AILA Review

Inclusive CLIL: Pre-vocational pupils' target language oral proficiency, fluency, and willingness to communicate

--Manuscript Draft--

Manuscript Number:	AILA-22020R3
Full Title:	Inclusive CLIL: Pre-vocational pupils' target language oral proficiency, fluency, and willingness to communicate
Short Title:	Inclusive CLIL: pre-vocational pupils' speaking skills
Article Type:	Special Issue Article
First Author:	Jenny Denman, M.A., M.Ed.
Corresponding Author:	Jenny Denman, M.A., M.Ed. Rotterdam University of Applied Sciences: Hogeschool Rotterdam Rotterdam, NETHERLANDS
Other Authors:	Erik van Schooten, PhD
	Rick de Graaff, PhD
Order of Authors Secondary Information:	
Funding Information:	
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 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, fluency, and Willingness to Communicate. Furthermore, the CLIL pupils showed significantly more growth than the non-CLIL control group in 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. 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
Manuscript Classifications:	Language acquisition; Language teaching and learning
Section/Category:	
Author Comments:	There are 3 Tables for inclusion in the article itself. The Appendix is uploaded as a separate document intended to be accessible online - as advised by the Guest Editors - and contains 3 Attachments (two new ones requested by the reviewers) and the remaining statistical tables. Thank you.

Suggested Reviewers:	
Opposed Reviewers:	

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

Jenny Denman, Rotterdam University of Applied Sciences, Research Centre Urban Talent
Erik van Schooten, Rotterdam University of Applied Sciences, Research Centre Urban Talent
Rick de Graaff, University of Utrecht, Faculty of Humanities, Department of Languages, Literature and Communication

Running head: Inclusive CLIL: pre-vocational pupils' speaking skills

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 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, fluency, and Willingness to Communicate. Furthermore, the CLIL pupils showed significantly more growth than the non-CLIL control group in 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. 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

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, 2013). Supported by several Council of Europe initiatives, the CLIL approach spread rapidly throughout Europe, particularly in secondary education (Nikula, 2017). 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 (Baïdak, Balcon, & Motiejunaite, 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.

One of the core questions is to what extent CLIL contributes to productive L2[1] development. As secondary school CLIL programs in Europe have often been offered only in the more academic school types (Feddermann, Möller, & Baumert, 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, Genesee, and Gorter, 2013), CLIL still increases not only the contact time compared to mainstream foreign language education, but also the quality of the interaction in the target language (Escobar Urmeneta, 2019). This is due in part to its focus on communication and meaning rather than form-focused accuracy, and because the CLIL approach should offer more opportunities for L2 interaction and authentic communication (Pérez-Vidal, 2009).

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, Clément, Dörnyei, and Noels (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, Baker, & MacIntyre, 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).

[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, Myles, & Marsden, 2019). Accordingly, we used L1 to indicate Dutch (the majority language and the language of school), still recognizing that over one-third of the pupils do not have Dutch as their first or home language.

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, Clément, & MacIntyre, 2003; MacIntyre, Baker, Clément, & Conrod, 2001). Individual variables such as gender, age, and prior L2 experience can influence a learner's WTC (MacIntyre, Baker, Clément, & Donovan, 2002; MacIntyre et al., 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, Vanhees, Smits, & Van De Putte, 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, Denessen, & Verhoeven, 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, Denessen, & Verhoeven, 2017). 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.

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, Westhoff, & De Bot, 2006; Lasagabaster, 2008; Lorenzo, Casal, & Moore, 2010; 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, Hüttner, Jexenflicker,

Schindelegger & Smit, 2008; Juan, 2010). 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 below. 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).

Comparisons and evaluations of the various empirical results for speaking are complicated because not all researchers have used an experimental design or stringently controlled for initial levels (Bruton, 2011a; Verspoor, de Bot, & Xu, 2015) or appropriately controlled for selection effects (Goris, Denessen, & Verhoeven, 2020; Paran, 2013; Pérez Cañado, 2020; Piesche, Jonkmann, Fiege, & Keßler, 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, 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, Jonkmann, Hollm, & Fiege, 2016; Schwab, Keßler, & Hollm, 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 unconsciously apply their own socio-cultural biases in the classroom (Van den Bergh, Denessen, Hornstra & Holland, 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, Jonkmann, & Hollm, 2018) - such as

gender, home language, standard achievement test scores, and prior L2 instruction - which might moderate the results.

CLIL selectivity

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).

Research on Canadian immersion has shown that pupils with low academic ability or low literacy development can benefit from bilingual immersion (Cummins, 1984), 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.

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, Granados, & Rico, 2021), Queensland and Victoria, Australia (Smala, 2021), the UK (Coyle, Bower, Foley, & Hancock, 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, van Schooten, & de Graaff, 2018), 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 (Bruton, 2011b; Broca, 2016; 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 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 (2007) 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.

The present study

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 sub-streams, 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 sub-streams (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 had until recently 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 stream, 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 350 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 (Standaard, 2020) 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 100 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.

Our research questions, therefore, are:

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

2. What is the effect of CLIL on pre-vocational pupils' growth in English speaking skills and WTC?

3. What are the differential effects of CLIL on growth in English 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 speaking proficiency, speaking 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 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.

Method

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 one, two and three of pre-vocational education and finishing respectively at the end of years two, three and four. 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.

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 was filled out once, at the start of the study. The WTC questionnaire was 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 the first WTC and speaking test measurements took place in the first weeks of the school year, the starting levels for WTC, speaking proficiency and speaking fluency 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 289 pupils who completed the speaking test at both measurement moments were included in the analyses. 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).

Participants

All fifteen pre-vocational schools with a 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 (see Table 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[2], 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 enrolment (CBS, 2016). In the study sample, there were 44 different non-Dutch languages spoken at home by either one or

[2] 2 For Dutch readers: *basisberoeps, kaderberoeps, gemengd, vmbo-t/mavo*

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.

@@ Insert Table 1 here

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 significant but low correlation was found for the home language environment: the experimental group (CLIL pupils) slightly more often had a non-Dutch home language environment ($r(563) = .117$, $p = .005$). In other words, the experimental and control group did not differ in individual pupil characteristics, except for a slight difference ($r^2 = .014$) in language background.

Instruments

Biodata questionnaire

The biodata questionnaire, filled in by all pupils at the start of the study, contained questions about gender, primary school scholastic aptitude test score[3], 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

3 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 sub-level.

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).

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 English '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 (Attachment 1, Appendix). 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 said 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, corresponding generally to the Common European Framework of Reference levels pre-A1 to C1 (CEFR, 2001). This holistic rubric was loosely based on the categories and descriptions of the Student Oral Proficiency Assessment Rating Scale (SOPA) (Boyson, Rhodes, & Thompson, 2009) and Cambridge English: Key for Schools (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 (Attachment 2, Appendix), based on the two sets of eight items for measuring WCT inside and outside the classroom (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).

Speaking Assessors

Over the two-year research period, the two measurements for speaking produced a total of 805 tests to be analyzed for proficiency and fluency. For proficiency, the speaking tests were made anonymous, transcribed, and assessed holistically by teams of six or seven assessors consisting of the head 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 was trained using the assessment rubric and a set of 'anchor' texts from the first measurement. 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 (hardly 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 speaking proficiency for scores of 0 through 4 are given in Attachment 3 (Appendix). For measuring speaking fluency, the number of English words in each text was counted.

Analyses

Reliability

Before conducting analyses to answer the research questions, the reliability of several instruments was calculated: Willingness to Communicate ('school' and 'London') and speaking proficiency and fluency. For the latter two instruments this was done by calculating the inter-rater reliability (IRR). For Willingness to Communicate, 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; see Table 2, Appendix).

To determine the assessors' individual inter-rater reliability (IRR) for the 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 (see Table 3, Appendix) because the number of rating categories was rather low since only the lower five of seven categories were actually used. The correlations found can be considered substantial to excellent (Albers, 2017): range Pearson's .701-.961; range Spearman's rho .707-.970 (see Table 3, Appendix).

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, dependent t-tests per pair of assessors show assessors were not equally strict: nearly half of the holistic speaking proficiency assessments showed significant differences in means in paired sample t-tests (see Tables 4 and 5, Appendix, for speaking proficiency). 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 so that all assessors scores are equally strict. This is permissible because of the random distribution of pupil tasks among assessors and because there were in all cases more than 100 tasks per assessor. The employment of a z-score here enables a comparison of the assessors' scores, making it 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.

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 (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 by means of the chi-square distributed difference in -2loglikelihood (Deviance) of nested models (one with and one without

the random intercept variance component tested). 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 (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).

Results

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

For speaking proficiency and speaking fluency (Table 6, Appendix), 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 6, Appendix).

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 7, model 7, Appendix), 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 8 (Appendix), 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 (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 were both more proficient and more fluent in speaking, and reported significantly more L2 WTC in a familiar school situation than their non-CLIL peers. For WTC-London there was no significant initial advantage for the CLIL pupils.

The second research question concerned the effect of CLIL on the growth in L2 oral proficiency, fluency, and WTC, controlling for initial starting difference in cohort 1. These results were obtained using the same instruments as the third research question (*Is there a differential effect of CLIL on growth in English 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.

Speaking proficiency, z-scores

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 9, model 8, Appendix), 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 10, Appendix). The most relevant models from these analyses of the Ancova and Change scores tables are given in Table 11.

@@ Insert Table 11 here

Tables 13 and 14 (Appendix) 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 15 & 16, Appendix).

Speaking fluency: number of English words

The analyses for speaking fluency followed the same pattern as for general speaking proficiency. The Ancova approach (Table 17, Appendix) 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 20.628 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 18, Appendix), 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 most relevant models from these analyses of the Ancova and Change scores tables are given in Table 12.

@@ Insert Table 12 here

The Ancova (Tables 19 & 20, Appendix) and the Change score (Tables 21 & 22, Appendix) 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 19, model 7 and 8 and Table 21, 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.

Willingness to Communicate: 'London' and 'School'

For WTC-London (Table 7), two random intercept levels are used in the analysis: repeated measures and pupil level. WTC-School (Table 8) 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 7, model 10; Table 8, model 10). Since there are no effects of CLIL on these measures, no moderation analyses with background variables were conducted.

Discussion and Conclusion

The aim of the present study was threefold. First, we wanted to ascertain whether at the start of CLIL the level of English speaking proficiency, 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 speaking 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.

Despite the lack of any formal selection criteria or procedures, 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 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. There was, however, no significant difference found between the two groups 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.

The second question concerned the longitudinal development of these skills for three cohorts of CLIL and non-CLIL pupils. 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). 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 language and 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, Denessen, and Verhoeven (2017). 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) 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.

We were also interested to know whether CLIL was more or less successful with pupils in different pre-vocational sub-stream 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 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, Hadjar, & Hascher, 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, 2017; 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.

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 speaking test 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 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 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, de Bot, & van Rein, 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 gains made over time in various linguistic domains, such as the use of chunks (Smiskova, Verspoor, & Lowie, 2012) or other analyses of complexity, accuracy, and fluency (Housen, Kuiken, & Vedder, 2012). Regarding the rating of 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, Schmid, & Xu, 2012). Finally, Mearns, de Graaff, and Coyle (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.

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 oral proficiency and oral fluency development in English as well as L2 willingness to communicate (WTC) inside and outside the classroom. The results 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 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.

References

ADiBE Project (Attention to Diversity in Bilingual Education, n.d.). Available from <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. DOI: 10.1080/13803610500392160

Albers, M.J. (2017). *Introduction to Quantitative Data Analysis in the Behavioral and Social Sciences*. Hoboken, NJ: John Wiley & Sons.

Allison, P.D. (1990). Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology*, *20*, 93-114. DOI: 10.2307/271083

Alonso, E., J. Grisaleña, & A. Campo (2008). Plurilingual education in secondary schools: Analysis of results. *International CLIL Research Journal* 1/1: 36–49. http://www.icrj.eu/11/article3.html

Baïdak, N., Balcon, M. P., & Motiejunaite, A. (2017). Key Data on Teaching Languages at School in Europe. 2017 Edition. Eurydice Report. *Education, Audiovisual and Culture Executive Agency, European Commission*. Available from <https://op.europa.eu/en/publication-detail/-/publication/73ac5ebd-473e-11e7-aea8-01aa75ed71a1/language-en/format-PDF>

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. DOI: 10.1111/0023-8333.00119

Boyson, B.A., Rhodes, N.C., & Thompson, L.E. (2009). *Administrator's manual for CAL Foreign Language Assessments, Grades K-8.* Center for Applied Linguistics, Washington, DC.

Broca, Á. (2016). CLIL and non-CLIL: differences from the outset. *Elt Journal*, *70*(3), 320-331. DOI: 10.1093/elt/ccw011

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. DOI: 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. DOI: 10.1016/j.system.2011.08.002

Bruton, A. (2013). CLIL: Some of the reasons why … and why not. *System, 41(3),* 587-597. DOI: 10.1016/j.system.2013.07.001

Cambridge English Key for Schools: Teacher's Handbook (2012). Cambridge: University of Cambridge ESOL Examinations.

Cenoz, J., Genesee, F., & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. *Applied linguistics*, *35*(3), 243-262. DOI: 10.1093/applin/amt011

CBS (Centraal Bureau voor de Statistiek) (2016). *VO; leerlingen, onderwijssoort, leerjar, herkomstgroepering, generatie* [data file]. Available from <http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=80043NED&D1=2-11,18&D2=0&D3=0&D4=a&D5=0&D6=0&D7=0&D8=9&HDR=G2,G5,G6,G7,G4,T&STB=G3,G1&VW=T>

CEFR (Common European Framework of Reference for Languages: Learning, Teaching, Assessment), 2001. Language Policy Unit, Strasbourg. Available from <https://www.coe.int/en/web/common-european-framework-reference-languages>

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. DOI: 10.1177/0261927x03022002003

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL quarterly*, *23*(3), 509-531. DOI: 10.2307/3586923

Coyle, D., Bower, K., Foley, Y., & Hancock, J. (2021). Teachers as designers of learning in diverse, bilingual classrooms in England: an ADiBE case study. *International Journal of Bilingual Education and Bilingualism*, 1-19. DOI: 10.1080/13670050.2021.1989373

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy.* Clevedon: Multilingual Matters. DOI: 10.3138/cmlr.42.1.137

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. DOI: 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. DOI: 10.1016/j.learninstruc.2015.09.003

Dalton-Puffer, C. (2008). Communicative Competence in ELT and CLIL Classrooms: Same or Different. *Views. Vienna English Working Papers*, *17*(3), 14-21.

Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, *31*, 182–204. DOI: 10.1017/s0267190511000092

Dalton-Puffer, C. (2017). CLIL in Practice: What does the research tell us? Available from <https://www.goethe.de/en/spr/unt/kum/clg/20984546.html>

Dalton-Puffer, C.; Hüttner, J., Jexenflicker, S., Schindlegger, V., & Smit, U. (2008) *CLIL an Österreichischen HTLs. Project Report.* University of Vienna/BMUKK

De Bot, K., & Maljers, A. (2009). De enige echte vernieuwing: Tweetalig onderwijs. In R. De Graaff en D. Tuin (Eds.), *De toekomst van het talenonderwijs: Nodig*? *Anders? Beter?* (pp. 131-146). IVLOS, Universiteit Utrecht.

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.* Unpublished doctoral dossertation, University of Groningen.

Denessen, E. J. P. G. (2017). *Dealing responsibly with differences: socio-cultural backgrounds and differentiation in education*. Inaugural Lecture, Leiden University. Available from <https://scholarlypublications.universiteitleiden.nl/access/item%3A2953124/view>

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. DOI: 10.1075/dujal.18005.den

Escobar Urmeneta, C. (2004). Content and language integrated learning: Do they learn content? Do they learn language? In J.D. Anderson, J.M. Oro and 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. (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. DOI: 10.5565/rev/clil.21

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

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, M. D. López-Jiménez, R. Chacón-Beltrán (Eds.), *Bilingual and Multilingual Education in the 21st Century: Building on Experience* (pp. 256-274). Bristol: Multilingual Matters. DOI: 10.21832/9781783090716-016

Genesee, F. (1987). *Learning through two languages: Studies of immersion and bilingual education.* Cambridge, MA: Newbury House.

Genesee, F. (2004). What do we know about Bilingual Education for Majority-Language Students? In T.K. Bhatia and W.C. Richie (Eds.), *The Handbook of Bilingualism* (pp. 547-576). Malden, MA: Blackwell. DOI: 10.1002/9780470756997.ch21

Genesee, F., & Fortune, T. W. (2014). Bilingual education and at-risk students. *Journal of Immersion and Content-Based Language Education, 2*(2), 196-209. DOI: 10.1075/jicb.2.2.03gen

Gierlinger, E.W., 2007. Modular CLIL in lower secondary education: some insights from a research project in Austria. In: Dalton-Puffer, C., Smit, U. (Eds.), *Empirical Perspectives on CLIL Classroom Discourse* (pp. 79-118). Frankfurt am Main: Peter Lang. DOI: 10.3726/978-3-653-01829-5/5

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. DOI: 10.1075/wll.16.2.03gor

Goris, J., Denessen, E., & Verhoeven, L. (2017). The contribution of CLIL to learners' international orientation and EFL confidence. *The Language Learning Journal, 47*(2), 246-256. DOI: 10.1080/09571736.2016.1275034

Goris, J. A., Denessen, E. J. P. G., & Verhoeven, L. T. W. (2019). Effects of content and language integrated learning in Europe A systematic review of longitudinal experimental studies. *European Educational Research Journal, 18*(6), 675-698. DOI: 10.1177/1474904119872426

Goris, J, Denessen, E., & Verhoeven, L. (2020). Determinants of EFL learning success in content and language integrated learning. *The Language Learning Journal*, DOI: 10.1080/09571736.2019.1709886

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. 1–20). Amsterdam: John Benjamins.

Hox, J.J. (2010). *Multilevel analysis. Techniques and applications. Quantitative methodology series*. (2nd ed.). New York: Routledge. DOI: 10.4324/9780203852279

Juan, M. 2010. Oral Fluency Development in Secondary Education CLIL Learners. *Vienna English Working Papers (Views)19 (3)*, 42–48.

Kang, S. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System*, 33: 227-92. DOI: 10.1016/j.system.2004.10.004

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: Breidbach, S. / Viebrock, B. (Eds.): *Content and language integrated learning (CLIL) in Europe. Research perspectives on policy and practice* (pp.285-296). Frankfurt am Main: Peter Lang.

Lahuerta, A. (2017). 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*, 1-12. DOI: 10.4067/s0718-48832017000100013

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. DOI: 10.1075/jicb.16013.lia

Lasagabaster, D. (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics Journal, 1*(1). p. 30-41. DOI: 10.2174/1874913500801010030

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. DOI: 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. DOI: 10.1093/applin/amaa037

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. DOI: 10.30827/digibug.54021

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, Sydney, Nova Scotia, Canada. Available from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.585.2454&rep=rep1&type=pdf>

MacIntyre, P. D., Baker, S. C., Clément, R., & Conrod, S. (2001). Willingness to communicate, social support, and language-learning orientations of immersion students. *Studies on Second Language Acquisition*, 23, 369-388. DOI: 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. DOI: 10.1111/1467-9922.00194

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. DOI: 10.1111/j.1540-4781.1998.tb05543.x

Marsh, D. (2002). *CLIL/EMILE-The European dimension: Actions, trends and foresight potential*. Jyväskula: UniCOM.

Marsh, D. (2013). *Content and Language Integrated Learning (CLIL). A Development Trajectory*. Universidad de Córdoba.

Mearns, T., 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, DOI: 10.1080/13670050.2017.1405906

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). Cham: Springer. DOI: 10.1007/978-3-319-11496-5

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. DOI: 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. DOI: 10.1016/j.tate.2006.04.024

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). Frankfurt am Main: Peter Lang. DOI: 10.3726/978-3-653-01829-5/7

Mitchell, R., Myles, F., & Marsden, E. (2019). *Second language learning theories* (4th ed.). New York: Routledge. DOI: 10.4324/9781315617046

Morinaj, J., Hadjar, A., & Hascher, T. (2020). School alienation and academic achievement in Switzerland and Luxembourg: a longitudinal perspective. *Social Psychology of Education 23*, 279–314. DOI: 10.1007/s11218-019-09540-3

Naber, R., & Lowie, W. (2012). Hoe vroeger, hoe beter? Een onderzoek naar de effectiviteit van vroeg vreemdetalenonderwijs. *Levende Talen Tijdschrift*, *13*(4), 13-21.

Nederlands Jeugdinstituut. *Cijfers over jeugd met een migratieachtergrond.* 7 June 2021. Available from <https://www.nji.nl/cijfers/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. DOI: 10.6018/ijes/2016/2/239611

Nikula T. (2017) CLIL: A European Approach to Bilingual Education. In: N. Van Deusen-Scholl & S. May (Eds.), *Second and Foreign Language Education. Encyclopedia of Language and Education* (3rd ed.). Cham: Springer. DOI: 10.1007/978-3-319-02246-8_10

Paran, A. (2013). Content and language integrated learning: Panacea or policy borrowing myth? *Applied Linguistics Review*, *4*(2), 317-342. DOI: 10.1515/applirev-2013-0014

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. Available from <https://dialnet.unirioja.es/servlet/articulo?codigo=6273210>

Pérez Cañado (2020): CLIL and elitism: myth or reality?, *The Language Learning Journal*, 48(1), 4-17. DOI: 10.1080/09571736.2019.1645872

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. Guerrini (Eds.), *CLIL across educational levels* (pp. 3–16). Oxford: Richmond.

Pérez-Vidal, C., & Roquet, H. (2015). CLIL in context: Profiling language abilities. In In M. Juan-Garau and J. Salazar Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 237-255). Cham: Springer. DOI: 10.1007/978-3-319-11496-5_10

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. DOI: 10.1016/j.learninstruc.2016.04.001

Rallo Fabra, L., and Jacob, K. (2015). Does CLIL enhance oral skills? Fluency and pronunciation errors by Spanish-Catalan learners of English. In M. Juan-Garau and J. Salazar Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 163-177*).* Cham: Springer. DOI: 10.1007/978-3-319-11496-5_10

Ruiz de Zarobe, Y. (2008). CLIL and foreign language learning: A longitudinal study in the Basque Country. *International CLIL Research Journal*, *1*(1), 60-73.

Rumlich, D. (2017). CLIL theory and empirical reality–Two sides of the same coin? *Journal of Immersion and Content-Based Language Education, 5*(1), 110-134. DOI: 10.1075/jicb.5.1.05rum

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). Newcastle upon Tyne: Cambridge Scholars.

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). Frankfurt am Main: Peter Lang.

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.

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*, 153-172. DOI:10.4995/rlyla.2019.10527

Smala, S. (2023). Situated emergence of CLIL. E. Codó, E. (Ed.), *Global CLIL: Critical, Ethnographic and Language Policy Perspectives* (pp. 52-73). New York: Routledge. DOI: 10.4324/9781003147374-4

Smiskova, H., Verspoor, M. H., & Lowie, W. (2012). Conventionalized ways of saying things (CWOSTs) and L2 development. *Dutch Journal of Applied Linguistics*, 1(1), 125–142. DOI: 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. DOI: 10.1007/s11159-017-9651-4

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. DOI: 10.1080/13670050.2018.1517722

Standaard Tweetalig vmbo. Nuffic: De Nederlandse organisatie voor internationalisering in onderwijs. Available from <https://www.nuffic.nl/onderwerpen/tweetalig-onderwijs/standaard-voor-t-vmbo>

Van Breukelen, G.J.P. (2013). ANCOVA Versus CHANGE From Baseline in Nonrandomized Studies: The Difference. *Multivariate Behavioral Research*, 48:895-922. DOI: 10.1080/00273171.2013.831743

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & W. Holland, R.W. (2010). The Implicit Prejudiced Attitudes of Teachers. *American Educational Research Journal*, 47:2, 497-527. DOI: 10.3102/0002831209353594

Verspoor, de Bot, & van Rein (2011). English as a foreign language. The role of out-of-school language input. In A. De Houwer and A. Wilton (Eds.) *English in Europe Today: Sociocultural and educational perspectives* (pp. 147-166) Amsterdam: John Benjamins.

Verspoor, M., Xu, X., & de Bot, C.J.L. (2013). *Verslag OTTO-2 aan Europees Platform*. Toegepaste Taalwetenschap, Universiteit Groningen.

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. DOI: 10.1075/jicb.3.1.01ver

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. DOI:

Westhoff, G. (2005). Talenquest: beloften en valkuilen. *Levende Talen Magazine*, *92*(4), 12-14.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. University of Hawaii at Manoa. DOI: 10.1017/s0272263101263050

Authors' postal and e-mail addresses:

Jenny Denman
Research Centre Urban Talent / Kenniscentrum Talentontwikkeling
Rotterdam University of Applied Sciences / Hogeschool Rotterdam
Postbus 25035
2001 HA Rotterdam
The Netherlands
J.L.Denman@hr.nl

Erik van Schooten
Research Centre Urban Talent / Kenniscentrum Talentontwikkeling
Rotterdam University of Applied Sciences / Hogeschool Rotterdam
Postbus 25035
2001 HA Rotterdam
The Netherlands
E.J.van.Schooten@hr.nl

Rick de Graaff
Department of Languages, Literature and Communication

University of Utrecht
Trans 10
3512 JK Utrecht
The Netherlands
R.degraaff@uu.nl

Table 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

Table 11: Simplified table, see Appendix for full tables. Results multi-level models for **Speaking proficiency (Ancova and Change)**, effects of CLIL and other variables. Standard errors between brackets. (posttest=z)

Model	**Model 8, Table 9**	**Model 8, Table 10**
Fixed part	**Ancova**	**Change**
Intercept	.060 (.080)	.080 (.076)
Speaking pretest (z)	.637*** (.044)	
non-CLIL (0=clil; 1=non-clil)	**-.214* (.094)**	**-.234* (.096)**
Grade 8 (grade 7=ref.cat.)	-.031 (.100)	-.038 (.115)
Grade 9	.272* (.117)	.292* (.127)
Pre-voc level (gm) (1=high; 5=low)	-.126** (.045)	-.144** (.041)
Random part		
School variance	.014 (.013)	
Class variance	.004 (.011)	.017 (.013)
Pupil variance	.341 (.030)	.341 (.030)
Total	.359	.358
prop. expl School variance	.176	
prop. expl Class variance	.556	.370
prop. expl Pupil variance	.003	.003
Prop. Expl total var	.024	.030
Deviance	518.296	
Model of reference and fit improvement	model 7	model 7
	X^2=4.875	X^2=5.475
	df=1	df=1
	p<.05	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) (Model of reference and fit improvement: see full tables 9 and 10 in appendix)

Table 12: Simplified table, see Appendix for full tables. Results multi-level models for **Speaking fluency (Ancova and Change)**, main effects of CLIL and other variables. Standard errors between brackets.

Model	**Model 8, Table 17**	**Model 8, Table 18**
Fixed part	**Ancova**	**Change**
Intercept	152.903 (8.756)	49.982 (9.071)
Words English pretest-gm	.864*** (.049)	
non-CLIL (0=clil; 1=non-clil)	**-20.628*** (8.550)**	**-13.969 (8.879)**
Grade 8 (grade 7=ref.cat.)	-18.828* (8.595)	-22.684* (9.408)
Grade 9	.088 (10.262)	-9.618 (10.603)
Pre-voc level (gm) (1=high: 5=low)	-9.674* (4.610)	-9.240# (4.845)
Random part		
School variance	279.436 (192.745)	263.569 (193.449)
Class variance	97.369 (85.283)	150.341 (103.688)
Pupil variance	1743.424 (151.583)	1769.842 (153.975)
Total variance	2120.229	2183.752
Deviance	3000.264	
Model of reference and fit improvement	model 7	model 7
	X^2=5.195	X^2=2.314
	df=1	df=1
	p<.05	p=n.s.
prop. expl. var. school level	.204	
prop. expl. var. class level	.394	
prop. expl. var. Pupil level	-	
prop. expl. var. total	.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) (Model of reference and fit improvement: see full tables 15 and 16 in appendix)

Appendix (online): Attachments 1, 2, & 3 and Tables 2 – 10 and 13 - 22

Attachment 1: Sample Speaking test

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.)

(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?

(Past) *What did you do last weekend?*
What did you do last summer?
What did you do yesterday evening?

(Future) *Tell me something you would like to do or try 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 2: 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 3: Sample Speaking anchor texts for holistic proficiency assessment (scores in unanimous agreement by all assessors in the assessment group)

SPEAKING ANCHOR TEXTS

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, 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 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[1]	7	.88	489	26.52 (6.38)	.94
1-M WTC London[1]	7	.89	490	27.32 (.29)	6.42
2-M WTC London[1]	7	.89	351	28.20 (.33)	6.18
0-M WTC School[1]	7	.87	489	16.45 (3.60)	1.09
1-M WTC School[1]	7	.87	488	16.94 (.16)	3.62
2-M WTC School[1]	7	.88	351	17.145 (.19)	3.51

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

Table 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

Table 3. 0-0.20 poor, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, 0.81-1 excellent (Albers, 2017)

Table 4: 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 5: 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 6: Results multi-level 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 between brackets)

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)
Prevoc 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	model 2	model 2		model 5	model 6	model 6
		X^2=78.705	X^2=.000	X^2=11.100		X^2=66.386	X^2=.000	X^2=7.860
		df=1	df=1	df=1		df=1	df=1	df=1
		p<.001	p=n.s.	p<.001		p<.001	p=n.s.	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 7: Results longitudinal multi-level models for **Willingness to Communicate in London**, effects of CLIL. Standard errors between brackets.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Fixed part	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**
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)
vmbo*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)
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 8: Results longitudinal multi-level models for **Willingness to Communicate at School**, effects of CLIL. Standard errors between brackets.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Fixed part	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**	**Coeff. (s.e.)**
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)
vmbo*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)
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 9: Results multi-level models for **Speaking proficiency (Ancova)**, effects of CLIL and other variables. Standard errors between brackets. (posttest=z)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**
Intercept	.035	.036 (.058)	-.017	-.062	-.008	.043	-.026	.060
	(.037)		(.088)	(.089)	(.071)	(.101)	(.077)	(.080)
Speaking pretest (z)	.732***	.695***	.697***	.679***	.690***	.679***	.670***	.637***
	(.038)	(.042)	(.040)	(.042)	(.040)	(.042)	(.042)	(.044)
non-CLIL (0=clil;						-.154		-.214*
1=non-clil)						(.106)		(.094)
Grade 8 (grade				.039			-.020	-.031
7=ref.cat.)				(.102)			(.106)	(.100)
Grade 9				.218			.234#	.272*
				(.127)			(.123)	(.117)
Pre-voc level (gm)					-.086#		-.102*	-.126**
(1=high; 5=low)					(.048)		(.048)	(.045)
Random part								
School variance			.041	.033	.021	.046	.017	.014
			(.029)	(.024)	(.019)	(.031)	(.016)	(.013)
Class variance		.054 (.024)	.012	.009	.014	.009	.009	.004
			(.014)	(.013)	(.015)	(.013)	(.013)	(.011)
Pupil variance	.390 .032)	.343 (.030)	.344	.343	.343	.342	.342	.341
			(.030)	(.030)	(.030)	(.030)	(.030)	(.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		model 1	model 2	model 3	model 3	model 3	model 3	model 7
and fit improvement		X^2=12.315	X^2=6.041	X^2=2.961	X^2=2.450	X^2=2.062	X^2=6.779	X^2=4.875
		df=1	df=1	df=2	df=1	df=1	df=3	df=1
		p<.001	p<.05	p=n.s.	p=n.s.	p=n.s.	p=n.s.	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 10: Results multi-level models for **speaking proficiency (change in z-scores)**, effects of CLIL and other variables. Standard errors between brackets.

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 13: Results multi-level models for **Speaking proficiency (Ancova)**, variables moderating the effect of CLIL. Standard errors between brackets. (Cito = scholastic aptitude test score)

Model **Fixed part**	**1** **Grade-1**	**2** **Grade-2**	**3** **Pre-voc levelrec -1**	**4** **Pre-voc levelrec -2**	**5** **Gender-1**	**6** **Gender-2**	**7** **Cito-1**	**8** **Cito-2**	**9** **Home language-1**	**10** **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)								
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 Model of reference and fit improvement	518.296	517.062 model 1 X^2=1.234 df=2 p=n.s.	518.296	518.068 model 1 X^2=.228 df=1 p=n.s.	517.037	516.576 model 4 X^2=.461 df=1 p=n.s.	274.267	274.173 model 6 X^2=.094 df=1 p=n.s.	517.464	517.191 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 14: Results multi-level models for **Speaking proficiency (Ancova)**, variables moderating the effect of CLIL. Standard errors between brackets. (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						
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 15: Results multi-level models for **Speaking proficiency (change scores)**, variables moderating the effect of CLIL. Standard errors between brackets. (Cito = scholastic aptitude test score)

Model **Fixed part**	**1** **Grade-1/ Pre-voc-levelrec-1**	**2** **Grade-2**	**3** **Pre-voc-levelrec-2**	**4** **Gender-1**	**5** **Gender-2**	**6** **Cito-1**	**7** **Cito-2**	**8** **Home language-1**	**9** **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)
Pupil variance	.409 (.036)	.409 (.036)	.409 (.036)	.409 (.036)	.408 (.036)	.334 (.039)	.332 (.039)	.411 (.036)	.410 (.036)
Deviance Model of reference and fit improvement	581.631	581.413 model 1 X^2=.218 df=2 p=n.s.	581.439 model 1 X^2=.192 df=1 p=n.s.	581.191	580.691 model 4 X^2=.500 df=1 p=n.s.	302.616	301.599 model 6 X^2=1.017 df=1 p=n.s.	580.588	580.184 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 16: Results multi-level models for **Speaking proficiency (change scores)**, variables moderating the effect of CLIL . Standard errors between brackets. (PE=Primary school English)

Model **Fixed part**	**1** **Years PE**	**2** **Years PE**	**3** **Frequency PE**	**4** **Frequency PE**	**5** **Years & freq: PE intensity**	**6** **Years & freq: PE intensity**	**7** **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
Deviance Model of reference and fit improvement	581.501	581.207 model 1 X^2=.294 df=1 p=n.s.	577.772	577.276 model 3 X^2=.496 df=1 p=n.s.	575.198	573.946 model 5 X^2=1.252 df=2 p=n.s.	571.388 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 17: Results multi-level models for **Speaking fluency (Ancova)**, main effects of CLIL and other variables. Standard errors between brackets.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**	**Ancova**
Intercept	142.284	142.055	137.331	141.164	138.720	144.095	144.490	152.903
	(2.885)	(6.082)	(9.136)	(9.982)	(8.471)	(9.627)	(8.987)	(8.756)
Words English	.933***	.910***	.890***	.898***	.890***	.867***	.897***	.864***
pretest-gm	(.046)	(.050)	(.046)	(.047)	(.046)	(.047)	(.047)	(.049)
non-CLIL						-16.989#		-20.628***
(0=clil; 1=non-clil)						(9.627)		(8.550)
Grade 8				-13.141			-17.594	-18.828*
(grade 7=ref.cat.)				(9.049)			(9.585)	(8.595)
Grade 9				-4.349			-3.160	.088
				(11.359)			(11.309)	(10.262)
Pre-voc level					-4.750		-8.087	-9.674*
(gm) (1=high: 5=low)					(5.014)		(5.104)	(4.610)
Random part								
School			470.957	513.610	376.358	454.273	351.361	279.436
variance			(313.277)	(328.322)	(263.675)	(294.699)	(241.657)	(192.745)
Class		759.809	203.532	157.429	211.729	147.052	160.591	97.369
variance		(260.881)	(121.498)	(105.871)	(123.851)	(102.354)	(106.621)	(85.283)
Pupil variance	2404.910	1742.464	1741.487	1742.272	1741.682	1744.046	1741.212	1743.424
	(200.062)	(151.658)	(151.496)	(151.545)	(151.550)	(151.694)	(151.476)	(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		model 1	model 2	model 3	model 3	model 3		model 7
reference and fit		X^2=49.026	X^2=11.496	X^2=1.979	X^2=.771	X^2=3.051		X^2=5.195
		df=1	df=1	df=2	df=1	df=1		df=1
improvement		p<.001	p<.001	p=n.s.	p=n.s.	p<.10		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 18: Results multi-level models for **Speaking fluency (change scores)**, main effects of CLIL and other variables. Standard errors between brackets.

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fixed part	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**	**Change**
Intercept	39.083	38.395	34.546	40.233	35.759	38.733	43.530	49.982
	(2.895)	(6.023)	(8.802)	(9.708)	(8.298)	(9.440)	(8.819)	(9.071)
non-CLIL (0=clil; 1=non-clil)						-10.245		-13.969
						(9.865)		(8.879)
Grade 8 (grade 7=ref.cat.)				-16.290			-20.857*	-22.684*
				(9.094)			(9.673)	(9.408)
Grade 9				-11.114			-10.200	-9.618
				(11.099)			(11.022)	(10.603)
Pre-voc level (gm) (1=high: 5=low)					-4.467		-7.897	-9.240#
					(5.076)		(5.107)	(4.845)
Random part								
School variance			417.087	475.851	345.574	391.820	329.171	263.569
			(291.347)	(310.479)	(253.849)	(274.745)	(232.004)	(193.449)
Class variance		740.834	241.459	169.007	247.430	224.022	171.961	150.341
		(256.270)	(134.989)	(110.559)	(136.367)	(129.052)	(111.072)	(103.688)
Pupil variance	2422.810	1767.300	1768.819	1769.857	1768.863	1769.778	1768.761	1769.842
	(201.551)	(153.811)	(153.893)	(153.956)	(153.945)	(153.977)	(153.898)	(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	model 2	model 3	model 3	model 3		model 7
		X^2=47.887	X^2=9.289	X^2=2.927	X^2=.693	X^2=1.051		X^2=2.314
		df=1	df=1	df=2	df=1	df=1		df=1
		p<.001	p<.01	p=n.s.	p=n.s.	p=n.s.		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 19: Results multi-level models for **Speaking fluency (number of words in English): Ancova,** variables moderating the effect of CLIL. Standard errors between brackets. (Cito = scholastic aptitude test score)

Model **Fixed part**	**1** **Grade-1**	**2** **Grade-2**	**3** **Pre-voc levelrec-1**	**4** **Pre-voc levelrec-2**	**5** **Gender-1**	**6** **Gender-2**	**7** **Cito-1**	**8** **Cito-2**	**9** **Home language-1**	**10** **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)								
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 Model of reference and fit improvement	3000.264	2998.581 model 1 X^2=1.683 df=2 p=n.s.	3000.264	3000.259 model 1 X^2=.005 df=1 p=n.s.	2999.945	2999.896 model 4 X^2=.049 df=1 p=n.s.	1772.543	1769.318 model 6 X^2=3.225 df=1 p<.10	2990.416	2990.376 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 20: Results multi-level models for **Speaking fluency (number of words in English): Ancova**, variables moderating the effect of CLIL. Standard errors between brackets. (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)
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)	1742.31 8 (151.488)	1731.14 9 (150.810)	1721.75 3 (150.014)	1727.61 4 (150.526)	1689.445 (147.197)	1682.980 (146.609)	1679.427 (146.303)
Total variance	2120.265	2116.37 0	2092.85 4	2067.74 9	2086.08 0	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	3000.09 1	2988.11 1	2986.46 5	2987.47 3	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 21: Results multi-level models for **Speaking fluency (number of words in English): change scores**, variables moderating the effect of CLIL. Standard errors between brackets. (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	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)		
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 Model of reference and fit improvement	3007.769	3004.694 model 1 X^2=3.075 df=2 p=n.s.	3007.705 model 1 X^2=.064 df=1 p=n.s.	3007.743	3007.724 model 4 X^2=.019 df=1 p=n.s.	1777.753	1774.138 model 6 X^2=3.615 df=1 p<.10	2998.116	2998.116 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 22: Results longitudinal multi-level models for **Speaking fluency (number of words in English): change scores**, variables moderating the effect of CLIL. Standard errors between brackets. (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)
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 Model of reference and fit improvement	3007.658	3007.333 model 1 X^2=.325 df=1 p=n.s.	2996.355	2994.043 model 3 X^2=2.312 df=1 p=n.s.	2995.308	2988.361 model 5 X^2=6.947 df=1 p<.01	2986.704 model 6 X^2=1.657 df=1 p=n.s.	2986.074 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)