

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.11: Results of multilevel models with pretest-scores sum speaking proficiency (models 1–4) and speaking fluency (number of words in English, models 5–8) as dependent variables (standard errors in parentheses)

Model	1	2	3	4	5	6	7	8
Fixed part	**sumspeaking_0**	**sumspeaking_0**	**sumspeaking_0**	**sumspeaking_0**	**sp.0.words.english**	**sp.0.words.english**	**sp.0.words.english**	**sp.0.words.english**
Intercept	-.332 (.059)	-.279 (.160)	-.279 (.160)	.037 (.153)	84.817 (3.670)	86.800 (10.352)	86.800 (10.352)	104.569 (10.525)
Pre-voc level (gm) (1=high: 5=low)	-.113** (.037)	-.169# (.095)	-.169# (.095)	-.246** (.079)	-4.898* (2.303)	-8.054 (6.150)	-8.054 (6.150)	-12.400* (5.404)
grade 8 (ref.=grade 7)	.437*** (.100)	.416 (.272)	.416 (.272)	.289 (.218)	26.367*** (6.237)	26.668 (17.592)	26.668 (17.592)	19.488 (14.993)
grade 9	.734*** (.111)	.861** (.291)	.861** (.291)	.796** (.231)	49.575*** (6.921)	62.175** (18.798)	62.175** (18.798)	58.379** (15.884)
non-CLIL (0=CLIL; 1=non-CLIL)				-.671** (.180)				-37.581** (12.389)
Random part								
School variance			.000 (.000)				.000 (.000)	
Class variance		.253 (.081)	.253 (.081)	.147 (.051)		1064.332 (337.985)	1064.332 (337.985)	718.590 (239.874)
Pupil variance	.785 (.051)	.590 (.040)	.590 (.040)	.591 (.040)	3028.412 (198.611)	2330.531 (157.155)	2330.531 (157.155)	2333.870 (157.393)
Total variance	.785	.843	.843	.738	3028.412	3394.863	3394.863	3052.460
Deviance	1207.116	1128.411	1128.411	1117.311	5046.957	4980.571	4980.571	4972.711
Model of reference and fit improvement		model 1 X^2=78.705 df=1 p<.001	model 2 X^2=.000 df=1 p=n.s.	model 2 X^2=11.100 df=1 p<.001		model 5 X^2=66.386 df=1 p<.001	model 6 X^2=.000 df=1 p=n.s.	model 6 X^2=7.860 df=1 p<.01
prop. expl. var. class level				.419				.325
prop. expl. var. Pupil level				-				-
Total				.125				.101
N	Npupils=465; Nclasses=25; Nschools=7				Npupils=465; Nclasses=25; Nschools=7			

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.12: Results of longitudinal multilevel models for Willingness to Communicate in London, effects of CLIL. Standard errors in parentheses.

Model	1	2	3	4	5	6	7	8	9	10
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	26.686 (.264)	26.729 (.300)	26.729 (.300)	27.334 (.341)	27.502 (.357)	26.952 (.442)	27.137 (.454)	26.773 (.541)	26.769 (.557)	26.779 (.557)
Time (in weeks)	.020*** (.004)	.020*** (.004)	.020*** (.004)	.020*** (.004)	.014** (.005)	.021*** (.004)	.015*** (.005)	.016** (.006)	.016* (.007)	.016* (.007)
non-CLIL (0=clil; 1=non-clil)				-1.385** (.464)	-1.775** (.526)	-1.472** (.480)	-1.914** (.543)	-1.187 (.741)	-1.190 (.781)	-1.192 (.781)
Grade 8 (grade 7=ref.cat.)						.614 (.590)	.605 (.590)	.490 (.943)	.132 (1.037)	.225 (1.047)
Grade 9						1.020 (.623)	1.053 (.623)	3.316* (1.103)	3.225# (1.181)	3.215 (1.181)
Pre-voc level (gm) (1=high: 5=low)						-.233 (.216)	-.257 (.216)	-.161 (.356)	-.082 (.383)	-.099 (.384)
2-way interactions										
non-CLIL*Time					.012# (.007)		.013# (.007)	.010 (.008)	.016 (.022)	.018 (.023)
Time*pre-voc level								-.002 (.003)	-.003 (.005)	-.003 (.005)
pre-voc level*grade 8								-.668 (.882)	-1.543 (1.154)	-1.340 (1.197)
pre-voc level*grade 9								-1.708 (1.117)	-2.174 (1.251)	-2.157 (1.251)
non-CLIL*pre-voc level								.050 (.450)	.058 (.538)	.096 (.541)
Time*grade 8								.002 (.009)	-.050 (.039)	-.031 (.049)
Time*grade 9								-.006 (.011)	.038 (.055)	-.037 (.055)
non-CLIL*grade 8								-.720 (1.254)	-.330 (1.989)	-.735 (2.089)
non-CLIL*grade9								-4.488# (1.860)	-4.347 (2.084)	-4.319 (2.085)
3-way interaction										
pre-voc level*time*grade 8									.022 (.016)	.014 (.021)
pre-voc level*time*grade 9									.014 (.017)	.014 (.017)
pre-voc level*time*non-CLIL									-.001 (.007)	-.002 (.007)

Table 4.12: Results of longitudinal multilevel models for Willingness to Communicate in London, effects of CLIL. Standard errors in parentheses. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
pre-voc level*grade 8*non-CLIL									.935 (1.860)	.400 (2.041)
pre-voc level*grade 9*non-CLIL									.000 (.000)	.000 (.000)
time *grade 8*non-CLIL									.013 (.020)	-.028 (.068)
time *grade 9*non-CLIL									-.013 (.030)	-.014 (.030)
4-way interaction										
pre-voc level*time*grade 8*non-CLIL										.021 (.033)
pre-voc level*time*grade 9*non-CLIL										.000 (.000)
Random part										
School variance			.000 (.000)							
Class variance		.499 (.522)	.499 (.522)							
Pupil variance	24.393 (1.904)	23.908 (1.914)	23.908 (1.914)	23.919 (1.876)	23.887 (1.873)	23.550 (1.856)	23.489 (1.851)	23.182 (1.832)	23.241 (1.832)	23.247 (1.833)
Repeated measures variance	15.745 (.815)	15.748 (.815)	15.748 (.815)	15.744 (.815)	15.710 (.813)	15.767 (.816)	15.734 (.814)	15.702 (.813)	15.597 (.807)	15.588 (.807)
Deviance	8303.948	8302.952	8302.952	8295.122	8292.636	8289.521	8286.517	8278.913	8274.085	8273.680
Model of reference and fit improvement		model 1 X^2=.996 df=1 p=n.s.	model 2 X^2=.000 df=1 p=n.s.	model 1 X^2=8.826 df=1 p<.01	model 4 X^2=2.486 df=1 p=n.s.	model 4 X^2=5.601 df=3 p=n.s.	model 6 X^2=3.004 df=1 p=n.s.	model 6 X^2=10.608 df=9 p=n.s.	model 8 X^2=4.828 df=7 p=n.s.	model 9 X^2=.405 df=2 p=n.s.
prop. expl. var. Pupil level				.019						
prop. expl. var. Rep. meas. level				.000						
Nrepmeas=1330; Npupils=590; Nclasses=25; Nschools=7										

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.13: Results of longitudinal multilevel models for Willingness to Communicate at School, effects of CLIL. Standard errors in parentheses.

Model	1	2	3	4	5	6	7	8	9	10
Fixed part	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**	**Coeff. (SE)**
Intercept	16.531 (.149)	16.604 (.223)	16.604 (.223)	17.085 (.236)	17.158 (.245)	16.838 (.264)	16.923 (.270)	16.637 (.304)	16.540 (.310)	16.540 (.310)
Time (in weeks)	.009*** (.002)	.010*** (.002)	.010*** (.002)	.010*** (.002)	.007* (.003)	.010*** (.002)	.008** (.003)	.010* (.004)	.012** (.004)	.012** (.004)
non-CLIL (0=clil; 1=non-clil)				-.1.155** (.338)	-1.323** (.369)	-1.294*** (.289)	-1.494*** (.323)	-1.003* (.417)	-.963 (.437)	-.963 (.437)
Grade 8 (grade 7=ref.cat.						.467 (.354)	.462 (.350)	.637 (.527)	.585 (.577)	.589 (.584)
Grade 9						.602# (.375)	.614 (.371)	1.758* (.614)	1.982# (.658)	1.982 (.658)
Pre-voc level (gm) (1=high: 5=low)						-.312* (.130)	-.321* (.129)	-.240 (.198)	-.043 (.213)	-.043 (.437)
2-way interactions										
non-CLIL*Time					.005 (.004)		.006 (.004)	.004 (.005)	-.016 (.013)	-.015 (.014)
Time*pre-voc level								-.001 (.002)	-.006** (.003)	-.006** (.003)
pre-voc*grade 8								-.438 (.488)	-.984 (.639)	-.976 (.667)
pre-voc level*grade 9								-.972 (.616)	-1.711# (.695)	-1.710 (.696)
non-CLIL*pre-voc								.026 (.250)	-.297. (.301)	-.296 (.303)
Time*grade 8								-.005 (.006)	-.025 (.023)	-.024 (.029)
Time*grade 9								.000 (.007)	-.062# (.033)	-.062# (.033)
non-CLIL*grade 8								-.639 (.696)	-.154 (1.099)	-.170 (1.162)
non-CLIL*grade9								-2.456* (1.029)	-3.060# (1.160)	-3.059 (1.160)
3-way interaction										
pre-voc level*time*grade 8									.006 (.010)	.006 (.012)
pre-voc level*time*grade 9									.022* (.010)	.022* (.010)
pre-voc level*time*non-CLIL									.007# (.004)	.007# (.004)

Table 4.13: Results of longitudinal multilevel models for Willingness to Communicate at School, effects of CLIL. Standard errors in parentheses. (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
pre-voc level*grade 8*non-CLIL									1.128 (1.018)	1.108 (1.135)
pre-voc level*grade 9*non-CLIL									.000 (.000)	.000 (.000)
time *grade 8*non-CLIL									.012 (.012)	.011 (.040)
time *grade 9*non-CLIL									.012 (.018)	.012 (.018)
4-way interaction										
pre-voc level*time*grade 8*non-CLIL										.001 (.020)
pre-voc level*time*grade 9*non-CLIL										.000 (.000)
Random part										
School variance			.000 (.000)							
Class variance		.686 (.311)	.686 (.311)	.292 (.198)	.298 (.200)	.077 (.133)	.069 (.130)	.011 (.112)	.003 (.109)	.003 (.109)
Pupil variance	7.279 (.596)	6.610 (.570)	6.610 (.570)	6.649 (.572)	6.634 (.571)	6.577 (.567)	6.563 (.566)	6.517 (.563)	6.494 (.560)	6.494 (.560)
Repeated measures variance	5.598 (.290)	5.614 (.291)	5.614 (.291)	5.609 (.290)	5.606 (.290)	5.619 (.291)	5.616 (.291)	5.606 (.290)	5.563 (.288)	5.563 (.288)
Deviance	6842.089	6828.504	6828.504	6819.188	6817.880	6807.165	6805.375	6797.410	6788.365	6788.364
Model of reference and fit improvement		model 1 X^2=13.585 df=1 p<.001	model 2 X^2=.000 df=1 p=n.s.	model 2 X^2=9.316 df=1 p<.01	model 4 X^2=1.308 df=1 p=n.s.	model 4 X^2=12.023 df=3 p<.01	model 6 X^2=1.790 df=1 p=n.s.	model 6 X^2=9.755 df=9 p=n.s.	model 8 X^2=9.045 df=7 p=n.s.	model 9 X^2=.001 df=2 p=n.s.
prop. expl. var. class level				.574		.736				
prop. expl. var. Pupil level				-		.011				
prop. expl. var. Rep. meas. level				.001		-				
		Nrepmeas=1328; Npupils=590; Nclasses=25; Nschools=7								

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.14: Results of longitudinal multilevel models for writing proficiency (z-scores), variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score)

Model	1	2	3	4	5	6
Fixed part	**Gender-1**	**Gender-2**	**Cito-1**	**Cito-2**	**Home language-1**	**Home language-2**
Intercept	-.028 (.127)	-.130 (.128)	-.155 (.146)	-.149 (.146)	-.195 (.124)	-.195 (.124)
Time (in weeks)	.006*** (.001)	.006*** (.001)	-.003 (.055)	.023 (.061)	.006*** (.001)	.006*** (.001)
non-CLIL (0=clil; 1=non-clil)	-.269 (.182)	-.264 (.184)	-.366 (.205)	-.374 (.205)	-.275 (.178)	-.275 (.178)
Grade 8 (grade 7=ref.cat.)	.281 (.220)	.280 (.220)	.308 (.266)	.295 (.265)	.259 (.222)	.259 (.222)
Grade 9	1.265* (.235)	1.266* (.235)	1.239* (.281)	1.232* (.281)	1.228* (.238)	1.227* (.238)
Pre-voc level (gm) (1=high: 5=low)	-.249* (.077)	-.249# (.077)	-.145 (.103)	-.149 (.103)	-.268* (.078)	-.268# (.078)
Gender (boy=1; girl=0)	-.230** (.081)	-.224** (.084)				
Cito (gm)			.021* (.009)	.023* (.010)		
Home language (gm)					-.012 (.030)	-.013 (.030)
2-way interactions						
non-CLIL*Time	-.004*** (.001)	-.004*** (.001)	-.004* (.002)	-.098 (.095)	-.004*** (.001)	-.004*** (.002)
Time*pre-voc level	-.003*** (.000)	-.003*** (.000)	-.003*** (.001)	-.003*** (.001)	-.003*** (.000)	-.003*** (.000)
Time*grade 8	-.016*** (.003)	-.016*** (.003)	-.013** (.005)	-.012* (.005)	-.016*** (.003)	-.016*** (.003)
Time*grade 9	-.009 (.006)	-.010 (.006)	-.011 (.009)	-.010 (.009)	-.009 (.006)	-.009 (.006)
pre-voc*grade 8	-.285 (.198)	-.286 (.198)	-.328 (.239)	-.327 (.238)	-.284 (.198)	-.284 (.198)
pre-voc*grade 9	-.482 (.240)	-.483 (.240)	-.564 (.294)	-.555 (.294)	-.496 (.241)	-.497 (.241)
pre-voc*non-CLIL	.213 (.108)	.213 (.108)	.201 (.138)	.200 (.138)	.232 (.110)	.232 (.110)
non-CLIL*grade 8	-.402 (.311)	-.401 (.311)	-.289 (.363)	-.271 (.363)	-.368 (.315)	-.368 (.315)
non-CLIL*grade 9	-.750 (.420)	-.751 (.419)	-.829 (.500)	-.801 (.499)	-.679 (.425)	-.679 (.425)
Gender*non-CLIL	.128 (.110)	.115 (.124)				
Gender*Time	-.001 (.001)	-.001 (.001)				
Cito*Time			.000 (.000)	-.000 (.000)		

Table 4.14: Results of longitudinal multilevel models for writing proficiency (z-scores), variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**
Cito*non-CLIL			-.000 (.014)	-.006 (.015)		
Home language*Time					.000 (.000)	.000 (.000)
Home language*non-CLIL					-.002 (.044)	-.001 (.049)
3-way interaction						
Time*pre-voc*grade 8	.004*** (.001)	.004*** (.001)	.003 (.002)	.003 (.002)	.004*** (.001)	.004*** (.001)
Time*pre-voc*grade 9	.002 (.002)	.002 (.002)	.002 (.003)	.002 (.003)	.002 (.002)	.002 (.002)
Time*non-CLIL*grade 8	.006** (.002)	.006** (.002)	.004 (.003)	.003 (.003)	.005* (.002)	.006** (.002)
Time*non-CLIL*grade 9	.000 (.003)	.000 (.003)	.003 (.004)	.002 (.004)	.000 (.003)	.000 (.003)
Time*gender*non-CLIL		.000 (.002)				
Time*cito*non-CLIL				.000 (.000)		
Time*home language*non-CLIL						-.000 (.001)
Random part						
Class variance	.061 (.023)	.061 (.023)	.071 (.030)	.070 (.030)	.059 (.023)	.059 (.023)
Pupil variance	.323 (.025)	.323 (.025)	.337 (.035)	.339 (.035)	.338 (.027)	.338 (.027)
Repeated measures variance	.206 (.010)	.206 (.010)	.209 (.014)	.208 (.014)	.204 (.010)	.204 (.010)
Deviance	2641.033	2640.984	1554.364	1553.410	2577.403	2577.400
Model of reference and fit improvement		model 1 X^2=.049 df=1 p=n.s.		model 3 X^2=.954 df=1 p=n.s.		model 5 X^2=.003 df=1 p=n.s.
N	Nrepmeas=1367; Npupils=594; Nclasses=25; Nschools=7		Nrepmeas=795; Npupils=333; Nclasses=25;Nschools=7		Nrepmeas=1332; Npupils=565; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.15: Results of longitudinal multilevel models for writing proficiency (z-scores), variables moderating the effect of CLIL. Standard errors in parentheses.

Model	7	8	9	10
Fixed part	**Years Primary English-1**	**Years Primary English-2**	**Freq.Primary English-1**	**Freq.Primary English-2**
Intercept	-.228 (.125)	-.227 (.125)	-.209 (.124)	-.208 (.124)
Time (in weeks)	.005*** (.001)	.005* (.002)	.006*** (.001)	.006*** (.001)
non-CLIL (0=clil; 1=non-clil)	-.252 (.178)	-.253 (.179)	-.260 (.178)	-.259 (.178)
Grade 8 (grade 7=ref.cat.)	.298 (.223)	.297 (.223)	.265 (.222)	.264 (.222)
Grade 9	1.321* (.239)	1.319* (.239)	1.229* (.237)	1.229* (.237)
Pre-voc level (gm) (1=high: 5=low)	-.259* (.078)	-.259# (.078)	-.267* (.078)	-.268# (.078)
Years Pri.English (gm)	.120** (.043)	.118** (.045)		
Frequency Pri.English (gm)			.110** (.040)	.105* (.041)
2-way interactions				
non-CLIL*Time	-.004*** (.001)	-.004 (.003)	-.004*** (.001)	-.003 (.002)
Time*pre-voc level	-.003*** (.000)	-.003*** (.000)	-.003*** (.000)	-.003*** (.000)
Time*grade 8	-.016*** (.003)	-.016*** (.003)	-.016*** (.003)	-.016*** (.003)
Time*grade 9	-.010# (.006)	-.010# (.006)	-.010# (.006)	-.010# (.006)
pre-voc*grade 8	-.285 (.198)	-.285 (.198)	-.294 (.198)	-.293 (.198)
pre-voc*grade 9	-.474 (.241)	-.475 (.241)	-.465# (.241)	-.465# (.241)
pre-voc*non-CLIL	.233 (.110)	.233 (.110)	.231 (.109)	.230 (.109)
non-CLIL*grade 8	-.390 (.316)	-.388 (.316)	-.370 (.315)	-.368 (.315)
non-CLIL*grade 9	-.744 (.425)	-.742 (.425)	-.667 (.424)	-.671 (.424)
Years Pri.English*Time	.000 (.000)	.000 (.001)		
Years Pri.English*non-CLIL	-.062 (.063)	-.055 (.071)		
Frequency Pri.English*Time			.000 (.000)	.000 (.001)
Frequency Pri.English*non-CLIL			-.028 (.062)	-.011 (.070)

Table 4.15: Results of longitudinal multilevel models for writing proficiency (z-scores), variables moderating the effect of CLIL. Standard errors in parentheses. (*continued*)

Model	7	8	9	10
3-way interaction				
Time*pre-voc*grade 8	.004*** (.001)	.004*** (.001)	.004*** (.001)	.004*** (.001)
Time*pre-voc*grade 9	.002 (.002)	.002 (.002)	.002 (.002)	.002 (.002)
Time*non-CLIL*grade 8	.006** (.002)	.006** (.002)	.006** (.002)	.006** (.002)
Time*non-CLIL*grade 9	.000 (.003)	.000 (.003)	.000 (.003)	.001 (.003)
Time*Years Pri.English*non-CLIL		-.000 (.001)		
Time*Freq.Pri.English*non-CLIL				-.000 (.001)
Random part				
Class variance	.060 (.023)	.060 (.023)	.060 (.023)	.060 (.023)
Pupil variance	.330 (.026)	.330 (.026)	.331 (.026)	.331 (.026)
Repeated measures variance	.204 (.010)	.204 (.010)	.203 (.010)	.203 (.010)
Deviance	2568.820	2568.782	2563.635	2563.360
Model of reference and fit improvement		model 7 X^2=.038 df=1 p=n.s.		model 9 X^2=.005 df=1 p=n.s.
N	Nrepmeas=1335; Npupils=566; Nclasses=25; Nschools=7		Nrepmeas=1332; Npupils=565; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.16: Results of longitudinal multilevel models for writing fluency, variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score)

Model	1	2	3	4	5	6
Fixed part	**Gender-1**	**Gender-2**	**Cito-1**	**Cito-2**	**Home language-1**	**Home language-2**
Intercept	99.813 (7.169)	98.809 (7.182)	100.455 (8.355)	100.788 (8.363)	94.680 (7.067)	94.690 (7.069)
Time (in weeks)	.611*** (.058)	.644*** (.062)	.542 (3.787)	1.829 (4.218)	.635*** (.085)	.649*** (.098)
non-CLIL (0=clil; 1=non-clil)	-15.774 (10.486)	-13.460 (10.585)	-23.395 (11.895)	-23.834 (11.903)	-15.044 (10.346)	-15.024 (10.349)
Grade 8 (grade 7=ref.cat.)	18.451 (12.472)	18.307 (12.430)	13.134 (15.291)	12.508 (15.308)	17.775 (12.785)	17.725 (12.790)
Grade 9	47.801* (13.598)	47.956* (13.556)	40.958# (16.500)	40.647 (16.496)	48.998* (13.955)	49.083* (13.962)
Pre-voc level (gm) (1=high: 5=low)	-7.962# (3.425)	-7.867 (3.415)	-6.075 (4.448)	-6.276 (4.454)	-7.905# (3.511)	-7.912 (3.512)
Gender (boy=1; girl=0)	-14.027** (4.764)	.157 (.112)				
Cito (gm)			1.275* (.558)	1.350* (.568)		
Home language (gm)					1.424 (1.723)	1.560 (1.792)
2-way interactions						
non-CLIL*Time	-.242** (.076)	-.314*** (.092)	-.249* (.105)	-4.776 (6.547)	-.281*** (.077)	-.315* (.146)
Time*pre-voc level	-.135*** (.027)	-.139*** (.027)	-.101* (.046)	-.095* (.047)	-.146*** (.028)	-.146*** (.028)
Time*grade 8	-.958*** (.232)	-.988*** (.233)	-.640# (.329)	-.612# (.332)	-.962*** (.233)	-.962*** (.233)
Time*grade 9	-1.171** (.412)	-1.216** (.413)	-.904 (.594)	-.851 (.598)	-1.085** (.417)	-1.086** (.417)
pre-voc*grade 8	-7.716 (11.482)	-8.119 (11.448)	-9.820 (14.029)	-9.759 (14.020)	-8.485 (11.699)	-8.503 (11.702)
pre-voc*grade 9	-21.971 (13.742)	-22.512 (13.705)	-24.983 (17.109)	-24.527 (17.112)	-25.075 (13.999)	-25.063 (14.003)
non-CLIL*grade 8	-23.404 (16.441)	-23.050 (16.387)	-12.251 (19.514)	-11.313 (19.547)	-23.803 (16.873)	-23.666 (16.885)
non-CLIL*grade 9	-38.845 (23.707)	-39.251 (23.634)	-29.166 (28.757)	-27.801 (28.806)	-40.108 (24.292)	-40.101 (24.298)
Gender*non-CLIL	7.489 (6.388)	2.063 (7.465)				
Gender*Time	-.145** (.056)	-.217** (.076)				
Cito*Time			-.000 (.007)	-.003 (.008)		

Table 4.16: Results of longitudinal multilevel models for writing fluency, variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**
Cito*non-CLIL			-.782 (.792)	-1.057 (.887)		
Home language*Time					-.025 (.021)	-.030 (.026)
Home language*non-CLIL					-.377 (2.515)	-.792 (2.929)
3-way interaction						
Time*pre-voc*grade 8	.240* (.099)	.252* (100)	.148 (.146)	.145 (.146)	.239* (.100)	.240* (.100)
Time*pre-voc*grade 9	.252* (.128)	.268* (.128)	.174 (.190)	.158 (.191)	.215# (.129)	.215# (.129)
Time*non-CLIL*grade 8	.352** (.136)	.348* (.136)	.242 (.184)	.206 (.191)	.361** (.137)	.356** (.138)
Time*non-CLIL*grade 9	.468* (.217)	.474* (.217)	.410 (.293)	.268 (.299)	.471* (.220)	.471* (.220)
Time*gender*non-CLIL		.157 (.112)				
Time*cito*non-CLIL				.009 (.012)		
Time*home language*non-CLIL						.011 (.042)
Random part						
Class variance	202.477 (75.421)	200.617 (74.858)	234.295 (102.148)	233.751 (101.848)	212.060 (79.076)	212.239 (79.125)
Pupil variance	965.864 (86.578)	962.435 (86.363)	1081.877 (126.551)	1082.088 (126.591)	985.705 (88.774)	985.593 (88.766)
Repeated measures variance	957.161 (48.370)	956.761 (48.349)	1007.159 (65.988)	1006.283 (65.915)	949.941 (48.252)	949.891 (48.249)
Deviance	13993.384	13991.427	8194.793	8194.317	13632.060	13631.983
Model of reference and fit improvement		model 1 X^2=1.957 df=1 p=n.s.		model 3 X^2=.476 df=1 p=n.s.		model 5 X^2=.177 df=1 p=n.s.
N	Nrepmeas=1367; Npupils=594; Nclasses=25; Nschools=7		Nrepmeas=795; Npupils=333; Nclasses=25; Nschools=7		Nrepmeas=1332; Npupils=565; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.17 (Continuation): Results of longitudinal multilevel models for writing fluency, variables moderating the effect of CLIL. Standard errors in parentheses.

Model	7	8	9	10
Fixed part	**Years Primary English-1**	**Years Primary English-2**	**Freq.Primary English-1**	**Freq.Primary English-2**
Intercept	93.213 (7.048)	93.311 (7.044)	93.949 (6.964)	93.932 (6.964)
Time (in weeks)	.472*** (.094)	.440*** (.111)	.523*** (.078)	.530*** (.088)
non-CLIL (0=clil; 1=non-clil)	-13.777 (10.297)	-14.004 (10.296)	-14.586 (10.195)	-14.608 (10.194)
Grade 8 (grade 7=ref.cat.)	19.504 (12.722)	19.416 (12.712)	18.173 (12.588)	18.198 (12.588)
Grade 9	51.694* (13.942)	51.408* (13.941)	47.821* (13.727)	47.822* (13.725)
Pre-voc level (gm) (1=high; 5=low)	-7.951# (3.471)	-7.951 (3.468)	-8.165# (3.446)	-8.155# (3.446)
Gender (boy=1; girl=0)				
Cito (gm)				
Home language (gm)				
Years Pri.English (gm)	5.114* (2.542)	4.695# (2.654)		
Frequency Pri.English (gm)			4.519# (2.366)	4.623 (2.457)
2-way interactions				
non-CLIL*Time	-.269*** (.077)	-.185 (.171)	-.263*** (.077)	-.281* (.136)
Time*pre-voc level	-.139*** (.027)	-.140*** (.027)	-.147*** (.027)	-.147*** (.027)
Time*grade 8	-.956*** (.233)	-.952*** (.233)	-.976*** (.233)	-.977*** (.233)
Time*grade 9	-1.068** (.416)	-1.071** (.415)	-1.097** (.417)	-1.095** (.417)
pre-voc*grade 8	-8.655 (11.625)	-8.612 (11.616)	-9.110 (11.534)	-9.122 (11.533)
pre-voc*grade 9	-23.943 (13.921)	-24.035 (13.911)	-23.647 (13.828)	-23.640 (13.826)
non-CLIL*grade 8	-25.194 (16.745)	-24.775 (16.748)	-23.373 (16.594)	-23.428 (16.595)
non-CLIL*grade 9	-41.415 (24.134)	-41.044 (24.124)	-37.434 (23.922)	-37.353 (23.924)
Years Pri.English*Time	.037 (.033)	.051 (.041)		
Years Pri.English*non-CLIL	-5.734 (3.620)	-4.513 (4.248)		

Table 4.17 (Continuation): Results of longitudinal multilevel models for writing fluency, variables moderating the effect of CLIL. Standard errors in parentheses. (*continued*)

Model	7	8	9	10
Frequency Pri.English*Time			.019 (.030)	.016 (.038)
Frequency Pri.English*non-CLIL			-1.021 (3.588)	-1.366 (4.206)
3-way interaction				
Time*pre-voc*grade 8	.240* (.100)	.239* (.100)	.245* (.100)	
Time*pre-voc*grade 9	.222# (.129)	.227# (.129)	.225# (.129)	
Time*non-CLIL*grade 8	.354** (.136)	.341* (.138)	.342* (.136)	
Time*non-CLIL*grade 9	.431* (.219)	.423# (.219)	.430* (.219)	
Time*Years Pri.English*non-CLIL		-.037 (.067)		
Time*Freq.Pri.English*non-CLIL				.010 (.064)
Random part				
Class variance	209.172 (78.007)	208.653 (77.857)	203.840 (76.546)	203.779 (76.524)
Pupil variance	969.266 (87.629)	969.347 (87.626)	972.277 (87.912)	971.875 (87.891)
Repeated measures variance	948.400 (48.108)	948.110 (48.093)	948.301 (48.164)	948.439 (48.171)
Deviance	13654.457	13654.156	13624.530	13624.504
Model of reference and fit improvement		model 7 X^2= .301 df=1 p=n.s.		model 9 X^2=.026 df=1 p=n.s.
N	Nrepmeaus=1335; Npupils=566; Nclasses=25; Nschools=7		Nrepmeas=1332; Npupils=565; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.18: Results of multilevel models for speaking proficiency (Ancova), effects of CLIL and other variables. Standard errors in parentheses. (posttest=z)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**
Intercept	.035 (.037)	.036 (.058)	-.017 (.088)	-.062 (.089)	-.008 (.071)	.043 (.101)	-.026 (.077)	.060 (.080)
Speaking pretest (z)	.732*** (.038)	.695*** (.042)	.697*** (.040)	.679*** (.042)	.690*** (.040)	.679*** (.042)	.670*** (.042)	.637*** (.044)
non-CLIL (0=clil; 1=non-clil)						-.154 (.106)		-.214* (.094)
Grade 8 (grade 7=ref.cat.)				.039 (.102)			-.020 (.106)	-.031 (.100)
Grade 9				.218 (.127)			.234# (.123)	.272* (.117)
Pre-voc level (gm) (1=high; 5=low)					-.086# (.048)		-.102* (.048)	-.126** (.045)
Random part								
School variance			.041 (.029)	.033 (.024)	.021 (.019)	.046 (.031)	.017 (.016)	.014 (.013)
Class variance		.054 (.024)	.012 (.014)	.009 (.013)	.014 (.015)	.009 (.013)	.009 (.013)	.004 (.011)
Pupil variance	.390 .032)	.343 (.030)	.344 (.030)	.343 (.030)	.343 (.030)	.342 (.030)	.342 (.030)	.341 (.030)
Total	.390	.397	.397	.385	.378	.397	.368	.359
prop. expl School variance								.176
prop. expl Class variance								.556
prop. expl Pupil variance								.003
Prop. Expl total var	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	.024
Deviance	548.306	535.991	529.950	526.989	527.500	527.888	523.171	518.296
Model of reference and fit improvement		model 1 X^2=12.315 df=1 p<.001	model 2 X^2=6.041 df=1 p<.05	model 3 X^2=2.961 df=2 p=n.s.	model 3 X^2=2.450 df=1 p=n.s.	model 3 X^2=2.062 df=1 p=n.s.	model 3 X^2=6.779 df=3 p=n.s.	model 7 X^2=4.875 df=1 p<.05
N	Npupils=289; Nclasses=25; Nschools=7							

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.19: Results of multilevel models for speaking proficiency (change in z-scores), effects of CLIL and other variables. Standard errors in parentheses.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**
Intercept	.052 (.040)	.039 (.060)	.025 (.079)	.071 (.082)	.037 (.059)	.022 (.078)	-.021 (.070)	.080 (.076)
non-CLIL (0=clil; 1=non-clil)						.039 (.120)		-.234* (.096)
Grade 8 (grade 7=ref.cat.)				-.087 (.137)			-.013 (.126)	-.038 (.115)
Grade 9				-.038 (.158)			.273# (.138)	.292* (.127)
Pre-voc level (gm) (1=high: 5=low)					-.036 (.046)		-.110* (.043)	-.144** (.041)
Random part								
School variance			.025 (.024)					
Class variance		.051 (.025)	.025 (.021)	.050 (.025)	.048 (.024)	.051 (.025)	.027 (.016)	.017 (.013)
Pupil variance	.457 (.038)	.409 (.036)	.409 (.036)	.408 (.036)	.409 (.036)	.409 (.036)	.342 (.030)	.341 (.030)
Total variance	.457	.460	.459	.458	.457	.460	.369	.358
prop. expl. Class variance								.370
prop. expl. Pupil variance								.003
prop. expl. total variance	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	.198	.030
Deviance	594.151	583.422	581.065	583.021	582.847	583.318	525.741	520.266
Model of reference and fit improvement		model 1 X^2=10.729 df=1 p<.001	model 2 X^2=2.357 df=1 p=n.s.	model 2 X^2=.041 df=2 p=n.s.	model 2 X^2=.174 df=1 p=n.s.	model 2 X^2=.104 df=1 p=n.s.	model 2 X^2=57.681 df=3 p<.001	model 7 X^2=5.475 df=1 p<.05
N	Npupils=289; Nclasses=25; Nschools=7							

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.20: Results of multilevel models for speaking proficiency (Ancova), variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Fixed part	**Grade-1**	**Grade-2**	**Pre-voc levelrec-1**	**Pre-voc levelrec-2**	**Gender-1**	**Gender-2**	**Cito-1**	**Cito-2**	**Home language-1**	**Home language-2**
Intercept	.060 (.080)	.021 (.084)	.060 (.080)	.061 (.079)	-.032 (.083)	.010 (.088)	-.062 (.088)	-.062 (.089)	.060 (.081)	.058 (.081)
Sumspeaking_0	.637*** (.044)	.630*** (.044)	.637*** (.044)	.632*** (.045)	.637*** (.044)	.639*** (.044)	.714*** (.052)	.717*** (.053)	.636*** (.044)	.637*** (.044)
non-CLIL (0=clil; 1=non-clil)	-.214* (.094)	-.130 (.118)	-.214* (.094)	-.202# (.098)	-.220* (.093)	-.174 (.113)	-.103 (.106)	-.106 (.107)	-.214* (.095)	-.208* (.096)
Grade 8 (grade 7=ref. cat.)	-.031 (.100)	.055 (.131)	-.031 (.100)	-.031 (.101)	-.045 (.098)	-.040 (.097)	-.020 (.114)	-.020 (.114)	-.028 (.101)	-.026 (.101)
Grade 9	.272* (.117)	.368* (.160)	.272* (.117)	.280* (.118)	.260* (.115)	.261* (.114)	.327* (.132)	.323* (.133)	.272* (.118)	.279* (.119)
Pre-voc level (gm) (1=high: 5=low)	-.126* (.045)	-.127* (.044)	-.126* (.045)	-.144* (.054)	-.134** (.045)	-.131* (.046)	-.177** (.057)	-.176** (.058)	-.127* (.046)	-.128* (.046)
Gender (boy=1; girl=0)					.081 (.071)	.124 (.095)				
Cito (gm)							.007 (.008)	.008 (.009)		
Home language (gm)									-.006 (.026)	.005 (.033)
2-way interactions										
non-CLIL*cito-gm								.004 (.014)		
Non-CLIL*gender (boy)						-.097 (.141)				
Non-CLIL*pre-voc level				.038 (.074)						
Non-CLIL*grade8		-.184 (.183)								
Non-CLIL*grade9		-.179 (.213)								

Table 4.20: Results of multilevel models for speaking proficiency (Ancova), variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Non-CLIL*home language										-.027 (.052)
Random part										
School variance	.014 (.013)	.010 (.011)	.014 (.013)	.011 (.012)	.014 (.013)	.015 (.014)	.012 (.013)	.013 (.014)	.014 (.013)	.013 (.013)
Class variance	.004 (.011)	.004 (.011)	.004 (.011)	,005 (.012)	.003 (.011)	.002 (.011)	.000 (.000)	.000 (.000)	.004 (.012)	.005 (.012)
Pupil variance	.341 (.030)	.340 (.030)	.341 (.030)	.340 (.030)	.340 (.030)	.340 (.029)	.286 (.032)	.285 (.032)	.342 (.030)	.341 (.030)
total variance	.359	.354	.359	.356	.357	.357	.298	.298	.360	.359
Deviance	518.296	517.062	518.296	518.068	517.037	516.576	274.267	274.173	517.464	517.191
Model of reference and fit improvement		model 1 X^2=1.234 df=2 p=n.s.		model 1 X^2=.228 df=1 p=n.s.		model 4 X^2=.461 df=1 p=n.s.		model 6 X^2=.094 df=1 p=n.s.		model 8 X^2=.273 df=1 p=n.s.
N		Npupils=289; Nclasses=25; Nschools=7					Npupils=170; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.21: Results of multilevel models for speaking proficiency (Ancova), variables moderating the effect of CLIL. Standard errors in parentheses. (PE=Primary school English)

Model	1	2	3	4	5	6	7
Fixed part	**Years PE**	**Years PE**	**Freq. PE**	**Freq. PE**	**Years & freq:PE intensity**	**Years & freq: PE intensity**	**Years & freq: PE intensity**
Intercept	.055 (.082)	.055 (.083)	.061 (.080)	.053 (.082)	.071 (.081)	.065 (.083)	.055 (.084)
Sumspeaking_0	.634*** (.044)	.633*** (.044)	.641*** (.045)	.638*** (.044)	.639 (.044)	.636*** (.044)	.640*** (.044)
non-CLIL (0=clil; 1=non-clil)	-.214* (.096)	-.215* (.097)	-.217* (.094)	-.238* (.095)	-.231* (.093)	-.253* (.094)	-.230* (.096)
Grade 8 (grade 7=ref.cat.)	-.021 (.102)	-.026 (.103)	-.036 (.099)	-.031 (.097)	.001 (.099)	.002 (.096)	.002 (.096)
Grade 9	.285* (.120)	.288* (.121)	.271* (.117)	.289* (.116)	.321* (.118)	.343** (.117)	.337** (.116)
Pre-voc level (gm) (1=high: 5=low)	-.125* (.046)	-.126* (.047)	-.123* (.046)	-.115* (.047)	-.111* (.046)	-.103* (.047)	-.099# (.047)
Years PE (gm)	.025 (.042)	.046 (.051)			.047 (.046)	.064 (.058)	.062 (.058)
Freq. PE (gm)			-.027 (.040)	.010 (.048)	-.036 (.044)	-.002 (.054)	-.006 (.054)
2-way interactions							
Years PE*Freq. PE					-.067# (.034)	-.074# (.035)	-.052 (.040)
Non-CLIL*Years PE		-.059 (.086)				-.056 (.092)	-.066 (.093)
Non-CLIL*Freq.PE				-.122 (.084)		-.114 (.089)	-.103 (.089)
3-way interaction							
Non-CLIL*Years PE*Freq. PE							-.083 (.077)
Random part							
School variance	.014 (.014)	.015 (.014)	.014 (.013)	.018 (.015)	.014 (.013)	.019 (.015)	.019 (.015)
Class variance	.005 (.012)	.006 (.012)	.004 (.011)	.002 (.011)	.003 (.011)	.001 (.010)	.001 (.010)
Pupil variance	.339 (.029)	.338 (.029)	.342 (.030)	.339 (.029)	.337 (.029)	.333 (.029)	.332 (.029)
Total variance	.358						

Table 4.21: Results of multilevel models for speaking proficiency (Ancova), variables moderating the effect of CLIL. Standard errors in parentheses. (PE=Primary school English) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Deviance	517.932	517.465	517.025	515.015	512.369	509.631	508.465
Model of reference and fit improvement		model 1 X^2=.467 df=1 p=n.s.		model 3 X^2=2.010 df=1 p=n.s.		model 5 X^2=2.738 df=2 p=n.s.	model 6 X^2=1.166 df=1 p=n.s.
N	Npupils=289; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7				

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.22: Results of multilevel models for speaking proficiency (change scores), variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score)

Model	1	2	3	4	5	6	7	8	9
Fixed part	**Grade-1/ Pre-voc-levelrec-1**	**Grade-2**	**Pre-voc-levelrec-2**	**Gender-1**	**Gender-2**	**Cito-1**	**Cito-2**	**Home language-1**	**Home language-2**
Intercept	.098 (.100)	.075 (.111)	.094 (.100)	.079 (.103)	-.056 (.108)	-.049 (.093)	-.056 (.093)	.098 (.100)	.096 (.101)
non-CLIL (0=clil; 1=non-clil)	-.001 (.121)	.051 (.164)	-.008 (.121)	-.007 (.120)	.045 (.141)	.104 (.109)	.106 (.108)	-.002 (.121)	.002 (.121)
Grade 8 (grade 7=ref. cat.)	-.164 (.149)	-.117 (.188)	-.161 (.149)	-.173 (.148)	-.168 (.148)	-.190 (.137)	-.185 (.137)	-.167 (.149)	-.164 (.150)
Grade 9	-.072 (.155)	-.027 (.205)	-.076 (.155)	-.077 (.154)	-.072 (.154)	.041 (.143)	.029 (.143)	-.069 (.155)	-.059 (.157)
Pre-voc level (gm) (1=high: 5=low)	-.061 (.052)	-.060 (.053)	-.044 (.065)	-.066 (.052)	-.064 (.052)	-.156* (.058)	-.151* (.058)	-.060 (.053)	-.060 (.053)
Gender (boy=1; girl=0)				.053 (.079)	.102 (.106)				
Cito-gm						-.011 (.009)	-.015 (.010)		
Home language (gm)								.007 (.029)	.022 (.038)
2-way interactions									
non-CLIL*cito-gm							.015 (.015)		
Non-CLIL*gender (boy)					-.112 (.158)				
Non-CLIL*pre-voc level			-.042 (.095)						
Non-CLIL*grade8		-.112 (.269)							
Non-CLIL*grade9		-.101 (.307)							
Non-CLIL*home language									-.038 (.059)
Random part									
Class variance	.044 (.023)	.044 (.023)	.043 (.023)	.043 (.023)	.042 (.022)	.016 (.019)	.015 (.018)	.044 (.023)	.045 (.023)

Table 4.22: Results of multilevel models for speaking proficiency (change scores), variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Pupil variance	.409 (.036)	.409 (.036)	.409 (.036)	.409 (.036)	.408 (.036)	.334 (.039)	.332 (.039)	.411 (.036)	.410 (.036)
Deviance	581.631	581.413	581.439	581.191	580.691	302.616	301.599	580.588	580.184
Model of reference and fit improvement		model 1 X^2=.218 df=2 p=n.s.	model 1 X^2=.192 df=1 p=n.s.		model 4 X^2=.500 df=1 p=n.s.		model 6 X^2=1.017 df=1 p=n.s.		model 8 X^2=.404 df=1 p=n.s.
N	Npupils=289; Nclasses=25; Nschools=7			Npupils=289; Nclasses=25; Nschools=7		Npupils=170; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.23: Results of multilevel models for speaking proficiency (change scores), variables moderating the effect of CLIL . Standard errors in parentheses. (PE=Primary school English)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Fixed part	**Years PE**	**Years PE**	**Frequency PE**	**Frequency PE**	**Years & freq: PE intensity**	**Years & freq: PE intensity**	**Years & freq: PE intensity**
Intercept	.102 (.100)	.102 (.101)	.104 (.098)	.100 (.099)	.116 (.097)	.116 (.100)	.097 (.099)
non-CLIL (0=clil; 1=non-clil)	-.003 (.120)	-.004 (.121)	-.021 (.119)	-.030 (.121)	-.032 (.117)	-.043 (.120)	-.003 (.121)
Grade 8 (grade 7=ref.cat.)	-.171 (.149)	-.176 (.151)	-.178 (.146)	-.177 (.148)	-.152 (.145)	-.156 (.149)	-.152 (.147)
Grade 9	-.078 (.155)	-.074 (.157)	-.063 (.152)	-.054 (.154)	-.029 (.152)	-.012 (.156)	-.017 (.154)
Pre-voc level (gm) (1=high: 5=low)	-.062 (.052)	-.063 (.053)	-.056 (.051)	-.054 (.052)	-.047 (.051)	-.045 (.052)	-.041 (.052)
Years PE (gm)	-.017 (.046)	.002 (.056)			.020 (.050)	.049 (.064)	.045 (.064)
Freq. PE (gm)			-.073# (.044)	-.053 (.052)	-.073 (.048)	-.062 (.060)	-.068 (.060)
2-way interactions							
Years PE* Freq. PE					-.060 (.038)	-.065 (.038)	-.029 (.045)
Non-CLIL*Years PE		-.052 (.095)				-.083 (.103)	-.097 (.103)
Non-CLIL*Freq.PE				-.067 (.094)		-.047 (.101)	-.031 (.101)
3-way interaction							
Years PE* Freq. PE*non-CLIL							-.137 (.086)
Random part							
Class variance	.043 (.023)	.045 (.023)	.041 (.022)	.043 (.023)	.038 ().021	.042 (.022)	.041 (.022)
Pupil variance	.409 (.036)	.408 (.035)	.408 (.036)	.406 (.035)	.405 (.035)	.401 (.035)	.398 (.035)
Total variance	.452	.453	.449	.449	.443	.443	.439

Table 4.23: Results of multilevel models for speaking proficiency (change scores), variables moderating the effect of CLIL . Standard errors in parentheses. (PE=Primary school English) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Deviance	581.501	581.207	577.772	577.276	575.198	573.946	571.388
Model of reference and fit improvement		model 1 X^2=.294 df=1 p=n.s.		model 3 X^2=.496 df=1 p=n.s.		model 5 X^2=1.252 df=2 p=n.s.	model 6 X^2=2.558 df=1 p=n.s.
N	Npupils=289; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7				

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.24: Results of multilevel models for speaking fluency (Ancova), main effects of CLIL and other variables. Standard errors in parentheses.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**
Intercept	142.284 (2.885)	142.055 (6.082)	137.331 (9.136)	141.164 (9.982)	138.720 (8.471)	144.095 (9.627)	144.490 (8.987)	152.903 (8.756)
Words English pretest-gm	.933*** (.046)	.910*** (.050)	.890*** (.046)	.898*** (.047)	.890*** (.046)	.867*** (.047)	.897*** (.047)	.864*** (.049)
non-CLIL (0=clil; 1=non-clil)						-16.989# (9.627)		-20.628*** (8.550)
Grade 8 (grade 7=ref.cat.)				-13.141 (9.049)			-17.594 (9.585)	-18.828* (8.595)
Grade 9				-4.349 (11.359)			-3.160 (11.309)	.088 (10.262)
Pre-voc level (gm) (1=high: 5=low)					-4.750 (5.014)		-8.087 (5.104)	-9.674* (4.610)
Random part								
School variance			470.957 (313.277)	513.610 (328.322)	376.358 (263.675)	454.273 (294.699)	351.361 (241.657)	279.436 (192.745)
Class variance		759.809 (260.881)	203.532 (121.498)	157.429 (105.871)	211.729 (123.851)	147.052 (102.354)	160.591 (106.621)	97.369 (85.283)
Pupil variance	2404.910 (200.062)	1742.464 (151.658)	1741.487 (151.496)	1742.272 (151.545)	1741.682 (151.550)	1744.046 (151.694)	1741.212 (151.476)	1743.424 (151.583)
Total variance	2404.910	2502.273	2415.976	2413.311	2329.769	2345.371	2253.164	2120.229
Deviance	3070.089	3021.063	3009.567	3007.588	3008.796	3006.516	3005.459	3000.264
Model of reference and fit improvement		model 1 X^2=49.026 df=1 p<.001	model 2 X^2=11.496 df=1 p<.001	model 3 X^2=1.979 df=2 p=n.s.	model 3 X^2=.771 df=1 p=n.s.	model 3 X^2=3.051 df=1 p<.10		model 7 X^2=5.195 df=1 p<.05
prop. expl. var. school level						.035		.204
prop. expl. var. class level						.277		.394
prop. expl. var. Pupil level						-		-
prop. expl. var. total						.029		.059
N			Npupils=289 ; Nclasses=25; Nschools=7					

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.25: Results of multilevel models for speaking fluency (change scores), main effects of CLIL and other variables. Standard errors in parentheses.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**
Intercept	39.083 (2.895)	38.395 (6.023)	34.546 (8.802)	40.233 (9.708)	35.759 (8.298)	38.733 (9.440)	43.530 (8.819)	49.982 (9.071)
non-CLIL (0=clil; 1=non-clil)						-10.245 (9.865)		-13.969 (8.879)
Grade 8 (grade 7=ref. cat.)				-16.290 (9.094)			-20.857* (9.673)	-22.684* (9.408)
Grade 9				-11.114 (11.099)			-10.200 (11.022)	-9.618 (10.603)
Pre-voc level (gm) (1=high: 5=low)					-4.467 (5.076)		-7.897 (5.107)	-9.240# (4.845)
Random part								
School variance			417.087 (291.347)	475.851 (310.479)	345.574 (253.849)	391.820 (274.745)	329.171 (232.004)	263.569 (193.449)
Class variance		740.834 (256.270)	241.459 (134.989)	169.007 (110.559)	247.430 (136.367)	224.022 (129.052)	171.961 (111.072)	150.341 (103.688)
Pupil variance	2422.810 (201.551)	1767.300 (153.811)	1768.819 (153.893)	1769.857 (153.956)	1768.863 (153.945)	1769.778 (153.977)	1768.761 (153.898)	1769.842 (153.975)
Total variance	2422.810	2508.134	2427.365	2414.715	2361.867	2385.620	2269.893	2183.752
Deviance	3072.232	3024.345	3015.056	3012.129	3014.363	3014.005	3010.083	3007.769
Model of reference and fit improvement		model 1 X^2=47.887 df=1 p<.001	model 2 X^2=9.289 df=1 p<.01	model 3 X^2=2.927 df=2 p=n.s.	model 3 X^2=.693 df=1 p=n.s.	model 3 X^2=1.051 df=1 p=n.s.		model 7 X^2=2.314 df=1 p=n.s.
N	Npupils=289; Nclasses=25; Nschools=7							

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.26: Results of multilevel models for speaking fluency (number of words in English): Ancova, variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Fixed part	**Grade-1**	**Grade-2**	**Pre-voc levelrec-1**	**Pre-voc levelrec-2**	**Gender-1**	**Gender-2**	**Cito-1**	**Cito-2**	**Home language-1**	**Home language-2**
Intercept	152.903 (8.756)	155.209 (9.455)	152.903 (8.756)	152.823 (8.833)	151.868 (8.847)	152.338 (9.078)	147.912 (8.833)	148.315 (9.043)	152.703 (8.821)	152.762 (8.851)
Sumwords English_0	.864*** (.049)	.877*** (.049)	.864*** (.049)	.863*** (.049)	.860*** (.049)	.860*** (.049)	.851*** (.062)	.858*** (.062)	.860*** (.049)	.860*** (.049)
non-CLIL (0=clil; 1=non-clil)	-20.628* (8.550)	-25.830* (10.926)	-20.628* (8.550)	-20.316# (9.823)	-21.041* (8.492)	-22.090* (9.713)	-12.124 (9.678)	-14.993 (9.891)	-20.745* (8.600)	-20.978# (8.687)
Grade 8 (grade 7=ref. cat.)	-18.828* (8.595)	-21.141# (10.926)	-18.828* (8.595)	-18.832* (8.592)	-19.263* (8.538)	-19.379* (8.558)	-27.051* (10.019)	-26.838* (10.012)	-18.155# (8.696)	-18.186# (8.697)
Grade 9	.088 (10.262)	-12.781 (13.895)	.088 (10.262)	.089 (10.258)	-.100 (10.163)	-.193 (10.176)	9.599 (11.970)	8.436 (12.015)	.390 (10.326)	.271 (10.343)
Pre-voc level (gm) (1=high: 5=low)	-9.674# (4.610)	-9.152# (4.601)	-9.674# (4.610)	-9.816# (5.184)	-10.001* (4.609)	-10.018* (4.607)	-14.240* (5.447)	-14.415* (5.525)	-9.771# (4.651)	-9.731# (4.658)
Gender (boy=1; girl=0)					2.967 (5.213)	1.964 (6.894)				
Cito (gm)							.039 (.673)	-.726 (.788)		
Home language (gm)									-1.161 (1.944)	-1.481 (2.506)
2-way interactions										
non-CLIL*cito-gm								2.155# (1.188)		
Non-CLIL*gender (boy)						2.288 (10.315)				
Non-CLIL*pre-voc level				.485 (7.601)						
Non-CLIL*grade8		5.443 (15.579)								

Table 4.26: Results of multilevel models for speaking fluency (number of words in English): Ancova, variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Non-CLIL*grade9		24.441 (18.400)								
Non-CLIL*home language										.779 (3.876)
Random part										
School variance	279.436 (192.745)	303.357 (201.160)	279.436 (192.745)	280.270 (193.142)	275.823 (189.773)	274.725 (189.248)	205.741 (166.636)	230.182 (179.550)	283.728 (195.553)	286.419 (196.963)
Class variance	97.369 (85.283)	73.740 (76.995)	97.369 (85.283)	97.078 (85.182)	91.424 (83.267)	91.793 (83.383)	49.969 (111.457)	55.284 (111.303)	99.988 (86.408)	99.871 (86.355)
Pupil variance	1743.424 (151.583)	1741.803 (151.433)	1743.424 (151.583)	1743.436 (151.584)	1744.754 (151.693)	1744.401 (151.662)	1839.522 (214.340)	1793.814 (209.091	1744.748 (151.992)	1744.222 (151.946)
Total variance	2120.229	2118.900	2120.229	2120.784	2111.648	2110.919	2095.232	2079.280	2128.464	2130.512
prop. expl. var. school level								-		
prop. expl. var. class level								-		
prop. expl. var. Pupil level								.025		
prop. expl. var. total								.008		
Deviance	3000.264	2998.581	3000.264	3000.259	2999.945	2999.896	1772.543	1769.318	2990.416	2990.376
Model of reference and fit improvement		model 1 X^2=1.683 df=2 p=n.s.		model 1 X^2=.005 df=1 p=n.s.		model 4 X^2=.049 df=1 p=n.s.		model 6 X^2=3.225 df=1 p<.10		model 8 X^2=.040 df=1 p=n.s.
N		Npupils=289; Nclasses=25; Nschools=7			Npupils=289; Nclasses=25; Nschools=7		Npupils=170; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.27: Results of multilevel models for speaking fluency (number of words in English): Ancova, variables moderating the effect of CLIL. Standard errors in parentheses. (PE=Primary school English)

Model	1	2	3	4	5	6	7	8
Fixed part	**Years PE**	**Years PE**	**Freq. PE**	**Freq. PE**	**Years & Freq: PE intensity**	**Years & Freq: PE intensity**	**Years & Freq: PE intensity**	**Years & Freq: PE intensity**
Intercept	152.922 (8.784)	152.985 (8.759)	152.796 (8.617)	153.387 (8.476)	153.358 (8.613)	155.495 (8.361)	155.873 (8.258)	155.482 (8.289)
Sumwords English_0	.864*** (.049)	.865*** (.049)	.857*** (.049)	.863*** (.049)	.860*** (.049)	.859*** (.048)	.864*** (.048)	.866*** (.048)
non-CLIL (0=clil; 1=non-clil)	-20.630* (8.550)	-20.617* (8.556)	-19.818* (8.618)	-18.387* (8.686)	-19.682* (8.606)	-21.150* (8.406)	-19.950* (8.490)	-19.040* (8.635)
Grade 8 (grade 7=ref. cat.)	-18.863* (8.685)	-18.706* (8.712)	-18.165# (8.727)	-18.666* (8.773)	-19.108* (8.796)	-17.074# (8.625)	-17.462# (8.681)	-17.425# (8.721)
Grade 9	.047 (10.366)	-.162 (10.389)	-.105 (10.357)	-1.723 (10.448)	-1.444 (10.479)	1.583 (10.295)	.302 (10.395)	.095 (10.434)
Pre-voc level (gm) (1=high: 5=low)	-9.679# (4.614)	-9.707# (4.608)	-10.417* (4.593)	-10.876* (4.548)	-10.723* (4.594)	-10.342* (4.460)	-10.688* (4.429)	-10.580* (4.441)
Years PE (gm)	-.085 (3.002)	-.980 (3.698)			-2.631 (3.290)	-2.419 (3.254)	-2.094 (4.169)	-2.220 (4.170)
Freq. PE (gm)			4.763# (2.895)	2.335 (3.458)	5.822# (3.179)	6.816* (3.162)	4.660 (3.918)	4.513 (3.922)
2-way interactions								
Years PE*freq PE						-6.679* (2.464)	-6.510* (2.482)	-5.548# (2.889)
Non-CLIL*Years PE		2.584 (6.229)					-.190 (6.646)	-.566 (6.666)
Non-CLIL*Freq.PE				7.963 (6.176)			6.600 (6.511)	7.036 (6.540)
3-way interactions								
Non-CLIL*Years PE*freq PE								-3.592 (5.548)
Random part								
School variance	279.502 (192.813)	275.677 (190.895)	256.508 (181.712)	237.339 (171.729)	253.443 (179.790)	232.404 (166.488)	218.325 (159.616)	216.608 (159.098)

Table 4.27: Results of multilevel models for speaking fluency (number of words in English): Ancova, variables moderating the effect of CLIL. Standard errors in parentheses. (PE=Primary school English) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Class variance	97.336 (85.298)	98.375 (85.592)	105.197 (87.764)	108.657 (88.506)	105.023 (87.418)	96.653 (83.373)	99.871 (84.456)	102.492 (85.241)
Pupil variance	1743.427 (151.579)	1742.318 (151.488)	1731.149 (150.810)	1721.753 (150.014)	1727.614 (150.526)	1689.445 (147.197)	1682.980 (146.609)	1679.427 (146.303)
Total variance	2120.265	2116.370	2092.854	2067.749	2086.080	2018.502	2001.176	1998.527
Prop expl school var						.083		
Prop expl class var						.080		
Prop expl pupil var						.022		
Prop expl total var						.032		
Deviance	3000.263	3000.091	2988.111	2986.465	2987.473	2980.230	2979.102	2978.684
Model of reference and fit improvement		model 1 X^2=.172 df=1 p=n.s.		model 3 X^2=1.646 df=1 p=n.s.		model 5 X^2=7.243 df=1 p<.01	model 6 X^2=1.128 df=2 p=n.s.	model 7 X^2=.418 df=1 p=n.s.
N	Npupils=289; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7					

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.28: Results of multilevel models for speaking fluency (number of words in English): change scores, variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score)

Model	1	2	3	4	5	6	7	8	9
Fixed part	**Grade-1/ Pre-voc-level-rec-1**	**Grade-2**	**Pre-voc-level-rec-2**	**Gender-1**	**Gender-2**	**Cito-1**	**Cito-2**	**Home language-1**	**Home language-2**
Intercept	49.982 (9.071)	52.857 (9.716)	50.258 (9.090)	49.685 (9.219)	49.985 (9.456)	45.180 (9.320)	45.266 (9.337)	49.878 (9.096)	49.879 (9.104)
non-CLIL (0=clil; 1=non-clil)	-13.969 (8.879)	-21.757# (11.293)	-15.226 (10.115)	-14.035 (8.866)	-14.690 (10.049)	-5.185 (9.908)	-7.893 (9.976)	-14.078 (8.933)	-14.078 (8.984)
Grade 8 (grade 7=ref. cat.)	-22.684* (9.408)	-24.943* (11.316)	-22.613* (9.416)	-22.833* (9.433)	-22.908* (9.452)	-32.252** (11.146)	-31.708* (11.047)	-22.336* (9.499)	-22.337* (9.506)
Grade 9	-9.318 (10.603)	-26.320# (13.305)	-9.566 (10.600)	-9.741 (10.603)	-9.810 (10.620)	-1.856 (12.419)	-2.750 (12.345)	-9.452 (10.669)	-9.452 (10.708)
Pre-voc level (gm) (1=high: 5=low)	-9.240# (4.845)	-8.408# (4.737)	-8.630 (5.458)	-9.326# (4.868)	-9.338# (4.868)	-13.984* (5.682)	-14.178* (5.679)	-9.250# (4.869)	-9.2516# (4.871)
Gender (boy=1; girl=0)				.848 (5.234)	.219 (6.959)				
Cito (gm)						-.065 (.678)	-.880 (.795)		
Home language (gm)								-.433 (1.951)	-.433 (2.512)
2-way interactions									
Non-CLIL*grade8		7.018 (16.288)							
Non-CLIL*grade9		34.668# (18.770)							
Non-CLIL*pre-voc level			-1.998 (7.919)						
Non-CLIL*gender (boy)					1.430 (10.445)				
non-CLIL*cito-gm							2.300# (1.202)		

Table 4.28: Results of multilevel models for speaking fluency (number of words in English): change scores, variables moderating the effect of CLIL. Standard errors in parentheses. (Cito = scholastic aptitude test score) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Non-CLIL*home language									-.001 (3.926)
Random part									
school variance	263.569 (193.449)	307.340 (207.436)	259.616 (191.727)	262.729 (193.069)	262.005 (192.753)	182.256 (170.111)	190.931 (173.343)	262.606 (193.528)	262.665 (194.088)
Class variance	150.341 (103.688)	93.624 (84.801)	150.425 (104.100)	148.529 (103.466)	148.945 (103.601)	133.821 (142.617)	132.332 (139.974)	153.583 (105.044)	153.757 (105.560)
Pupil variance	1769.842 (153.975	1770.077 (153.897)	1769.899 (153.933)	1770.525 (153.988)	1770.324 (153.970)	1853.974 (216.480)	1811.023 (211.491)	1773.209 (154.567)	1773.130 (154.504)
Total variance	2183.752	2171.041	2179.940	2181.783	2181.274	2170.051	2134.286	2189.398	2189.552
prop. expl. var. school level							-		
prop. expl. var. class level							.011		
prop. expl. var. Pupil level							.023		
prop. expl. var. total							.016		
Deviance	3007.769	3004.694	3007.705	3007.743	3007.724	1777.753	1774.138	2998.116	2998.116
Model of reference and fit improvement		model 1 X^2=3.075 df=2 p=n.s.	model 1 X^2=.064 df=1 p=n.s.		model 4 X^2=.019 df=1 p=n.s.		model 6 X^2=3.615 df=1 p<.10		model 8 X^2=.000 df=1 p=n.s.
	Npupils=289; Nclasses=25; Nschools=7					Npupils=170; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7	

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Table 4.29: Results of longitudinal multilevel models for speaking fluency (number of words in English): change scores, variables moderating the effect of CLIL. Standard errors in parentheses. (PE=Primary school English)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Years PE**	**Years PE**	**Freq. PE**	**Freq. PE**	**Years & freq: PE intensity**	**Years & freq: PE intensity**	**Years & freq: PE intensity**	**Years & freq: PE intensity**
Intercept	50.210 (9.101)	50.292 (9.062)	49.887 (8.972)	50.551 (8.774)	50.635 (8.966)	52.763 (8.724)	53.175 (8.567)	52.694 (8.591)
non-CLIL (0=clil; 1=non-clil)	-14.083 (8.881)	-14.152 (8.875)	-13.130 (8.988)	-11.843 (8.920)	-13.184 (8.978)	-14.648 (8.728)	-13.534 (8.744)	-12.507 (8.879)
Grade 8 (grade 7=ref. cat.)	-23.057* (9.469)	-22.797* (9.488)	-22.195* (9.552)	-22.549* (9.466)	-22.330* (9.613)	-21.346* (9.444)	-21.567* (9.387)	-21.465* (9.435)
Grade 9	-9.985 (10.652)	-10.178 (10.656)	-10.112 (10.729)	-11.515 (10.652)	-11.578 (10.817)	-8.682 (10.636)	-9.813 (10.587)	-9.933 (10.630)
Pre-voc level (gm) (1=high: 5=low)	-9.309# (4.850)	-9.340# (4.835)	-9.779# (4.828)	-10.253* (4.739)	-10.208* (4.829)	-9.774# (4.696)	-10.135* (4.625)	-10.027* (4.632)
Years Pri.English (gm)	-1.002 (3.011)	-2.235 (3.705)			-3.391 (3.310)	-3.184 (3.275)	-2.935 (4.203)	-3.087 (4.202)
Frequency Pri.English (gm)			3.994 (2.918)	1.115 (3.470)	5.383# (3.211)	6.364* (3.195)	3.814 (3.954)	3.651 (3.956)
2-way interactions								
Years PE*freq PE						-6.613** (2.492)	-6.396* (2.509)	-5.202# (2.920)
Non-CLIL*Years PE		3.589 (6.286)					.261 (6.718)	-.192 (6.735)
Non-CLIL*Freq. PE				9.523 (6.236)			7.917 (6.586)	8.420 (6.610)
3-way interaction								
Non-CLIL*Years PE*freq. PE								-4.456 (5.605)
Random part								
School variance	264.351 (193.767)	259.282 (191.143)	243.477 (184.374)	225.058 (173.571)	239.523 (182.281)	219.854 (169.497)	205.814 (161.272)	202.561 (160.061)

Table 4.29: Results of longitudinal multilevel models for speaking fluency (number of words in English): change scores, variables moderating the effect of CLIL. Standard errors in parentheses. (PE=Primary school English) (*continued*)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Class variance	149.803 (103.491)	150.360 (103.649)	159.621 (106.841)	155.031 (104.982)	160.233 (106.864)	150.677 (102.387)	147.046 (100.977)	150.605 (102.015)
Pupil variance	1769.251 (153.923)	1767.565 (153.773)	1761.844 (153.562)	1751.335 (152.639)	1755.380 (152.997)	1717.774 (149.725)	1710.889 (149.116)	1705.964 (148.690)
Total variance	2183.405	2177.207	2164.942	2131.424	2155.136	2088.305	2063.749	2059.130
prop. expl. var. school level						.082		
prop. expl. var. class level						.060		
prop. expl. var. Pupil level						.021		
prop. expl. var. total						.031		
Deviance	3007.658	3007.333	2996.355	2994.043	2995.308	2988.361	2986.704	2986.074
Model of reference and fit improvement		model 1 X^2=.325 df=1 p=n.s.		model 3 X^2=2.312 df=1 p=n.s.		model 5 X^2=6.947 df=1 p<.01	model 6 X^2=1.657 df=1 p=n.s.	model 7 X^2=.630 df=1 p=n.s.
N	Npupils=289; Nclasses=25; Nschools=7		Npupils=288; Nclasses=25; Nschools=7					

#=sig at 10% (=5% one sided); *=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non-significant)

Appendix 5

Attachment 5.1: **CLIL teacher questionnaire**

Research Questionnaire TTO subject and English teachers (answer in English or in Dutch!)

This questionnaire is designed to gather information about the TTO situation and practice at T-vmbo schools in the Netherlands. All of your answers will remain anonymous in the data analysis and in any discussions or publication of results. Please return your paper questionnaire to your CLIL coordinator, or digitally to J.L.Denman@XX.nl Please write your name only so we can keep track of which teachers and which schools have returned the questionnaires. Your name will be deleted for data entry.

Name_____School_____
Teacher of subject(s) _____
Mother tongue_____ Other languages _____
How many lessons do you teach per week (**total**)? _____ How many **TTO** lessons per week? _____
Which **TTO** class groups did you teach this year?: _____
TTO level(s) you teach (circle all that apply): **mavo/havo mavo/vmbo-t vmbo-t vmbo-t/gemengd gemengd gemengd/kader/basis kader kader/basis basis**
Number of years you have been a teacher: _____ Number of years you have been teaching vmbo: _____ Number of years you have been teaching **TTO:** _____
Have you taught TTO at havo/vwo level? _____ If yes, for how many years? _____
Have you had any specific TTO / CLIL training? If so, what kind and where?
_____

Please estimate/circle your English level on the ERK/CEFR scale. (B2 is approximately *havo* end level.)

A1 A2 B1 B2 C1 C2 native speaker

If you have a Cambridge/Anglia/IELTS English language certificate, which one/which level?
_____

In your opinion, what are the most **positive aspects** of teaching TTO?
_____
_____

In your opinion, what are the most **negative aspects** of teaching TTO?

_____

_____

What **two things** would make teaching TTO even more effective for you as a teacher, or for your pupils, or for your school?

_____

_____

_____

What percent of the time do you use English for these different activities? Tick one box on each line.					
Note that each different activity (each line) represents a percentage from 0 – 100%.	**80 – 100%**	**60 – 80%**	**40 – 60%**	**20 – 40%**	**0 – 20%**
Of all the time spent on **classroom management** *(orde houden)*, what percentage do you do **in English**?					
Of all the time spent on **giving instructions and directions**, what percentage do you do **in English**?					
Of all the time spent on **lesson content and subject teaching**, what percentage do you do **in English**?					
Of all the time spent on **attention to language and corrective feedback**, what percentage do you do **in English**?					
Of all the time spent on **personal interaction with pupils**, what percentage do you do **in English**?					

Attachment 5.2: **Focus group interview protocol**

Vragen voor focus group interview SCHOOL_____KLAS_____

1. Hoe heb je besloten om naar een TTO-school te gaan? (Of heb jij zelf niet gekozen, maar je ouders?)
   Wat waren de belangrijkste redenen voor jou om een TTO-school te kiezen?
   Hadden jullie al Engels op de basisschool?
2. In welke vakken hebben jullie les in het Engels?
   Zijn ze allemaal even moeilijk/makkelijk in het Engels, of is er een verschil? Wat is het verschil?
3. Hoe werkt TTO in de les? Praat de leraar altijd Engels? wanneer wel, wanneer niet?
   Praten de leerlingen altijd Engels? Met de leraar? Met elkaar? Hoe vinden jullie dat? Zijn er regels voor?
   Hoeveel Engels praat je op school, gemiddeld? (Hoeveel woorden ongeveer op een normale dag?)
4. Als je een les hebt in het Engels, wat doet de leraar dan om de taal begrijpelijk te maken?
5. Wat vind je het moeilijkst en het makkelijkst in het Engels, en waarom:
   (prompt: lezen, schrijven, luisteren, spreken, anders)
6. Werken de TTO-leraren samen, of de leraar Engels samen met de TTO-leraren, volgens jou? Hoe?
7. Is er een verschil tussen de TTO-lessen en de gewone lessen, behalve het Engels? Hoe?
8. Wat doen jullie aan EIO/internationale orientatie? (reizen, excursies, contact met buitenlandse klassen of leerlingen, bezoek op school)
9. Wat moet een TTO-leraar heel goed kunnen, volgens jou? (anders dan een niet-TTO docent)
10. Kun je iets bedenken dat een TTO-leraar doet dat je niet goed vindt?
11. Kun je iets uitleggen dat een TTO-leraar doet dat je juist heel goed vindt?
12. Als jij de leraar was voor een TTO-vak in het Engels, wat zou jij zelf doen om de TTO-les heel boeiend én begrijpelijk te maken?
13. Denk je dat de TTO-leraren meer Engels zouden moeten spreken? (deze vraag toegevoegd naar aanleiding van opmerking over weinig Engels)

Algemene vervolgvragen: Kun je daar een voorbeeld van geven? Kun je dat beschrijven? Ben jij/zijn jullie ermee eens? Wil je nog iets toevoegen?

Attachment 5.3: **Codebook** for the thematic content analysis of the focus group interview transcriptions: 70 pupils from 15 classes in 6 schools

Categories	Definition	THEME	Sub-categories	Codes
1. CLIL specific subjects	Which CLIL subjects are mentioned, and what do pupils say about EIO?	**CLIL content subjects (not EFL lessons)**	Variety of subjects offered in the 6 schools Relative difficulty/ease EIO (special status)	CLIL subjects considered more difficult CLIL subjects considered easier EIO activities – positive EIO activities - neutral EIO activities - negative
2. English lessons & skills	What do CLIL pupils say about their EFL lessons and specific English skills?	**Learning English – EFL lessons and skills**	English lessons Specific English skills	English lessons - general Speaking Listening Reading Writing Vocabulary Grammar
3. Teachers	What do CLIL pupils say about their CLIL teachers' didactic/ pedagogic/interpersonal behavior?	**Teacher pedagogical skills and behavior**	Positive strategies/behavior Negative teacher behavior Teacher collaboration	Strategies used by teachers to make content/ language more accessible (20 different strategies named,e.g.: *word strategies: synonyms, descriptions, summaries *use English before translating/avoid Dutch where possible *give multimodal imput (write it, speak it, visual aids, realia, smartboard, apps) *slow speaking tempo, repetition *sign language, acting, mime *corrective feedback *formative assessment *extra explanation with examples *individual attention *fun educational games Frustrating teacher behavior: Interpersonal Frustrating teacher behavior: lesson/language Perception of teacher collaboration

Categories	Definition	THEME	Sub-categories	Codes
4. Peers	What do CLIL pupils say about their CLIL peers?	**Pupil peer behavior**	English level Frustration Class (group) dynamics	Insufficient English level Pupils not serious Pupils refuse to speak English Class is more powerful than teacher
5. CLIL general	What do the pupils say about CLIL in the lessons?	**Aspects of CLIL methodology**	Differences between CLIL lessons and regular lessons CLIL - general	Lesson activities Materials Curricular integration Teacher behavior No difference CLIL general comments
6. Use of L1 vs L2 in the classroom	What do pupils say about the use of Dutch vs. English?		Teacher use of L1/L2 & pupil use of L1/L2	Frustration: not enough L2 Frustration: too much L2 Non-Dutch speaking teacher issues
7. Policy/logistics/ administration	What comments do pupils make regarding CLIL policies?		Year 4 exams in Dutch CLIL program promises at the outset	Dutch exams in year 4 - negative CLIL program promises not kept – e.g. trips, contacts, not enough English
8. Suggestions for improvement	What skills do pupils say a good CLIL teacher should have and what are the pupils' suggestions for improvement?	**Suggestions for improvement**	Basic skills for a pre-voc CLIL teacher (some overlap with 'Improvements') – no particular order after the first two.	*good English, with no 'annoying' Dutch accent; know Dutch *good explanations; allow questions & repetition *give corrective feedback *obvious subject competence *personable/friendly/kind/ patient *personal stories/anecdotes
			Improvements: lesson activities	*DO-activities: moving, go outside * make things, creative * more fun and games *multimodal presentation: visual material, films, realia, music and lyrics, IWB *country and culture info
			Improvements: L1/L2 language use	*let pupils speak English more *make a strict rule: English only *speak English consistently/speak less Dutch *be able to speak enough Dutch to explain

Categories	Definition	THEME	Sub-categories	Codes
			Improvements: teacher didactic behavior	*less Teacher Talking Time *tell stories and jokes *clear lesson structure, lesson goals, summaries *differentiation, mixed groups, more group work *don't leave anyone behind; explain thoroughly
			Improvements: teacher pedagogic behavior	*be patient, don't get angry *be nice, friendly, helpful, persistent *be interesting, show personality *be active and energetic

Appendix 6

Attachment 6.1: **Practitioner and pupil recommendations** for strengthening CLIL in *vmbo* (summarized from Chapter 5*: 'CLIL can be improved'* theme). This attachment complements the key recommendations presented in Section 6.5D.

Classroom level

- Increase the consistent use of English, while ensuring comprehensibility through scaffolding (visuals, modelling, structured tasks, peer support).
- Provide more interactive, hands-on, and creative activities tailored to *vmbo* learners as 'doers' (e.g., role-plays, projects, excursions).
- Use authentic materials and realia to make lessons relevant and engaging.
- Offer individualized feedback to support both language and content learning.
- Maintain clear language modeling and avoid confusing code-switching within utterances.

School/organizational level

- Ensure that prospective CLIL pupils and their parents are well-informed about what CLIL entails, so that the choice for the CLIL stream is conscious and deliberate rather than incidental.
- Allocate time for collaboration between subject and English teachers in planning and lesson design.
- Ensure that internationalization activities (projects, exchanges, excursions) are reliably implemented, highly attractive, and accessible to all, as pupils attach high value to these.
- Provide sustained CLIL-specific professional development for teachers, combining language training with methodological training adapted to weaker learners.
- Support teacher well-being by recognizing and addressing workload demands of CLIL.
- Ensure that the teachers chosen to teach in the CLIL program fully support the school's goals for CLIL.

Policy level

- While the Year 3 switch back to Dutch for exam subjects is unlikely to change, policymakers and school leaders could extend the current CLIL opportunities in non-exam subjects (e.g., arts, PE, electives, or projects) to maintain target language input throughout the school career.
- Consider developing optional CLIL subject examinations in English, alongside the required Dutch exams, for pupils who wish to demonstrate their bilingual competence formally.
- Provide dedicated funding and frameworks for CLIL in vocational/pre-vocational education, recognizing that these contexts have distinct learner needs compared to academic tracks.
- Involve pupils as stakeholders in CLIL policy development, given the educational merit and practicality of their suggestions.

Nederlandse Samenvatting

1. Achtergrond en probleemstelling

In de afgelopen drie decennia is Content and Language Integrated Learning (CLIL) uitgegroeid tot een van de meest zichtbare vormen van tweetalig onderwijs in Europa. Terwijl vroeg onderzoek (Admiraal et al., 2006) vooral de successen van CLIL in academisch selectieve leerwegen belichtte, bleef grotendeels onbekend of en hoe CLIL ook zou kunnen werken voor leerlingen met een lagere schoolse belastbaarheid. In het sterk gedifferentieerde Nederlandse onderwijssysteem ontstond tweetalig onderwijs (TTO) met de CLIL-aanpak in de jaren negentig als een curriculaire innovatie die vooral gemotiveerde, academisch sterke leerlingen in vwo en later havo moest aantrekken. Het voorbereidend middelbaar beroepsonderwijs (vmbo)—goed voor ongeveer de helft van de Nederlandse middelbare-schoolpopulatie—bleef daarentegen vrijwel afwezig binnen het tweetalige aanbod.

Deze scheefgroei riep een bredere ethische en pedagogische vraag op: als CLIL de motivatie en de Engelse taalontwikkeling van leerlingen kan bevorderen, is het dan rechtvaardig om deze benadering alleen te reserveren voor academisch selectieve leerwegen? Met andere woorden: kan CLIL zowel talig haalbaar als motivationeel duurzaam zijn voor minder academisch georiënteerde leerlingen? Dit proefschrift onderzocht die vraag door na te gaan of de kernmechanismen van CLIL—betekenisvolle blootstelling (exposure) aan Engels, een dubbele focus op vakinhoud en taal, en actieve leerlingparticipatie—ook effectief kunnen werken in niet-selectieve onderwijscontexten met lagere aanvangsniveaus.

Omdat de gegevens zijn verzameld tussen 2012 en 2014, toen vmbo-CLIL zich nog in een beginfase bevond, documenteert het project de eerste generatie scholen die met deze aanpak experimenteerden. Deze programma's werkten op basis van de eerste landelijke standaarden voor TTO op vmbo-niveau (Nuffic, n.d.-b; Europees Platform, 2009), waarin eindtermen waren aangepast aan de realistische talige en cognitieve mogelijkheden van vmbo-leerlingen. Veel leerlingen hadden een meertalige achtergrond en weerspiegelden de diversiteit die kenmerkend is voor Nederlandse vmbo-scholen.

De studie plaatst het Nederlandse vmbo-experiment daarmee binnen bredere Europese discussies over inclusiviteit en onderwijsgelijkheid in tweetalig onderwijs (Pérez Cañado, 2023), met als doel te verkennen wat haalbaar is voor leerlingen die veelal uitgesloten zijn van academisch georiënteerde vormen van CLIL.

2. Potentiële relevantie en onderzoekslacune

Lager-presterende CLIL-leercontexten zoals het vmbo zijn schaars onderzocht en blijven ondervertegenwoordigd in de CLIL-literatuur, die grotendeels afkomstig is uit selectieve

of hogerpresterende onderwijsomgevingen. Deze scheve onderzoeksbasis heeft het zicht beperkt op de vraag in welke mate CLIL kan bijdragen aan onderwijsgelijkheid over het volledige spectrum van het voortgezet onderwijs. Dit proefschrift levert daarom relevantie op vijf niveaus.

Ten eerste biedt het proefschrift theoretische relevantie door kernprincipes van CLIL—zoals taal van/voor/door het leren (language of/for/through learning) en het 4Cs-framework van inhoud, communicatie, cultuur, en context (content, communication, culture, context) (Coyle, 2007)—toe te passen op een leerlingpopulatie die tot nu toe nauwelijks is onderzocht, en door na te gaan of deze principes houdbaar blijven wanneer cognitieve en talige eisen moeten worden aangepast.

Ten tweede levert het proefschrift beleidsrelevantie door bevindingen op klasniveau te verbinden met nationale ontwikkelingen rond tweetalig onderwijs. Door vmbo-CLIL te bestuderen binnen aangepaste eindtermen laat de studie zien hoe rechtvaardigheid en realistische voortgang kunnen samengaan wanneer verwachtingen worden afgestemd op leerweg- en niveau-specifieke uitkomsten .

Ten derde toont het proefschrift praktische relevantie voor scholen, lerarenopleidingen en beleidsmakers die duurzame CLIL-routes willen ontwikkelen voor heterogeen samengestelde leerlinggroepen. De bevindingen maken duidelijk onder welke voorwaarden CLIL effectief kan zijn in de praktijkgerichte onderwijstrajecten zonder dat academische voorselectie noodzakelijk is.

Ten vierde heeft het proefschrift maatschappelijke relevantie, omdat vmbo-CLIL opereert binnen een breder onderwijslandschap waarin de vmbo-sector nog steeds te maken heeft met een hardnekkig negatief publiek imago. Deze maatschappelijke context—zichtbaar in nationale debatten en media—staat in contrast met herhaalde signalen van werkgevers en onderwijsexperts dat de praktijkgerichte competenties van vmbo-leerlingen essentieel zijn voor de Nederlandse samenleving. Door de haalbaarheid van CLIL juist binnen deze context te onderzoeken, laat het proefschrift zien hoe functionele beheersing van Engels en internationale oriëntatie vmbo-leerlingen toegang kunnen bieden tot talig en intercultureel kapitaal dat in onderwijs, werk en burgerschap steeds waardevoller wordt.

Ten vijfde levert het proefschrift empirische relevantie door de eerste longitudinale meting van L2-ontwikkeling en attitudinale groei in vmbo-CLIL te presenteren, evenals de eerste systematische analyse van leerlingperspectieven (pupil voice) en stakeholderperspectieven in deze context. Deze meting ondersteunt toekomstige evaluaties van het Nederlandse TTO-veld en vormt een basis voor vergelijkend onderzoek, waarmee de internationale empirische onderbouwing van inclusieve vormen van CLIL wordt versterkt.

3. Theoretisch kader

Het conceptuele fundament van dit proefschrift steunt op drie onderling verbonden domeinen: CLIL-pedagogiek, sociaal-cognitieve modellen van leermotivatie, en internationaal onderzoek naar diversiteit en onderwijsgelijkheid binnen tweetalig onderwijs.

Binnen de CLIL-pedagogiek vormt het 4Cs-framework (Coyle, 2007) een centraal uitgangspunt, waarin content, communication, cognition en culture geïntegreerd worden. Leerlingen verwerven vakinhoud en taal gelijktijdig via betekenisvol, contextueel ingebed taalaanbod. In de vmbo-context vereist CLIL echter een zorgvuldige afstemming van deze elementen: taken moeten cognitief uitdagend blijven, maar tegelijkertijd voldoende talige scaffolding bieden om begrip en actieve deelname te ondersteunen.

Het tweede theoretische spoor is gebaseerd op Ajzen's (1991) model van gepland gedrag (Model of Planned Behavior – MPB), dat verklaart hoe attitudes, waargenomen handelingsmogelijkheden (*perceived behavioural control*) en sociale normen intentioneel gedrag mede vormgeven. Dit model ligt ten grondslag aan de analyse van attitudinale haalbaarheid in dit proefschrift: de vraag of CLIL duurzame, positieve attituden kan versterken bij leerlingen die vaak beperkte eerdere schoolse ervaring met Engels hebben.

Ten slotte sluit de studie aan bij een internationale discussie over diversiteit, onderwijsgelijkheid en inclusie binnen tweetalig onderwijs, en over de potentie van immersie- en CLIL-programma's voor uiteenlopende leerlingpopulaties. Onderzoek van onder meer Genesee en Fortune (2014), Harrop (2012), Somers (2017) en Pérez Cañado (2023) betwist de aanname dat tweetalig onderwijs alleen geschikt is voor hoogpresterende of talig bevoorrechte leerlingen. Deze onderzoekers laten zien dat zorgvuldig ontworpen en ondersteunde CLIL-omgevingen cognitieve, talige en motivationele voordelen kunnen bieden aan leerlingen met zeer verschillende profielen, waaronder leerlingen met lagere schoolse capaciteiten of een meertalige thuisomgeving.

In dit proefschrift vormt deze literatuur de basis voor een tweevoudige haalbaarheidslens:

- talige haalbaarheid – haalbare en meetbare vooruitgang in Engelse taalvaardigheid, en
- attitudinale haalbaarheid – de mate waarin leerlingen en docenten CLIL ervaren als motiverend, beheersbaar en de moeite waard.

4. Populatie en context

Het onderzoek vond plaats binnen het voorbereidend middelbaar beroepsonderwijs (vmbo), het grootste segment van het Nederlandse voortgezet onderwijs, dat ongeveer de helft van alle leerlingen bedient binnen vier leerwegen. Deze leerwegen bereiden

leerlingen voor op een mbo-route en latere arbeidsmarktdeelname, en omvatten een brede variatie aan schoolse capaciteiten en talige achtergronden.

Tijdens de onderzoeksperiode bevonden vmbo-CLIL-programma's zich in een vroege, experimentele fase. Sommige deelnemende scholen waren pioniers met meerdere jaren ervaring, terwijl andere juist begonnen met het invoeren van CLIL in één specifieke jaarlaag of cohort. Deze variatie tussen scholen maakte het mogelijk om programma's met verschillende niveaus van expertise, continuïteit en organisatorische ondersteuning met elkaar te vergelijken.

Vmbo-CLIL functioneert binnen aangepaste eindtermen (Nuffic, n.d.-b), die de verwachtingen afstemmen op haalbare voortgang voor vmbo-leerlingen. Het aandeel Engelstalige lessen is doorgaans lager dan in havo en vwo, en de nadruk ligt op communicatieve vaardigheid, woordenschatuitbreiding en functionele grammatica op Europees Referentiekader (ERK) niveaus A2 tot B2 (afhankelijk van vaardigheid en leerweg), eerder dan op beheersing van academische taal.

In totaal namen ongeveer 600 leerlingen deel aan de vier empirische deelstudies, verdeeld over 25 klassen in zeven scholen, samen met 53 docenten Engels en vakdocenten binnen de tweetalige trajecten. De leerlingenpopulatie weerspiegelde de meertalige diversiteit die kenmerkend is voor veel vmbo-scholen: in totaal werden 44 thuistalen genoemd, waarvan Nederlands, Turks en Marokkaans Arabisch het meest voorkwamen.

Deze heterogeniteit maakte de onderzoeksetting bijzonder geschikt voor het bestuderen van inclusieve CLIL-praktijken. Het gezamenlijke doel van de programma's — een evenwicht bereiken tussen begrijpelijkheid en uitdaging — bood een realistische toetsing van de kernprincipes van CLIL in een niet-selectieve onderwijscontext.

5. Onderzoeksopzet en methodologie

Het proefschrift hanteerde een mixed-methods onderzoeksopzet met meerdere deelstudies, waarin longitudinale kwantitatieve analyses werden gecombineerd met aanvullende kwalitatieve gegevens. Elke deelstudie onderzocht een afzonderlijk aspect van haalbaarheid—talig, attitudinaal of perceptueel—zodat zij samen een samenhangende evaluatie vormden van hoe CLIL werkt binnen vmbo-contexten.

5.1 Onderzoeksdesign en procedure

De gegevens werden verzameld uit een gelegenheidssteekproef (convenience sample) van meerdere cohorten in leerjaar 1, 2 en 3, zowel in CLIL- als in niet-CLIL-klassen. Het onderzoeksdesign combineerde ecologische validiteit—werken met intacte schoolprogramma's—met de noodzaak tot statistische vergelijkbaarheid tussen groepen en over tijd. In Studies 1–3 maakte de longitudinale opzet het mogelijk om ontwikkeling in de tijd

te analyseren, terwijl Studie 4 gebruikmaakte van cross-sectioele gegevens om percepties van betrokkenen binnen vmbo-CLIL-programma's te onderzoeken.

5.2 Instrumenten en dataverzameling
De kwantitatieve instrumenten bestonden uit woordenschat- en grammaticatoetsen, schrijf- en spreekopdrachten, een 'bereidheid om te communiceren' (*Willingness to Communicate,* WtC) vragenlijst, en een zelfrapportage-instrument waarin docenten hun eigen gebruik van de doeltaal inschatten. Attitudinale vragenlijsten, gebaseerd op Ajzen's Model of Planned Behavior, maten de constructen cognition, affect, subjective norm, perceived behavioural control en intention met betrekking tot het leren van Engels.

Het kwalitatieve onderdeel bestond uit leerlingvragenlijsten, leerling-focusgroepinterviews en open-eindreflecties van docenten, gericht op klaspraktijk, motivatie en programmatische implementatie.

5.3 Analytische procedures
De statistische analyses waren afgestemd op de aard van de verschillende datasets:
- Confirmatory factor analysis (CFA) en structural equation modeling (SEM) valideerden het MPB-meetmodel en onderzochten de relaties tussen attitudinale constructen, aangevuld met multilevelanalyses voor verschillen tussen groepen (Hoofdstuk 2).
- Multilevel growth modelling analyseerde de ontwikkeling van woordenschat en grammatica over tijd (Hoofdstuk 3).
- ANCOVA en change-score analyses, aangevuld met multilevel regressiemodellen, onderzochten schrijf-, spreek- en WtC-uitkomsten (Hoofdstuk 4).
- Thematic analysis (Braun & Clarke, 2006, 2022) werd toegepast op kwalitatieve gegevens van leerlingen en docenten (Hoofdstuk 5).

5.4 Integratie
Door statistische en thematische benaderingen te combineren konden de kwantitatieve bevindingen worden geïnterpreteerd in samenhang met de ervaringen van leerlingen en docenten. Zo biedt het proefschrift een geïntegreerd beeld van de talige en attitudinale dimensies van vmbo-CLIL.

6. Uitkomsten van de deelonderzoeken

Studie 1 (Hoofdstuk 2): Attitudinal factors and intention to learn English
Hoofdstuk 2 onderzoekt hoe de attitudes van vmbo-leerlingen ten aanzien van het leren van Engels zich ontwikkelden over drie meetmomenten gedurende twee schooljaren, in zowel CLIL- als niet-CLIL-klassen. Hiervoor werd het Model of Planned Behavior (MPB) (Ajzen, 1991) toegepast op een steekproef van 488 leerlingen in leerjaar 1–3. De attitudinale constructen werden gemeten met een gestandaardiseerde vragenlijst die was afgestemd

op het MPB-raamwerk. De analyses combineerden ontwikkelingsgerichte vergelijkingen (change scores over de drie meetmomenten) met structurele modellering binnen het MPB-kader. Confirmatory factor analyses bevestigden een goede modelfit en toonden meetinvariatie aan tussen CLIL- en niet-CLIL-groepen.

Binnen dit MPB-model bleek het construct cognitie - het waargenomen nut van inspanning leveren om Engels te leren - de sterkste voorspeller van de intentie om die inspanning daadwerkelijk te leveren. Het structurele model liet verder zien dat affect – de gevoelens van de persoon over het uitvoeren van het gedrag - en subjective norm – de morele ideeën van de persoon over de wenselijkheid van het gedrag – sterk correleerden met zowel cognitie als intentie, maar geen extra verklaarde variantie toevoegden zodra voor cognition werd gecontroleerd. Waargenomen gedragscontrole (perceived behavioural control; PBC) – de mate waarin de persoon belemmeringen ervaart bij het uitvoeren van het gedrag - vertoonde een zwakke correlatie met intentie, en het directe pad in het structurele model was niet significant, wat erop wijst dat waargenomen obstakels slechts een beperkte onafhankelijke invloed hadden op de attitudinale uitkomsten. Hoewel correlaties en SEM-paden uit verschillende analytische stappen komen, richten zij zich op verwante modelaspecten en leiden zij tot eenzelfde interpretatie.

CLIL-leerlingen scoorden significant hoger dan niet-CLIL-leerlingen op vier van de vijf MPB-constructen—cognitie, affect, PBC en intentie. Deze verschillen werden vastgesteld via multilevelanalyses van variantie op klasniveau. Ongeveer 40% van de tussen-klassenvariantie in cognitie, affect en intentie, en 25% van de variantie in PBC, hield verband met CLIL-klassen. Belangrijk is dat beide groepen aan het begin van de brugklas vergelijkbaar waren (cohort 1): noch cognitie noch affect verschilde significant op baseline, wat aangeeft dat het CLIL-voordeel zich tijdens de schoolperiode ontwikkelde en niet het gevolg was van voorselectie. Het verschil in affect bereikte een piek in leerjaar 2 en nam weer af aan het einde van leerjaar 3, parallel aan een vermindering in Engelstalige instructie. Een klein maar aanhoudend voordeel in PBC suggereert dat winst in zelfeffectiviteit en autonomie langer behouden bleef, ook na afname van programma-intensiteit.

Gezamenlijk wijzen deze bevindingen erop dat deelname aan CLIL zowel de waargenomen waarde als het plezier in het leren van Engels versterkte, met name tijdens de intensieve eerste twee leerjaren van het programma. Hoewel het affectieve voordeel afnam naarmate de taalblootstelling kleiner werd, bleven de motivationale voordelen die samenhangen met waargenomen nut en controle bestaan. Al met al duidt dit patroon erop dat CLIL-deelname de attitudinale componenten versterkte die bepalend zijn voor de intentie om Engels te leren, ook bij leerlingen met minder gunstige schoolse uitgangsposities.

Studie 2 (Hoofdstuk 3): Woordenschat (receptief) en grammatica (gecontroleerde productie)
De resultaten laten zien dat CLIL-leerlingen significant hoger scoorden dan hun niet-CLIL-peers in zowel woordenschat als grammatica, na correctie voor klas- en schoolverschillen. Woordenschat werd gemeten met een aangepaste woordenschattoets; grammatica werd gemeten via een lexicogrammaticale invultoets waarin leerlingen de ontbrekende grammaticale vorm moesten produceren op basis van syntactische en contextuele informatie. Dit betreft dus gecontroleerde grammaticale productie, niet louter receptieve kennis.

Op het eerste meetmoment bedroeg het verschil 5,18 punten (max. 48) voor woordenschat en 2,0 punten (max. 36) voor grammatica. CLIL-leerlingen breidden deze voorsprong gedurende twee jaar verder uit. Groeianalyses toonden aan dat het voordeel in woordenschat stabiel en sterker was: CLIL-leerlingen wonnen gemiddeld 0,016 punten per week meer dan niet-CLIL-leerlingen, wat neerkomt op 0,8% van de totale variantie in woordenschat-groei en bijna 7% van de variantie tussen klassen. De groei in grammatica was bescheidener (0,5% van de totale variantie; 3,7% van de groei-variantie). De sterkste vooruitgang trad op in de eerste twee jaren van het programma, de periode met de meeste CLIL-taalblootstelling.

Er waren daarnaast kleine, verwachte effecten van leerjaar en vmbo-leerweg: leerlingen in hogere leerwegen en in leerjaar 2 scoorden gemiddeld iets hoger. Zoals te verwachten viel, gingen leerlingen in hogere leerwegen sneller vooruit. Tegelijkertijd liet een moderatie-effect zien dat het relatieve CLIL-voordeel in grammatica juist iets groter was in de lagere leerwegen, ook al bleven de absolute scores daar lager. In het derde cohort daalden de grammaticascores van beide groepen licht, parallel aan een daling in CLIL-intensiteit.

Samengevat was CLIL-deelname geassocieerd met kleine maar statistisch betrouwbare voordelen in zowel woordenschat (receptief) als grammatica (gecontroleerde productie). Vooral woordenschatgroei was duidelijk gekoppeld aan perioden met hogere blootstelling aan Engelstalige instructie. Deze patronen laten zien dat blootstelling en taalaanbod sterk bepalend waren voor de ontwikkeling van leerlingen binnen vmbo-CLIL.

Studie 3 (Hoofdstuk 4): Productieve vaardigheden — schrijven, spreken en bereidheid om te communiceren
Deze derde studie breidde het onderzoek uit naar productieve Engelse vaardigheden en naar *willingness to communicate* (bereidheid om te communiceren, WtC; MacIntyre et al., 1998), met de vraag of CLIL-leerlingen in het vmbo sterker vooruitgingen in schrijf- en spreekvaardigheid en in vloeiendheid dan hun niet-CLIL-peers. Schrijf- en spreekvaardigheid werden gemeten via aangepaste performance-taken op drie meetmomenten. WtC werd gemeten met aanvullende zelfrapportagevragenlijsten. Voor zowel schrijven als spreken werd gekeken naar totale vaardigheid en vloeiendheid (het totaal aantal

geproduceerde Engelse woorden). De analyses maakten gebruik van multilevel regressiemodellen die rekening hielden met de geneste structuur van leerlingen in klassen en scholen. Voor WtC werden afzonderlijke schalen gebruikt voor de vertrouwde schoolcontext en een onbekende taalgebruikscontext.

De resultaten lieten een significant voordeel zien voor de CLIL-leerlingen voor schrijfvaardigheid en schrijfvloeiendheid, al waren de effectgroottes klein: CLIL verklaarde ongeveer 2 procent van de relatieve groei in schrijfvaardigheid en 1 procent in schrijfvloeiendheid. Spreekvaardigheid liet eveneens een klein maar consistent CLIL-voordeel zien, in zowel de ANCOVA-analyses als de change-score analyses. Spreekvloeiendheid vertoonde echter een genuanceerder patroon: in de ANCOVA produceerden CLIL-leerlingen gemiddeld circa 20 woorden meer per taak, terwijl de change-score analyse dezelfde positieve richting liet zien maar niet significant was. Toch wijst het patroon op een voordeel in vloeiendheid vooral bij de jongste CLIL-leerlingen (cohort 1), die over tijd de grootste toename in mondelinge output lieten zien. Een marginale moderatie-effect suggereerde bovendien dat leerlingen met lagere Cito-vaardigheidsscores uit het basisonderwijs iets sterker profiteerden in spreekvloeienheid dan hun hoger scorende peers.

CLIL-leerlingen gaven zichzelf ook hogere WtC-scores in de schoolcontext, maar niet in een onbekende taalgebruikscontext. Beide WtC-schalen namen significant toe in alle groepen, maar de groeipatronen verschilden niet significant tussen CLIL- en niet-CLIL-leerlingen wanneer interactietermen werden toegevoegd. CLIL-leerlingen behielden dus een hogere zelfgerapporteerde mate van vertrouwen in schoolse communicatiesituaties, maar ontwikkelden geen extra voordeel in onbekende contexten. Kleine contextuele invloeden—zoals leerjaar en vmbo-leerweg—droegen bij aan verschillen tussen scholen. De studie leverde bovendien enigszins uiteenlopende resultaten op afhankelijk van de analysemethode (ANCOVA versus change-score), een methodologisch punt dat later in Sectie 6.7 wordt besproken.

Samenvattend laten de bevindingen zien dat vmbo-CLIL-leerlingen bescheiden maar statistisch aantoonbare voordelen behaalden in productieve Engelse vaardigheden, vooral in schrijven en spreken, terwijl de voordelen in vloeiendheid kleiner en contextafhankelijk waren. Hoewel ANCOVA- en change-score uitkomsten licht van elkaar verschilden, wezen beide in dezelfde richting. Het totale patroon suggereert dat CLIL geassocieerd was met kleine maar consistente voordelen in productieve vaardigheden en in het zelfvertrouwen van leerlingen om Engels te gebruiken in schoolse situaties, ook al bleven de statistische effecten over het algemeen beperkt.

Studie 4 (Hoofdstuk 5): Perspectieven van leraren en leerlingen op CLIL-implementatie
Hoofdstuk 5, gebaseerd op een vragenlijst onder CLIL-docenten en een vragenlijst plus vijftien focusgroepgesprekken met CLIL-leerlingen in zes vmbo-scholen, onderzocht uitsluitend de perspectieven van direct betrokkenen bij CLIL-programma's. In tegenstelling tot de voorafgaande kwantitatieve studies bevatte deze deelstudie geen vergelijking met niet-CLIL-groepen, maar richtte zich op de ervaringen van degenen die dagelijks met CLIL werken. Deze kwalitatieve inzichten verdiepen de statistische bevindingen doordat zij laten zien hoe sterke punten en beperkingen van vmbo-CLIL in de praktijk worden ervaren.

Beide groepen noemden vooruitgang in het Engels het meest positieve aspect. Leerlingen ervoeren CLIL-lessen vaak als leuk, anders en actief en verbonden deze met zelfvertrouwen, interactie en plezier, ondanks dat leren via Engels soms moeilijk was. Tegelijkertijd vonden sommige leerlingen dat lessen te weinig doeltaalinput of communicatieve focus boden. Bijna driekwart zou opnieuw voor CLIL kiezen, wat suggereert dat ervaren voordelen doorgaans zwaarder wogen dan de uitdagingen. Docenten beschreven CLIL eveneens als motiverend en activerend, maar wezen op inconsistent taalgebruik, werkdruk en de behoefte aan samenwerking.

Begrip van taal en vakinhoud vormde een terugkerende spanningsbron. Sommige docenten betwijfelden de haalbaarheid van CLIL in de lagere leerwegen, terwijl veel leerlingen juist een betere balans tussen Engels en Nederlands nodig achtten voor begrip. Zij waardeerden docenten die duidelijk communiceerden en Engels veelvuldig of strategisch afwisselden met Nederlands; inconsistent Engels of beperkte uitleg in beide talen leidde tot frustratie of verlies van betrokkenheid. Deze percepties sloten aan bij de grote variatie in doeltaalgebruik die docenten zelf rapporteerden.

Docenten verschilden sterk in hun gebruik van het Engels—van bijna volledig tot minder dan 20%—door lokale verschillen in taalvaardigheid, schoolbeleid en aandacht voor vakinhoudelijke begrijpelijkheid. Leerlingen bevestigden dit beeld: consequent Engels kon motiveren, maar wisselende of minimale doeltaalpraktijken werden als verwarrend of demotiverend ervaren.

Beide groepen benadrukten het belang van interactie en taakvariatie. Communicatieve en actieve taken werden door leerlingen gekoppeld aan plezier en vooruitgang; docenten waardeerden afwisseling en authenticiteit, ondanks extra voorbereidingstijd. Tegelijkertijd gaven sommige leerlingen aan dat feedback op taal zelden voorkwam en aandacht voor taalvormen (focus on form) beperkt was, wat de taalontwikkeling mogelijk beperkte.

Leerlingen benadrukten daarnaast de relatie tussen docent en leerling. Effectief CLIL-onderwijs werd geassocieerd met geduld, persoonlijke aandacht en begrip voor de uitdaging van leren via Engels. Docenten noemden op hun beurt empathie belangrijk en merkten op dat gedragsproblemen vaak voortkwamen uit frustratie bij afnemend begrip. Deze relationele factoren bleken cruciaal voor het zelfvertrouwen van leerlingen.

De beschreven ervaringen lieten zien dat de waargenomen effectiviteit van vmbo-CLIL vooral afhing van het consequente gebruik van begrijpelijk Engels, naast de hoeveelheid input die zij ontvingen. Docenten verschilden aanzienlijk in hun doeltaalgebruik—van zeer uitgebreid tot zeer beperkt—waardoor kwaliteit en kwantiteit van de input sterk uiteenliepen. Vrijwel alle focusgroepleerlingen wilden meer Engels of waardeerden dat docenten al (bijna) consequent 'doeltaal-voertaal' toepasten, ook wanneer dat tot begripsspanningen leidde.

Leerlingen hechtten daarnaast veel waarde aan internationale activiteiten, zoals reizen en uitwisselingen, die zij belangrijk vonden voor hun motivatie en de identiteit van CLIL. Wanneer dergelijke kansen uitbleven of slechts voor een kleine groep toegankelijk waren, leidde dat tot diepe teleurstelling en een gevoel van ongelijkheid.

Ten slotte gaven verschillende leerlingen aan dat zij zich in hun leerpotentieel soms onderschat voelden of onvoldoende uitgedaagd, wat tot demotivatie en verminderde betrokkenheid leidde. Gezamenlijk suggereren deze ervaringen dat leerlingen zich het meest competent en gemotiveerd voelen wanneer begrijpelijke taalinput, relationele ondersteuning en realistische maar stimulerende verwachtingen elkaar versterken binnen het CLIL-programma.

7. Geïntegreerde bevindingen: wat we nu weten

In synthese suggereren de bevindingen uit de vier studies dat de didactische kern van CLIL—het geïntegreerd leren van vakinhoud en taal—ook in minder academische contexten kan functioneren, maar dat de daadwerkelijke realisatie ongelijk en sterk lokaal afhankelijk is. In verschillende deelnemende scholen gaven leerlingen aan dat zij beperkte en inconsistente doeltaalinput, weinig communicatieve oefenkansen en relatief weinig expliciete aandacht voor taal binnen vaklessen kregen. Deze omstandigheden helpen verklaren waarom de kwantitatieve studies (Hoofdstukken 3 en 4) slechts bescheiden of niet-significante verschillen vonden tussen CLIL- en niet-CLIL-leerlingen: de talige voordelen die vaak aan CLIL worden toegeschreven, kunnen moeilijk ontstaan wanneer taalblootstelling, feedback en oefening beperkt blijven. Tegelijkertijd laat het kwalitatieve materiaal zien dat CLIL wél duidelijk potentieel heeft wanneer de instructie responsief wordt ondersteund, verwachtingen realistisch zijn en docenten diverse leerlingen begeleiden met helderheid en empathie. De daaropvolgende secties bespreken daarom de

bredere theoretische, beleidsmatige en pedagogische implicaties van deze resultaten voor toekomstbestendige en meer rechtvaardige vormen van vmbo-CLIL.

Over de vier deelstudies heen bleek CLIL haalbaar voor leerlingen met een lagere schoolse uitgangspositie, mits programma's werden aangepast aan de vmbo-context. Vier brede conclusies kwamen naar voren:

1. Attitudinale groei zonder voorselectie – Positieve motivatie ontwikkelde zich in inclusieve klassen, wat aantoont dat de motiverende effecten van CLIL niet beperkt zijn tot academisch sterke leerlingen.
2. Meetbare taalprogressie ondanks beperkte taalblootstelling – Leerlingen gingen vooruit in de meeste receptieve en productieve vaardigheden, zij het binnen leerwegpassende grenzen.
3. Leerlingdiversiteit vormt geen belemmering – Meertalige achtergronden en uiteenlopende beginsituaties bleken goed verenigbaar met gestage ontwikkeling.
4. Implementatie doet ertoe – Verschillen in doeltaalgebruik en in de mate waarin scholen CLIL-principes toepasten, beïnvloedden hoe leerlingen en docenten CLIL ervoeren in de dagelijkse praktijk en beperkten waarschijnlijk de kansen voor taalontwikkeling in sommige programma's.

Motivatie volgde een tijdspatroon: sterke groei in de eerste jaren van intensieve blootstelling, gevolgd door stabilisatie naarmate nieuwigheid en intensiteit afnamen. Taalwinst was soms bescheiden, maar over het algemeen positief over de verschillende vaardigheden heen, in overeenstemming met wat leerlingen en docenten rapporteerden over wisselende kansen voor duurzame communicatie en (incidentele) aandacht voor taalvorm.

Gezamenlijk wijzen deze patronen erop dat voortgezette talige en attitudinale groei afhankelijk is van gevarieerde, uitdagende, plezierige en communicatief rijke Engelstalige leerkansen—voorwaarden die leerlingen zelf omschreven als wisselend of inconsistent tussen scholen.

Succes in vmbo-CLIL moet daarom worden beoordeeld aan de hand van haalbare voortgang binnen aangepaste eindtermen, niet door vergelijking met hogere leerwegen. Dit perspectief bevordert onderwijsrechtvaardigheid doordat het verwachtingen afstemt op wat realistisch en passend is binnen de vmbo-context.

8. Implicaties voor theorie, beleid en onderwijskundige praktijk

8.1 Theoretische implicaties

De bevindingen dragen bij aan lopende theoretische discussies over de haalbaarheid van CLIL voor leerlingen buiten academisch selectieve contexten. Fundamentele model-

len zoals het 4Cs-framework (Coyle, 2007; Coyle et al., 2010) en het dual focus-principe (Marsh et al., 2009a) gaan ervan uit dat vakinhoudelijk leren en taalverwerving elkaar wederzijds kunnen versterken. Het huidige onderzoek bevestigt dat deze mechanismen ook binnen vmbo-contexten effectief kunnen functioneren, maar alleen wanneer cognitieve uitdaging, talige toegankelijkheid en contextuele randvoorwaarden zorgvuldig op elkaar zijn afgestemd.

De resultaten verfijnen het begrip haalbaarheid door te laten zien dat ontwikkeling binnen vmbo-CLIL wordt gestuurd door onderling verweven talige, attitudinale en contextuele factoren. Groei in Engelse taalvaardigheid bleek haalbaar, maar de ontwikkeling kan mogelijk worden verklaard door verschillen in gerapporteerde blootstelling aan Engels in de lessen; attitudinale uitkomsten werden bepaald door waargenomen nut, hanteerbaarheid en pedagogische ondersteuning. Deze dubbele lens sluit aan bij de nadruk op perceived behavioural control in het Model of Planned Behavior (Ajzen, 1991) en bij breder onderzoek naar zelfvertrouwen in tweedetaalcommunicatie (MacIntyre et al., 1998). Belangrijk is dat de bevindingen de stelling van auteurs zoals Pérez Cañado (2018, 2023) ondersteunen dat het potentieel van CLIL voor een diverse leerlingpopulatie meer afhangt van de kwaliteit en consistentie van de implementatie dan van leerlingselectie op zich.

8.2 Beleidsmatige en organisatorische implicaties

Op beleidsniveau onderstrepen de resultaten dat duurzaam vmbo-CLIL niet uitsluitend kan steunen op lokale innovatie of individuele docentexpertise, maar structurele ondersteuning, professionele erkenning en continuïteit vereist. De variatie tussen scholen toont het belang aan van institutionele kaders die minimale doeltaalblootstelling, taalvaardigheidseisen voor docenten en pedagogische standaarden vastleggen. De herziene TTO 2.0-standaard (Nuffic, n.d.-b) beweegt reeds in deze richting door verwachtingen te specificeren voor doeltaalgebruik, CLIL-methodologie en professionalisering van docenten.

Tegelijk laat de geconstateerde variabiliteit zien waar nationaal beleid de vmbo-dimensie van tweetalig onderwijs verder zou kunnen versterken. Gericht investeren in lesmateriaal, kleinere klassen, en gedifferentieerde scholingsroutes kan ongelijkheid door verschillen in middelen of docentervaring helpen verminderen. Intensere samenwerking tussen lerarenopleidingen en vmbo-CLIL-programma's kan daarnaast de continuïteit van expertise bevorderen.

De Ontwikkelagenda TTO 2025–2030 (Nuffic, 2025) legt bovendien nadruk op het versterken van TTO binnen het vmbo, het vergroten van de aantrekkelijkheid van het programma voor een diverse leerlingenpopulatie, het positioneren van Engels als instrument voor global citizenship en het uitbreiden van internationaliseringsmogelijkheden.

Deze prioriteiten sluiten nauw aan bij de conclusie van het proefschrift dat vmbo-CLIL vooral kan floreren wanneer talige doelen, leerlingendiversiteit en bredere onderwijsdoelen coherent worden geïntegreerd.

8.3 Pedagogische implicaties

Op klasniveau wijzen de bevindingen op verschillende voorwaarden waardoor CLIL toegankelijk werd zonder dat de verwachtingen hoefden te worden verlaagd. Leerlingen uit meerdere scholen rapporteerden dat begrijpelijke input, duidelijke uitleg en strategische translanguaging (functioneel meertalig handelen) hun begrip ondersteunden; sommige docenten gaven aan deze praktijken flexibel in te zetten. Samenwerkingsopdrachten, creatieve activiteiten en andere gestructureerde mogelijkheden voor taalproductie werden door leerlingen vaak genoemd als momenten waarop de lessen 'leuk', activerend en vertrouwenwekkend aanvoelden. Tegelijkertijd werden deze praktijken onregelmatig toegepast: sommige docenten—met name in de lagere vmbo-leerwegen—vertrouwden sterk op het Nederlands of boden minder communicatieve opdrachten aan, wat leerlingen koppelden aan minder begrip of gemiste kansen voor zinvol gebruik van het Engels.

Gezamenlijk sluiten deze patronen aan bij gevestigde CLIL-pedagogiek: interactieve taken, regelmatige mogelijkheden voor output en lichte, tijdige feedback op taalvorm ondersteunen zowel talige participatie als affectieve betrokkenheid. Samenwerking tussen docenten Engels en vakdocenten kan de coherentie verder vergroten door talige doelen expliciet te verbinden aan vakinhoud en door ervoor te zorgen dat leerlingen in meerdere vakken Engelse input tegenkomen in plaats van in geïsoleerde momenten.

Cruciaal is dat de kwalitatieve bevindingen het belang van interpersoonlijke dynamiek benadrukt. Leerlingen wezen herhaaldelijk op het belang van duidelijkheid, geduld en de responsiviteit van docenten op frustratie—voorwaarden die hen helpen doorzetten wanneer taken overweldigend lijken. Zoals Schwab et al. (2014) stellen, kan affectieve ondersteuning talige uitdaging omzetten in haalbare groei. Leerlingen maakten bovendien duidelijk dat plezier ertoe doet: voor vmbo-leerlingen droegen lessen die als leuk werden ervaren—gevarieerd, actief en sociaal—bij aan zelfvertrouwen, minder defensieve reacties op gevoelens van onderschatting of onvoldoende erkenning van hun moeite, en meer ruimte voor authentieke communicatie. Deze bevindingen stroken met onderzoek waaruit blijkt dat positieve klasemoties de *willingness to communicate* en mondelinge risicobereidheid in een tweede taal bevorderen (Boudreau et al., 2018; Dewaele & MacIntyre, 2014), evenals met CLIL-studies die laten zien dat goed ondersteunde pre-vocational programma's motivatie en zelfconcept kunnen versterken bij leerlingen met lagere schoolse uitgangsniveaus (Denman et al., 2013; Denman et al., 2018; zie ook Hoofdstuk 2 van dit proefschrift).

Samengevat suggereren de pedagogische implicaties dat effectieve vmbo-CLIL-praktijken berusten op een combinatie van talige scaffolding, interactieve taakontwerpen, een ondersteunende relationele sfeer en kansen voor betekenisvolle betrokkenheid. Een terugkerend thema in zowel leerling- als docentperspectieven is de noodzaak van ruime Engelse input én output, in lijn met aanbevelingen van Denman et al. (2013) en overeenkomend met de eigen wens van de leerlingen voor meer systematische doeltaalblootstelling in de lessen. Deze voorwaarden helpen leerlingen CLIL te ervaren als zowel hanteerbaar als de moeite waard.

9. Huidige situatie en toekomstperspectief

Hoewel de gegevens voor de empirische studies in dit proefschrift tussen 2012 en 2014 zijn verzameld, bieden ontwikkelingen in het daaropvolgende decennium belangrijke context voor de interpretatie van de bevindingen. De zes deelnemende scholen volgden geen gelijklopend langetermijntraject. Eén school wist zijn tweetalig onderwijs te versterken en uit te breiden, gesteund door een stabiel docententeam, een duidelijke CLIL-visie en samenhangende institutionele structuren. De overige scholen hebben hun vmbo-CLIL-programma's uiteindelijk ingeperkt of beëindigd—meestal als gevolg van personeelsverloop, financiële druk, fusies of organisatorische keuzes die niet direct met didactiek samenhingen. In meerdere gevallen hield de beëindiging verband met een onvoldoende of instabiele instroom van leerlingen in de CLIL-stroom, wat de structurele of financiële haalbaarheid van voortzetting ondermijnde. Gezamenlijk onderstrepen deze uiteenlopende trajecten een centrale conclusie van dit proefschrift: vmbo-CLIL is haalbaar, maar duurzame borging vereist structurele continuïteit, voldoende personele capaciteit en een stabiele leerlinginstroom—niet louter enthousiasme of kortstondige innovatie.

Nationale ontwikkelingen tussen 2012 en 2025 laten een geleidelijke verschuiving zien naar een meer gestructureerde en duurzame benadering van tweetalig onderwijs. Opeenvolgende updates van de vmbo TTO-Standaard 2.0 (Nuffic, n.d.-b) introduceerden duidelijkere verwachtingen ten aanzien van doeltaalgebruik, CLIL-methodologie en docentcertificering, in combinatie met expliciete erkenning van leerwegspecifieke taalniveaus. Het aantal scholen met een vmbo-CLIL programma is in deze periode meer dan verdubbeld, van 14 naar 32, inclusief twee in het Nederlands Caribisch gebied. De toenemende nadruk op functionele Engelse taalvaardigheid, communicatieve competentie en interculturele oriëntatie sluit aan bij zowel de bevindingen van dit proefschrift als bij bredere discussies over de maatschappelijke waarde van beroepsgericht onderwijs.

Vooruitkijkend legt de Ontwikkelagenda TTO 2025–2030 een sterker accent op een inclusieve en toekomstgerichte tweetalige pedagogiek. Centrale ambities zijn het vergroten van de zichtbaarheid en aantrekkelijkheid van vmbo-TTO, het borgen van toegankelijkheid voor diverse leerlingpopulaties, en het positioneren van Engels primair als middel

voor *global citizenship* in plaats van als einddoel. De agenda benadrukt verder de uitbreiding van internationalisering en Europees en Internationale Oriëntatie componenten via uitwisselingen, projectmatige samenwerking en digitale of hybride vormen, en pleit voor meer samenhang tussen scholen door gedeelde professionalisering en onderlinge partnerschappen. Deze prioriteiten resoneren met patronen die in dit proefschrift naar voren komen: leerlingenwens voor betekenisvolle Engelse input en internationale kansen, het belang van relationele ondersteuning, en de noodzaak van haalbare verwachtingen die aansluiten bij lokale contexten. Naarmate vmbo-CLIL deze nieuwe ontwikkelfase ingaat, lijkt het nationale beleid steeds beter aan te sluiten bij principes van duurzame en rechtvaardige vormen van tweetalig onderwijs.

Vervolgonderzoek

De beperkingen van dit onderzoek wijzen bovendien op duidelijke richtingen voor toekomstig onderzoek. De niet-gerandomiseerde aard van de steekproeven vereist terughoudendheid bij het interpreteren van groepsverschillen, en de afhankelijkheid van docent-zelfrapportages beperkt de precisie waarmee klassentaalgebruik kan worden vastgesteld. Toekomstige studies zouden baat hebben bij meer systematische steekproeftrekking waar mogelijk, aangevuld met klasobservaties, fijnmazige analyses van gesproken interactie en nauwkeuriger monitoring van de Engelse taalvaardigheid van de docenten. Langetermijnonderzoek - bijvoorbeeld het volgen van vmbo-CLIL-leerlingen in vervolgonderwijs zoals het mbo - kan inzicht bieden in de vraag of de attitudinale en talige patronen die hier zijn vastgesteld, doorwerken in latere onderwijstrajecten en zelfvertrouwen in Engelstalige omgevingen. Naarmate vmbo-CLIL zich verder ontwikkelt onder de Ontwikkelagenda TTO 2025–2030, ontstaan nieuwe kansen om te onderzoeken hoe blootstelling, pedagogiek en relationele praktijk samen vorm krijgen in contexten die aanzienlijk verschillen van de academisch selectieve settings die in de CLIL-literatuur vaak centraal staan.

Gezamenlijk bevestigen deze ontwikkelingen de functie van dit proefschrift als startmeting voor het begrijpen van de vroege implementatie van vmbo-CLIL en verduidelijken zij de bredere voorwaarden voor duurzaamheid. Zij laten zien dat vmbo-CLIL, wanneer het is ingebed in realistische verwachtingen en samenhangende institutionele structuren, kan uitgroeien tot een stabiel onderdeel van het voortgezet onderwijs—gevoed door de lessen uit deze experimentele beginfase.

10. Slotconclusie

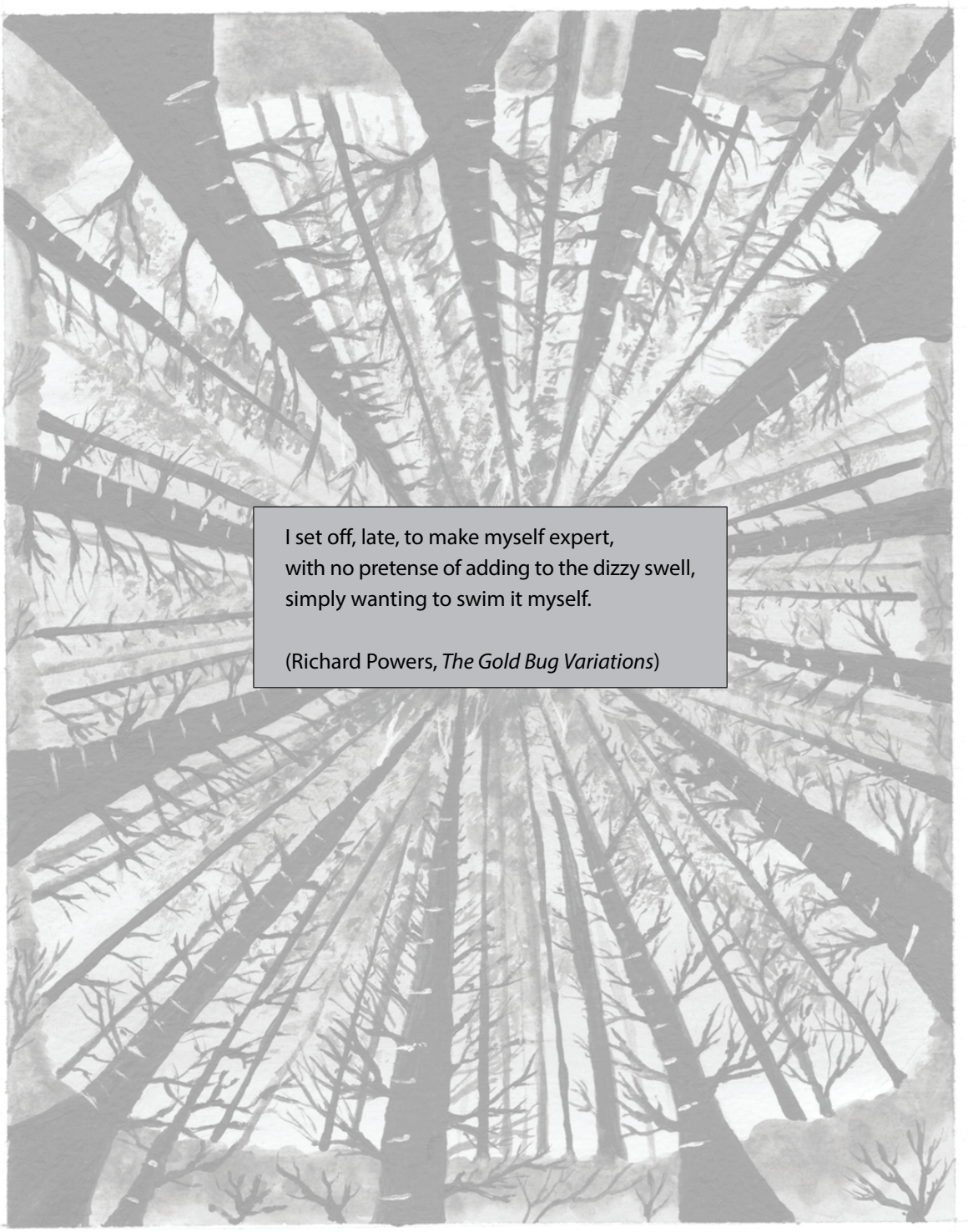
De vier deelstudies in dit proefschrift laten zien dat CLIL zowel haalbaar als rechtvaardig kan zijn voor vmbo-leerlingen, mits het wordt uitgevoerd binnen leerwegpassende verwachtingen en wordt ondersteund door relationele en pedagogische responsiviteit. Hoewel de taalwinst soms bescheiden was, was deze consistent en betekenisvol, en weerspiegelde de ervaringen van leerlingen van groeiend zelfvertrouwen en actieve

betrokkenheid. Attitudinale en talige ontwikkeling versterkten elkaar wederzijds, en de empathie en helderheid van docenten bleken cruciale voorwaarden voor blijvende betrokkenheid van leerlingen.

Gezamenlijk tonen de bevindingen aan dat CLIL in het voorbereidend beroepsonderwijs meetbare en betekenisvolle taalontwikkeling kan realiseren, naast positieve attitudinale groei, zelfs zonder academische voorselectie—mits de pedagogische basis stevig is en docenten voldoende ondersteuning krijgen. Door de vormende beginjaren van vmbo-CLIL te documenteren en te plaatsen binnen een zich ontwikkelend landelijk beleidskader, biedt dit proefschrift een fundament om het begrijpen hoe tweetalig leren gestalte kan krijgen in niet-selectieve contexten.

Inclusieve CLIL wordt daarbij niet gedefinieerd door het reproduceren van resultaten uit meer elitaire en academische contexten, maar door het mogelijk maken van haalbare voortgang, het versterken van leerlingenidentiteit, en het bevorderen van duurzame tweetalige praktijken in het voorbereidend beroepsonderwijs. In bredere zin geldt dit ook voor vergelijkbare internationale contexten.

Tot slot dank ik de deelnemende vmbo-scholen, hun schoolleiding, docenten en leerlingen voor hun medewerking aan dit onderzoek. Mijn bijzondere dank gaat uit naar de leerlingen, die met grote inzet en geduld deelnamen aan de testafnames en herhaalde metingen en hun ervaringen openhartig deelden in de focusgroepgesprekken. Ook de docenten ben ik erkentelijk voor de tijd die zij uittrokken en voor hun waardevolle perspectieven op de dagelijkse praktijk van CLIL. Verder dank ik Nuffic voor de ondersteuning bij de werving van scholen, het verstrekken van informatie en updates, en de mogelijkheid om aan brainstormsessies en accreditatiebezoeken deel te nemen. Hun bijdragen waren essentieel voor de totstandkoming van dit onderzoek.



I set off, late, to make myself expert,
with no pretense of adding to the dizzy swell,
simply wanting to swim it myself.

(Richard Powers, *The Gold Bug Variations*)

Acknowledgements

Without the cooperation of the CLIL coordinators, teachers, and the hundreds of pupils at the research schools, there would simply have been no data for this study. I am deeply grateful for the organizational support of the schools, the persistence of the pupils when faced with repeated test batteries, and especially for the frank and forthcoming contributions of the focus group CLIL pupils.

Immense thanks go to my supervisor, Rick de Graaff, and my co-supervisor, Erik van Schooten, for their faith in my ability to finish this dissertation—even when I myself wavered. Rick, invariably calm, patient, and encouraging, you were instrumental in fuelling my motivation with opportunities for workshops, presentations, conferences, and interaction with other CLIL researchers. Erik, we found common points of intersection in music, literature, and politics, if not in statistics! Your vision from the start of this project, and your commitment to it—and especially your expert and comprehensive statistical analyses—were the *sine qua non* of this research.

Several prominent researchers in the field generously shared their time and expertise with me, which was not only informative but also tremendously inspiring. I am indebted to Anthony Bruton (Spain), Fred Genesee and Robert Gardner (Canada), Simone Smala, Margaret Gearon, and Russell Cross (Australia), Dominik Rumlich (Germany), and Tessa Mearns, Leo van Putten, and Marjolein Verspoor (the Netherlands).

Closer to home, my thanks go to the Board of Hogeschool Rotterdam for granting me a 'promotievoucher' (PhD grant) which enabled me not only to keep working and teaching, but also to be part of the Kenniscentrum Talentontwikkeling. Frans Spierings, the inspiring KCTO director, generously made funds available for my international conference presentation trips to Belgium, Spain, and Australia. Thanks, too, to Nuffic, especially to Onno van Wilgenburg, Ineke Swaan, and Lineke Owendijk, and for the opportunities afforded to me as a member of team for TTO *vmbo* visitation and assessment which allowed me to see later *vmbo* CLIL programs at work.

A resounding shout-out to my colleagues in the English Department of the Lerarenopleiding, Hogeschool Rotterdam, who continued to show support and interest for the seemingly interminable duration of this research project. Special thanks to Claire and Lana for their listening ears and good-natured sense of perspective.

A condition of the HR 'promotievoucher' was that HR students be included in the research process in some way. I was fortunate to have several teams of eager and dedicated English teacher training students who helped assess the thousands of writing and speaking tests. The process was as much a journey of discovery for you as it was for me! Thank

you, Andrej, Astrid, Bart, Bernadine, Coen, Erik, Ian, Kirsten, Laura, Leroy, Maartje, Malou, Michael, Naomi, Rob, Sandra, Saskia, Sherleen, Theresa, Yardéna, Zayenne, and Zoe!

To Joke and Simone, gallery owners of Galerie Terra Delft, I want to express my thanks for the decades-long trust you have placed in me to be the in-house translator for your peerless contemporary ceramics gallery, allowing me to hone the translation skills which I also used in this project. Congratulations with Terra 40!

An intensive professional project needs to be balanced with extracurricular passion. Thanks to all of the second violins of the Delft Symphonie Orkest for providing a joyful and harmonious social and musical foil to my PhD work, and for allowing me to be your section leader. A standing ovation for Anja, Catherine, Courtney, Eri, Hilde, Ingeborg, Joep, Kitty, Liedieke, Liesbeth, Maria, Mieke, Monica, Renske, and Tanja!

Last, but in no way least, to my family here and abroad, especially Katja, Mathieu, Benno, Hester, Toby, Kaðlín, Sander, and Tamar. Specific and invaluable contributions were made by Toby (logistical help at the schools and technical help with splitting the recordings), Sander (cover painting and speaking test transcriptions), Benno (creating Figure 5.1), and Hester (indispensable qualitative research advice). To all of you: your support, patience, and love have meant more to me than I can express here. Finally, to Pieter, for his partnership and for offering me in life—if not with this dissertation—in his own immortal words at our wedding, "help, help, and help".

Curriculum Vitae

Jenny Denman grew up in New Jersey, Pennsylvania, New York, Vermont, and Colorado, USA. She spent a year at Lehrerseminar Chur, Switzerland (primary school teacher training) under the auspices of the American Field Service, and attended the Aspen Music School, the Vienna International Music Center, and the Vienna Conservatory (now the Music and Arts University of the City of Vienna, Austria) while studying music and comparative literature at Swarthmore College in Pennsylvania, where she graduated with a B.A. in German. She returned there to work while singing with professional church and synagogue choirs, the Philadelphia Singers, and the Opera Company of Philadelphia.

In 1985 she moved to the Netherlands and joined the extra chorus of the Dutch National Opera. Eventually discovering the incompatibility of rehearsals, performances, touring, and family life, she worked as a freelance translator and as an editorial assistant for Elsevier before retraining as an English teacher. She received her B.Ed. and M.Ed. from the Rotterdam University of Applied Sciences and her M.A. in English Language and Literature *cum laude* from Leiden University. Alongside her graduate studies and teaching career, she worked as a Cambridge Oral Examiner for the British Council and wrote English secondary-level teaching materials for the Dutch market, including *New Interface* and *Wasp Reporter,* and also experienced the joys and challenges of raising four bilingual children. She has taught English at primary, lower secondary, upper secondary, and tertiary level in the Netherlands, and was a teacher trainer in the Institute for Teacher Education of Rotterdam University of Applied Sciences for over 20 years. Most recently she taught academic English to physics graduate students at Sun Yat-sen University in Guangzhou, China.

This dissertation examines the opportunities and challenges of Content and Language Integrated Learning (CLIL) in Dutch pre-vocational secondary education (vmbo). While bilingual education has often been associated with academically high-performing pupils, this study investigates whether CLIL can also function successfully in more diverse and practically oriented educational contexts. Drawing on longitudinal quantitative and qualitative research conducted in seven Dutch secondary schools, the dissertation compares the English language development of CLIL and non-CLIL vmbo pupils in vocabulary, grammar, writing, speaking, and willingness to communicate. It also explores the perspectives of CLIL teachers and pupils themselves. The findings indicate that inclusive forms of CLIL can support language development and positive attitudes, provided that programs take account of learners' needs, educational context, and the importance of meaningful communication and support.

Jenny Denman is a former teacher educator at Rotterdam University of Applied Sciences with a professional background in music, language teaching, teacher training, translation, and educational publishing. She has taught English in primary, secondary, and tertiary education for over twenty-five years.