A showdown between bilingual and mainstream education: the impact of language of instruction on learning subject content knowledge

Huub Oattes, Ruben Fukkink, Ron Oostdam, Rick de Graaff & Arie Wilschut

Routledge
Taylor & Francis Group

Check for updates

A showdown between bilingual and mainstream education: the impact of language of instruction on learning subject content knowledge

Huub Oattes [a], Ruben Fukkink [a,b], Ron Oostdam [a,b], Rick de Graaff [c] and Arie Wilschut [a]

[a]Centre for Applied Research in Education (CARE), Amsterdam University of Applied Sciences, Amsterdam, The Netherlands; [b]Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands; [c]Faculty of Humanities, Languages and Communication, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Bilingual education has become popular in many countries in the last two decades. It is generally acknowledged that learning a second language (L2) through subject content has a positive impact on students' L2 learning, but there is less agreement on whether this also applies to learning subject content knowledge in and through L2. This cross-sectional study compared Dutch pre-university mainstream and bilingual education students in grades 7 and 9 on a history knowledge test, taking into consideration the language of instruction and testing. Students were also tested on their motivation to learn and affinity with history, because of the alleged higher motivation bilingual education students bring to the classroom. Multilevel analyses showed that bilingual education students in grade 7 lagged behind in the English part of the test but performed at the same level in the Dutch part. 9th bilingual education graders on the other hand performed significantly better on the knowledge test than 9th mainstream graders in both L2 and L1, thus providing evidence for the non-detrimental effect of bilingual education on the acquisition of subject content knowledge.

Introduction

Bilingual education (BE) comes in many shapes, forms and sizes (Eurydice 2017) and Dutch bilingual secondary education also has some particular features that distinguish it from other countries. In the Dutch junior BE program (grades 7, 8, and 9) of secondary schools with a BE department, at least 50% of the curriculum is in English (L2).

In BE the preferred pedagogy is content and language integrated learning (CLIL) which implies that a foreign language is learned through learning subject content. Use of the mother tongue (L1) is actively discouraged in Dutch BE except for the first few months in grade 7, when BE students sometimes need the support of L1 to help them understand the teacher or express themselves (Oattes et al. 2018a). BE students get one extra English lesson per week in order to support their L2 development.

Over the past 20 years (inter)national research into BE has mainly focused on the development of BE students' L2-proficiency, and the evaluation of CLIL pedagogy by (subject) teachers and students. The focus in this study, however, is on learning subject content knowledge in Dutch BE. The main goal

---

CONTACT Huub Oattes ✉ h.oattes@hva.nl

is to determine the impact of the language of instruction on the acquisition of subject content knowledge in grades 7 and 9 of bilingual secondary education.

Studies focusing on the development of subject content knowledge in L2 show a diversity of results (Dalton-Puffer 2011; Perez-Canado 2017). Several examples can illustrate these different research conditions, and reported effects of BE on students' language and subject content learning. San Isidro and Lasagabaster (2018) conclude that various European studies confirm mostly positive or non-detrimental effects of BE. They performed a small-scale (N = 44) plurilingual Spanish study and they found no differences between BE and MS (mainstream) students in subject content learning. Stohler (2006) also showed no significant differences between Swiss BE and MS students on developing subject knowledge. However, the sample was very heterogeneous, containing students with French or German as L1 and test results from different grades and subjects. In addition, a Finnish study (Jappinen 2005) reported positive students' results on learning mathematics and science through CLIL. The longitudinal study involved 669 students between the ages of 7 and 15 (BE, N = 335, MS, N = 334) who were tested four times over a period of two years. All testing was done in L1 and participating schools had a lot of freedom in how to meet the demands as only the English language, the (curriculum) topics and the testing time slots were mandatory.

Recent German research (Dallinger et al. 2016) did not find differences in acquired academic history knowledge between MS and BE students. Yet, they considered this to be a negative result because the BE students received one extra lesson per week and therefore had more instruction to reach the same level of history knowledge as the MS students. Another German study among 191 grade 10 students in three Berlin schools showed a significant positive correlation between L2 proficiency and knowledge of basic concepts from a mix of three school subjects. However, the level of L2 proficiency caused some students to have problems understanding detailed subject content knowledge (Zydatiss 2007). This L2 proficiency problem also emerged in the amount and quality of subject knowledge obtained through L2 by German BE students who were clearly surpassed by MS students using their mother tongue (Coetzee-Lachmann 2009).

The above-mentioned BE outcomes on subject content learning ranges from positive to neutral to negative. The multitude and diversity of settings, measurement instruments, sample sizes and student populations make it difficult to reach general conclusions on the impact of the language of instruction on BE students' learning of subject content knowledge. Moreover, Bruton (2011) criticised the methodological inaccuracy and therefore the quality of BE studies. He mentions incorrect interpretation of outcome scores, neglecting extra CLIL support and a lack of comparability of the study samples, for instance in the selection of non-BE control groups.

Present study

The first research on BE subject learning outcomes dates from the early days of Dutch BE at the beginning of the millennium (Huibregtse 2001). It compared four cohorts of MS and BE students' final exam results in grade 12. Results showed that BE students performed better in L2 (English), and performed just as well when subject content knowledge (history and geography) was tested in L1. However, the impact of BE students being taught subject content in Dutch in the three upper years of Dutch secondary education was not taken into consideration.

Later research confirmed that CLIL students' English proficiency exceeds that of the MS students and found no negative effects for Dutch, geography and history (Admiraal, Westhoff, and De Bot 2006; Verspoor, de Bot, and Xu 2015). They compared BE students (N = 157) to MS students (N = 242) in grades 7 and 9 and gave special attention to the lead the BE students had in L2 development, scholastic aptitude, initial L2 proficiency, motivation, and out of school contact with L2.

The aim of the present study is to contribute to the existing general knowledge on the impact of Dutch BE on academic outcomes. Our goal is to determine if being taught in L2 has an impact on Dutch pre-university level students' development of historical subject content knowledge in grades 7 and 9. History was chosen because the particular language intensity of this school

subject makes it challenging to teach and to learn in L2 (Coffin 2006; De Oliveira 2011; Van Drie, Braaksma, and Van Boxtel 2015). Dutch BE students in grades 7 through 9 are taught history exclusively in English for three years, implying that they actively develop a historical vocabulary in L2, not in L1. This leads to the first research question: To what extent does the language of instruction influence the test performance of grade 7 and grade 9 bilingual education and mainstream students on history content knowledge? This question is answered by testing subject content knowledge in BE grades 7 and 9 in both English and Dutch, and only in the mother tongue in MS grades 7 and 9.

Many authors have suggested that BE students are a particular group because of their above-average intrinsic motivation to learn, their positive attitude towards school, their academic capacity and their willingness to invest extra time in mastering a second language (Mearns, De Graaff, and Coyle 2017; Sylvén and Thompson 2015; Verspoor, de Bot, and Xu 2015). Based on these findings about possible differences between MS and BE students, a second research question was formulated: Are there differences between mainstream and bilingual education students in motivation to learn and attitude towards the school subject history, and if so, is this related to bilingual education students' performance when tested on history content knowledge?

Method

Design

In this study we applied a cross-sectional design to examine whether the use of English (L2) as the sole language of instruction and communication in history class influenced BE students' development of curricular historical knowledge, and their ability to apply this knowledge in both English and Dutch.

Sample

Twelve schools, approximately 10% of all Dutch BE schools, participated in the study. The schools were spread across the Netherlands and form a representative sample of Dutch BE schools. They have a mix of BE experience (between 6 and 19 years, average 12.8 years), and BE teachers with level C1 or C2 (CEFR) in English language proficiency and CLIL certification. Within each school 2–3 teachers participated, summing up to a total of 28 teachers.

Within schools some participating history teachers taught in one grade, others in both (only in grade 7, $N = 8$; only in grade 9, $N = 14$; in both grades, $N = 6$). Some of these teachers only taught history in BE classes (in grade 7, $N = 3$; in grade 9, $N = 5$; in both grades, $N = 3$), some only in MS classes (in grade 7, $N = 3$; in grade 9, $N = 5$; in both grades, $N = 2$), and others taught both BE and MS students (in grade 7, $N = 2$; in grade 9, $N = 4$; in both grades, $N = 1$).

Some BE teachers who also taught in MS used the same textbook, the MS version in Dutch and an identical English version for BE students for practical reasons (in grade 7, $N = 6$; in grade 9, $N = 6$). These history textbooks are all based on the Dutch history curriculum. The other BE teachers used alternative English, Anglo-American, and/or self-developed textbooks often adding self-developed teaching materials to stay in line with MS and the Dutch history curriculum (in grade 7, $N = 8$; in grade 9, $N = 15$). Both MS and BE streams spend an equal amount of time (100–120 min) per week on the history curriculum.

The final study sample consisted of 1127 students; 19 classes of 7th graders and 24 classes of 9th graders, both of pre-university level, which were divided into four subgroups. Grade 7 was split up in mainstream (Dutch, $N = 176$, male: 51%), and bilingual stream (English, $N = 324$, male: 55%). Grade 9 also was divided into mainstream (Dutch, $N = 291$, male 46%), and bilingual stream (English, $N = 336$, male 47%). In the mainstream, three grade 7 classes ($N = 86$) were excluded from participating because they unexpectedly also contained unknown numbers of students of a non-pre-university level. This exclusion resulted in a somewhat skewed distribution of MS versus BE students in grade 7.

The students in grade 7 and grade 9 are from two different cohorts and the test results therefore represent two different groups of students. The age of students in grade 7 is 12-13, and 14–15 in grade 9. The distribution of boys and girls in MS and BE within grades was mostly evenly matched, in both grade 7 (MS girls 49%; BE girls 44%) and grade 9 (MS girls 54%; BE girls 53%). This also goes for the distribution of country of birth of parents (Dutch born versus non-Dutch born): in grade 7 (MS Dutch born 76%; BE Dutch born 75%) and grade 9 (MS Dutch born 81%; BE Dutch born 77%). Concerning the home language (Dutch or other) the pattern is almost the same: grade 7 (MS Dutch 76%; BE Dutch 75%) and grade 9 (MS Dutch 81%; BE Dutch 77%) respectively.

Measurements

History Knowledge Test (HKT)

For this study, two history knowledge tests (HKTs) were developed for grade 7 and 9. These consisted of multiple-choice questions about subject content knowledge fitting in with the Dutch history curriculum. Multiple-choice items were chosen so all students could fully concentrate on reading subject content and prevent possible cognitive overload by having to write answers. The main attainment goal in upper secondary history education is to develop students' ability in historical thinking and reasoning, for example understanding causal connections. They therefore need to be able to use a chronological frame of reference that is central in history teaching in grades 7 through 9 (Wilschut 2015). This framework consists of ten eras, each with four or five typical general features that in turn are made up of multiple historical events, developments, concepts and actors. For example, era 3 is called Era of Monks and Knights, 500–1000 AD and is characterised by four general features: Christianisation of Europe, Rise of Islam, Domains and serfdom, and Feudalism. Each of these topics is taught within a chronological frame and based on related key concepts using historical actors and their involvement in historical events to demonstrate relevant causal connections. In general, the chronological framework determines the content of the curriculum, but textbook writers are free to choose their own accents, and that also applies to the history teachers.

The HKT in grade 7 focused on three eras that were taught that same year. In grade 9 the HKT focused on seven eras that were taught in earlier years (grade 7, grade 8) and two in the year of the test (grade 9). Both HKTs also contained questions with a focus on initial historical reasoning, or the ability to understand and use historical knowledge (historical events, developments, concepts and actors) in a new context (see examples in Table 1).

With the help of three experienced history teachers, multiple-choice items were developed that enabled comparison of the acquired historical knowledge of students in grade 7 and grade 9. Each question had three answer options. Nearly all items were derived from the chronological frame of reference with known eras, and their related general features (see Table 1). Because historical subject matter is inherently diverse and heterogeneous the HKT covers a broad and multifaceted reference area in both time and content. The test was not created for measuring in depth knowledge of underlying constructs or specific knowledge domains, but for testing a representative selection of historical subject matter from a broad curricular domain.

The HKT was first pilot tested in ten schools (139 7th graders and 216 9th graders). After analysing the results from the pilot phase (p- and Rit-values for individual items and Cronbach's alphas), and the feedback from experts (the history teachers involved), a set of 40 test items per grade was established. Next, the HKT for each grade was subdivided into two sets (part A and part B) of comparable size (20 test items) and similar content. The items were evenly distributed over the various eras that were tested, i.e. three eras in grade 7 and seven eras in grade 9.

For both grades part A was translated from Dutch into English. To check whether the test halves of part A within grades are identical in L1 (Dutch) and L2 (English), a back translation was done to check for possible deviations. For both grades, this procedure resulted in parallel versions of Part A (in Dutch and English) and one version of Part B (in Dutch).

Table 1. Examples of items from the History Knowledge Test (HKT), Personal Motivation Scale (PMS) and School Subject Experience Scale (SSES).

	Example	Answer options: multiple choice or Likert scale
HKT grade 7	The oldest civilisations first formed along which rivers?	(1) Danube, Nile, Amazon (2) Euphrate, Danube, Indus (3) Nile, Euphrate, Tigris*
HKT grade 7	Which groups of subjects did the Egyptian pharaoh use to govern his land?	(1) Farmers, soldiers, priests (2) Priests, slaves, soldiers (3) Soldiers, civil servants, priests*
HKT grade 9	Why did hunter-gatherers have to live as nomads?	(1) To survive they had to travel around following herds of wild animals* (2) Since there wasn't enough food, they travelled from city to city (3) In the winter there are no carrots, berries or mushrooms to be found
HKT grade 9	Which statement about the First World War is true?	(1) The First World War created the direct cause for the start of the Second World War (2) During the First World War Germany was occupied by the allies (3) Shortly after the First World War the League of Nations was established*
HKT Historical reasoning grade 7 and 9	During the German occupation the Jewish Frank family went into hiding in Amsterdam. Young Anne kept a diary during two years until the family was arrested by the Germans. For which research is her diary NOT suitable?	(1) Life in the city for Jewish persons in hiding (2) How Jewish youngsters experienced living in hiding (3) German sanctions against people assisting Jews in hiding*
PMS Self-efficacy	I am sure that this year everything will work out fine at school.	1 = strongly disagree - 6 = strongly agree
PMS Intrinsic value	I prefer challenging assignments so that I can learn new things.	1 = strongly disagree - 6 = strongly agree
PMS Performance approach	I enjoy being the only one who knows an answer.	1 = strongly disagree - 6 = strongly agree
SSES History & Enjoyment	I enjoy history lessons.	1 = strongly disagree - 6 = strongly agree
SSES History & Difficulty & Anxiety	I find it difficult to understand the history textbook.	1 = strongly disagree - 6 = strongly agree

Note: *correct answer option HKT.

Per grade, all forty items in both streams were administered in the same order. In MS all items were in Dutch, whereas in the BE stream the items in Part A were in English and the items in Part B were in Dutch. This approach created the opportunity to examine BE student's ability to deploy their mother tongue in answering questions about history content knowledge. Per grade several items with low item total correlations ($r_{it} < .30$) were removed. Close inspection of these items revealed that in some cases more than one answer options was justifiable. In others the level of complexity was too high, for instance when the application of both detailed subject and general academic knowledge was required to select the correct answer. This eventually resulted in an HKT for grade 7 containing 31 items ($\alpha = .64$) and an HKT for grade 9 with 30 items ($\alpha = .69$) with adequate internal consistency (Taber 2016) and with coverage of all eras.

Personal Motivation Scale (PMS)

For the personal motivation test, a survey was put together that consisted of three previously validated and often used motivation constructs: (a) Self-efficacy: academic self-efficacy enhances personal accomplishments and a person's sense of well-being, consisting of 6 items, α = .77 (Midgley et al. 2000); (b) Intrinsic value: relates to the value that students attribute to learning new subject knowledge and skills at school, consisting of 7 items, α = .83 (Pintrich and De Groot 1990), and (c) Performance approach, or the drive to try to outperform other students, consisting of 6 items, α = .85 (Seegers, Van Putten, and De Brabander 2002). Following the recommendation of Sturgis, Roberts, and Smith (2014), we used a 6-point Likert scale (1 = strongly disagree, 6 = strongly agree) to avoid students choosing a neutral position when rating the statements.

School Subject Experience Scale (SSES)

Students' attitude towards a school subject may influence their test results. We therefore conducted a small-scale survey on the personal appreciation of the school subject history. The widely used and validated School Subject Experience Scale (SSES) (Martinot, Kuhlemeier, and Feenstra 1988; Otten and Boekaerts 1990) addresses four affective attitudes students may have towards a school subject. Two affective subscales were included: perceived enjoyment (8 items, α = .88) and perceived difficulty & anxiety (8 items α = .80). Again, a 6-point Likert scale (1 = strongly disagree, 6 = strongly agree) was used.

Procedure

The HKT was directly based on one and three years of history teaching respectively. Students participating in the study were not informed beforehand about the HKT and had to rely solely on their ready knowledge during test taking. BE students were told at the beginning of the test that part B would be in Dutch.

All participants had sufficient time to complete the test and questionnaires, which were administered in May 2018 by the history teacher who had received instructions on how to supervise the test procedure. It took students about 50 min to complete the entire test and rate thirty-five statements that were combined in a booklet. Participants were guaranteed that schools, history teachers and the results of the individual students would be anonymised.

Data analysis

In our sample, students (level 1) were nested in classes (level 2), which were nested in schools (level 3). We therefore analysed our data with multilevel regression analysis, which takes into account the hierarchical, clustered structure of the data, for grade 7 and grade 9 separately.

The two outcome measures were Historical Knowledge part A (HKT-A) and Historical knowledge part B (HKT-B). To adjust for possible differences between the bilingual education sample and the mainstream group, we included student motivation, perceived self-efficacy, intrinsic value, performance approach, enjoyment (in history) and difficulty & anxiety (with history) as covariates. We also included students' gender, parents' country of birth, and home language as demographic covariates, as these could influence the outcome of the test results.

We tested the statistical significance of the differences in outcome measures at the conventional alpha level of .05. An a priori analysis indicated that power was adequate (β= .80) for finding small-to-medium differences between the BE group and MS group ($ES = 0.35$) at the conventional level of α = .05 (two-sided) with a sample of at least 130 students in each condition. The clustering effect and the effect of covariates were not yet included in this a priori analysis. In view of the fact that in the current study the final size of the sample per condition was much larger than 130 students, there was sufficient power to demonstrate small-to-medium effects (Cohen 1988).

We tested the statistical significance of BE (i.e. is there a difference between BE and MS students?) and the interaction effect of BE with gender (i.e. are there differential, gendered BE effects?) in each model at .05. Finally, model fit was determined with the log-likelihood ratio (-2LL), Aikake Information Criterion and Bayesian Information Criterion (BIC). The fit of a final model with BE as main variable of interest and all covariates was compared with the fit of a basic model without any predictor ('null model') by comparing the differences in model fit (i.e. Δ-2LL, ΔAIC, ΔBIC). Cohen's d was calculated as the effect size, based on the observed means and standard deviations. Following Cohen's rule of thumb, effect sizes can be classified as small, medium or large with corresponding values of 0.20, 0.50 and ≥0.80 (Cohen 1988).

Results

Table 2 shows the HKT results for MS and BE students in grade 7 and grade 9 and Table 3 indicates the mean differences between the MS and BE group for the affective measures. In the statistical analysis related to our research questions, we take differences between the BE and MS sample into account.

Multilevel regression analyses were used to test differences between the BE and MS groups on the HKT (part A and B), taking into account the (generally small) differences between the groups. Table 4 shows the results of the final model in which the five affective variables were added as covariates; three related to motivation (self-efficacy, intrinsic value and performance approach) and two related to subject appreciation (enjoyment and difficulty & anxiety). The predictors from our model improved the model fit significantly, as indicated by the statistically significant decrease of the log-likelihood, AIC and BIC scores ($p < .001$).

The results show that in grade 7 MS students outperformed BE students on HKT-A (i.e. with the items for the BE group in English). For HKT-B (i.e. the items for both groups in Dutch), no significant difference was found. Results in grade 9, however, show a significantly different outcome as the BE students outperformed the MS students in both part A (in English) and part B (in Dutch) of the HKT.

In our models, the individual student background variables 'place of birth of parents', and 'home language' showed no significant relationship with the HKT results, not for part A nor part B, neither for grade 7 nor for grade 9. Boys performed significantly better than girls in grade 7 on the A part (English) and in grade 9 on the B part (Dutch) of the HKT. For both grades some significant correlations can be observed for 'self-efficacy', 'enjoyment', and perceived 'difficulty & anxiety'.

Conclusion and discussion

The first research question of this study was posed to determine to what extent the language of instruction influences the test performance of grade 7 and grade 9 BE and MS students on history content knowledge.

Results show that in grade 7 BE students were outperformed by MS students on part A of the HKT which was taken in Dutch for MS or English for BE. Concerning part B (in Dutch for all students) no differences were found. In grade 9 significant differences in favour of the BE students were found on both parts of the HKT and thus also in both languages. Overall, boys performed better than girls, while no systematic patterns occurred in relation to the background variables 'parents born in the Netherlands' and 'home language'. This is in line with other studies that also showed that boys perform slightly better in history than girls in primary and secondary education (cf. Cito 2019; Van der Kaap and Visser 2016).

The second research question focused upon students' motivation to learn and attitude towards the school subject history in relation to their performance on the knowledge test. Results of the multilevel regression analyses, which take into account possible differences in students' affective measures between the BE and the MS samples, showed some significant but very small correlations that have little meaning. BE students perceived a slightly stronger sense of 'self-efficacy', 'enjoyment', and 'difficulty & anxiety' than their MS peers. Considering the scores on the HKT, we established that

Table 2. Descriptive statistics for outcome measures for mainstream (MS) and Bilingual Education (BE) in grades 7 and 9: Means (and standard deviations) broken down by HKT-A (Dutch or English), HKT-B (Dutch), and student background variables.

	Grade 7				Grade 9			
	HKT-A (Dutch/English)		HKT-B (Dutch)		HKT-A (Dutch/English)		HKT-B (Dutch)	
Min-max	0–14		0–17		0–15		0–15	
	MS N = 176	BE N = 324	MS N = 176	BE N = 324	MS N = 291	BE N = 336	MS N = 291	BE N = 336
Mean *(SD)*	9.27 (2.10)	7.50 (2.04)	10.81 (2.97)	10.38 (2.61)	8.79 (2.57)	9.78 (2.22)	9.42 (2.57)	10.14 (2.46)
Girl	9.13 (1.94)	7.22 (1.98)	10.51 (2.89)	10.18 (2.51)	8.52 (2.43)	9.46 (2.16)	8.95 (2.62)	9.69 (2.25)
Boy	9.38 (2.21)	7.71 (2.04)	11.06 (3.04)	10.57 (2.68)	9.11 (2.16)	10.14 (2.23)	9.92 (2.43)	10.61 (2.61)
Dutch-born parents	9.33 (2.11)	7.57 (2.06)	10.86 (2.87)	10.45 (2.55)	8.78 (2.30)	9.75 (2.21)	9.37 (2.53)	10.21 (2.46)
Non-Dutch born parents	9.09 (2.07)	7.30 (1.97)	10.65 (3.31)	10.18 (2.77)	8.84 (2.41)	9.87 (2.27)	9.65 (2.74)	9.91 (2.44)
Dutch home language	9.36 (2.08)	7.48 (2.01)	10.81 (2.95)	10.40 (2.59)	8.77 (2.38)	9.76 (2.25)	9.47 (2.60)	10.15 (2.47)
Non-Dutch home language	8.71 (2.18)	7.81 (2.29)	10.79 (3.18)	10.19 (2.77)	9.03 (1.75)	10.00 (1.89)	9.00 (2.27)	10.10 (2.40)

Table 3. Descriptive statistics for mainstream (MS) and bilingual education (BE) in grades 7 and 9: Means (and standard deviations) for self-efficacy, intrinsic value, performance approach, enjoyment, and difficulty & anxiety.

Min-max	Self-efficacy 6–36		Intrinsic value 7–42		Performance approach 6–36		Enjoyment 8–48		Difficulty & Anxiety 8–48	
	MS N = 176	BE N = 324	MS N = 176	BE N = 324	MS N = 176	BE N = 324	MS N = 176	BE N = 324	MS N = 176	BE N = 324
Grade 7										
Mean *(SD)*	27.54 (3.70)	27.72 (3.96)	30.44 (5.34)	31.52 (4.56)	23.09 (5.72)	22.74 (6.46)	29.68 (8.61)	32.72 (7.79)	34.86 (6.10)	34.68 (6.56)
Girl	27.54 (3.71)	27.46 (4.19)	30.96 (5.05)	32.11 (4.43)	22.63 (5.69)	21.94 (6.60)	29.04 (8.82)	32.62 (8.16)	33.68 (6.13)	33.34 (6.88)
Boy	27.64 (3.74)	27.89 (3.76)	30.06 (5.60)	30.95 (4.60)	23.58 (5.81)	23.44 (6.30)	29.94 (8.61)	32.72 (7.50)	35.98 (5.93)	35.72 (6.09)
Dutch-born parents	27.71 (3.45)	27.60 (4.10)	30.17 (5.09)	31.62 (4.38)	22.96 (5.89)	22.32 (6.60)	29.95 (8.82)	32.58 (7.50)	35.13 (6.11)	34.50 (6.53)
Non-Dutch born parents	27.02 (4.41)	28.06 (3.55)	31.36 (6.09)	31.24 (5.09)	23.49 (5.20)	23.95 (5.92)	28.56 (8.24)	33.15 (8.66)	34.05 (6.08)	35.18 (6.69)
Dutch home language	27.74 (3.49)	27.76 (3.97)	30.56 (5.17)	31.56 (4.54)	23.10 (5.73)	22.76 (6.39)	30.02 (8.97)	32.76 (7.61)	34.99 (6.15)	34.71 (6.46)
Non-Dutch home language	26.33 (4.74)	27.33 (3.92)	29.62 (6.51)	31.17 (4.79)	23.00 (5.76)	22.58 (7.16)	27.17 (6.32)	32.39 (9.45)	34.09 (5.82)	34.40 (7.60)
Grade 9	MS N = 291	BE N = 336	MS N = 291	BE N = 336	MS N = 291	BE N = 336	MS N = 291	BE N = 336	MS N = 191	BE N = 336
Mean *(SD)*	26.98 (4.46)	26.99 (3.91)	28.89 (5.16)	29.78 (4.48)	23.05 (6.20)	24.23 (5.75)	30.20 (8.24)	31.15 (8.61)	34.58 (6.24)	35.12 (6.31)
Girl	25.69 (4.50)	26.78 (3.97)	29.45 (4.46)	30.77 (3.99)	22.42 (6.27)	23.93 (5.77)	28.83 (8.45)	29.69 (8.85)	33.20 (6.20)	33.08 (6.23)
Boy	26.68 (4.37)	27.21 (3.84)	28.27 (5.93)	28.65 (4.83)	23.92 (5.98)	24.60 (5.69)	32.03 (7.51)	32.53 (8.04)	36.43 (5.82)	37.36 (5.67)
Dutch-born Parents	25.81 (4.57)	26.72 (3.83)	28.65 (5.23)	29.53 (4.29)	22.88 (6.26)	23.92 (5.86)	30.29 (8.31)	31.00 (8.50)	34.83 (6.26)	35.07 (6.14)
Non-Dutch born parents	27.20 (3.82)	27.90 (4.06)	30.00 (4.72)	30.62 (4.98)	23.76 (5.91)	25.33 (5.25)	29.84 (7.99)	31.64 (8.98)	33.54 (6.07)	35.28 (6.90)
Dutch home language	25.99 (4.47)	26.83 (3.84)	28.70 (5.13)	29.57 (4.44)	22.79 (6.21)	24.05 (5.70)	29.94 (8.38)	30.85 (8.61)	34.46 (6.39)	34.87 (6.24)
Non-Dutch home language	26.80 (4.38)	28.59 (4.32)	30.50 (5.20)	31.93 (4.42)	25.23 (5.71)	26.03 (6.03)	32.46 (6.68)	34.07 (8.16)	35.55 (4.78)	37.62 (6.52)

Table 4. Outcomes for multilevel models: Historical Knowledge Tests in Dutch/English (HKT-A) and Dutch (HKT-B) in Bilingual Education Grades 7 and 9.

	Grade 7		Grade 9	
	HKT-A (Dutch/English)	HKT-B (Dutch)	HKT-A (Dutch/English)	HKT-B (Dutch)
Fixed part				
Intercept	4.89***(0.83)	6.80*** (1.11)	4.12*** (0.92)	6.26*** (1.00)
Type of school – (BE)	−1.54*** (0.36)	−0.54 (0.57)	0.85** (0.36)	0.68* (0.29)
Gender (boy)	0.40* (0.24)	−0.01 (0.32)	0.27 (0.27)	0.61** (0.22)
Type of school*Gender	0.04 (0.37)	0.05 (0.48)	−0.02 (0.37)	−0.36 (0.40)
Parents born in Netherlands	0.18 (0.24)	0.35 (0.32)	0.21 (0.26)	0.09 (0.28)
Home language Dutch	−0.42 (0.35)	−0.44 (0.46)	0.24 (0.35)	0.59 (0.38)
Self-efficacy	0.03 (0.03)	0.03 (0.04)	0.07** (0.03)	0.10** (0.03)
Intrinsic value	0.00 (0.02)	−0.04 (0.03)	0.02 (0.02)	−0.04 (0.03)
Performance approach	−0.03 (0.02)	−0.02 (0.02)	0.00 (0.02)	−0.02 (0.02)
Enjoyment	0.04** (0.01)	0.08*** (0.02)	0.02 (0.01)	0.04* (0.02)
Difficulty & Anxiety	0.04 (0.02)	0.06* (0.02)	0.08*** (0.02)	0.06* (0.02)
Random part				
Residual	3.61*** (0.24)	6.21*** (0.42)	4.16*** (0.27)	5.01*** (0.32)
School*Class	0.39* (0.16)	1.22** (0.46)	0.37* (0.17)	0.29 (0.15)
Model fit				
−2LL (df = 3) basic model	2506.9	2836.1	2794.7	2919.9
−2LL (df = 13) final model	1927.2	2187.8	2243.3	2333.8
Δ-2LL(Δdf = 10)	579.7***	648.3***	551.4***	586.1***
AIC (df = 3) basic model	2512.9	2842.1	2800.7	2925.9
AIC (df = 13) final model	1953.2	2213.8	2269.3	2359.8
ΔAIC (Δdf = 10)	559.7***	628.3***	531.4***	566.1***
BIC (df = 3) basic model	2526.0	2855.2	2814.0	2939.2
BIC (df = 13) final model	2006.9	2267.5	2324.6	2415.1
ΔBIC (Δdf = 10)	519.1***	587.7***	489.4***	524.1***

Note: * = .05, ** = .01, *** = .001.

BE students in grade 7 had some difficulty when tested in English and this negatively affected their score. However, when the language of testing changed to L1 they were able to keep up with MS 7th graders. The lower test results of the BE students after nearly one year of CLIL history teaching in L2, seem to be the result of BE students' difficulties with the language of testing. A likely explanation is that they lagged behind in their ability to fully comprehend subject-specific English questions and pre-selected answers. The first year of BE is challenging since most students arrive with a limited L2 vocabulary and have to process a great deal of new information. Developing subject-specific and general academic vocabulary in English in grade 7 takes up a lot of time and energy, for both students and teacher (Oattes et al. 2018a). However, the result that grade 7 BE students performed at a similar level on the Dutch part of the knowledge test as the MS students, indicates that the focus on L2 apparently does not negatively affect their understanding of historical knowledge. This substantiates the assumption of a language backlog in English in grade 7.

A language backlog can no longer be observed in grade 9 since BE students performed significantly better on both parts of the HKT than the MS students. They performed better in L2 than MS students did in their L1 (Part A). On top of that the BE students also outperformed the MS students in L1 (Part B) despite being taught history only in English as the language of instruction for three years. Teaching history in L2 therefore does not seem to impede these students' use of the mother tongue when it comes to understanding subject-specific knowledge. It seems that BE students develop understanding of the subject-specific vocabulary in L1 autonomously, e.g. at home or outside the classroom, although some BE history teachers may occasionally apply code-switching to support their CLIL teaching (Oattes et al. 2018b).

The finding that BE students in grade 9 outperformed their MS counterparts in both parts of the knowledge test, also provides evidence for the non-detrimental effect of BE on the acquisition of subject content knowledge. Therefore, it could be concluded that CLIL teaching in BE has a positive effect on learning subject content in L2. These results may support earlier research suggesting that

CLIL pedagogy stimulates deep learning (Coyle, Hood, and Marsh 2010; Lublin 2003) as the use of L2 in the classroom requires students to focus on the content and the L2, to process content in the L2 (meaning) and to use both adequately (output) to communicate on content matters (Bertaux et al. 2010). However, as no classroom observation or curriculum plans were analysed in this study, we cannot attribute the positive outcomes for the BE group to specific CLIL pedagogy.

In the multilevel regression analyses with the scores on the knowledge test as outcome measures and the affective variables as covariates, we observed some significant, although very small, relationships for grade 7 and/or grade 9 students between their HKT scores and the scores on the scales for self-efficacy (grade 9), enjoyment and perceived difficulty & anxiety (grades 7 and 9). For example, students who considered the subject history more difficult, performed, as expected, somewhat lower on the knowledge test. However, students in grade 9, who assessed their self-efficacy to be high, scored slightly lower than expected on the knowledge test. Perhaps BE students had overconfidence in their self-efficacy (cf. Dunlosky and Rawson 2012) or underestimated the difficulty of the subject-specific knowledge.

Limitations

The design of this study was cross-sectional which made it impossible to follow the development of individual students regarding the development of their subject content knowledge in order to discover longitudinal relations between background and school variables. A longitudinal design could have offered greater insight into the development of the BE students' subject content knowledge.

Furthermore, we did not map out the content of the history lessons exhaustively. Some teachers may have focused more strongly on certain historical topics than others. Likewise, we do not know whether individual students had access to additional support to improve their performance level, for example through parental coaching or professional homework guidance. Although most teachers follow the curriculum fairly closely (Oattes et al. 2018a), making systematic bias between MS and BE unlikely, deviations at school and/or teacher level cannot be completely excluded.

In this study, a multiple-choice test was used to determine the acquired history knowledge of students. In this way speaking and writing in L2 were deliberately excluded to prevent students being distracted or slowed down due to a possible lack of active L2 proficiency. It is conceivable that other test formats in which a greater appeal is made on L2 proficiency will show different results. Future research into the effects of different test formats on students' performance regarding their acquisition of subject content knowledge can provide more insight. Also, the HKT was held unannounced, which is not common in Dutch education, and test results could have been different if students had had time to prepare for the HKT. No standardised L2 measure was performed and the obtained student grades for English and history were incomparable due to schools using different grading systems.

This study seems consistent with earlier research claiming that BE students bring a stronger pre-existing motivation to learn to the classroom than MS students (Admiraal, Westhoff, and De Bot 2006; Mearns, De Graaff, and Coyle 2017; Verspoor, de Bot, and Xu 2015). The intrinsic value of BE students in both grade 7 and grade 9 is significantly stronger than of the MS peers (respectively $M = 31.52$, $SD = 4.56$; $M = 30.44$, $SD = 5.34$; $t(468) = -2.32$, $p = .021$ in grade 7 and $M = 29.78$, $SD = 4.48$; $M = 28.89$, $SD = 5.16$; $t(598) = -2.26$, $p = .024$ in grade 9). BE students in grade 7 further derive significantly more pleasure from history, but show comparable scores to the MS 7th graders when it comes to 'self-efficacy', 'performance approach' and 'difficulty & anxiety'. BE students in grade 9 show an even greater distinction compared to the MS 9th graders by also outperforming them significantly on 'self-efficacy' (respectively $M = 26.99$, $SD = 3.91$; $M = 26.08$, $SD = 4.46$; $t(600) = -2.68$, $p = .008$) and 'performance approach' (respectively $M = 24.23$, $SD = 5.75$; $M = 23.05$, $SD = 6.20$; $t(602) = -2.44$, $p = .015$).

On the other hand, they have lost their sense of 'pleasure' compared to BE grade 7 and show no difference with the mainstream group on 'difficulty & anxiety'. However, the measurement of

affective variables in the present study was primarily examined in relation to the school subject history. In other contexts, in relation to other subjects, other differences between MS and BE students may be observed.

Implications

Considering the limitations mentioned, the results of this study support the conclusion that BE students can apply subject content knowledge in another language than the language of instruction. Although BE students learned history content knowledge using L2 as language of instruction, they were able to apply this knowledge not only in L2 but also in L1.

The skills and language of testing in L2 in grade 7 need the BE teachers' special attention as the BE students seem to find this difficult. When subject knowledge is tested in L2 a flexible language approach is advisable as some students (partly) need a dictionary to get a grip on a L2 text or the permission to use L1 to demonstrate their subject knowledge.

In this study, content knowledge seems to be acquired independently of the language of instruction. It raises the question whether a more gradual transition is preferable after grade 9 when the language of instruction becomes exclusively L1 for all exam subjects and all streams in the Netherlands. Given the demonstrated non-disruptive effect of BE on the learning of subject content knowledge in this particular research setting (passive language use, multiple-choice questions), it would be interesting to find out if the outcome holds when BE students need to demonstrate an active command of L2 and L1, for instance when questions require written answers in the appropriate academic and subject-specific language.

In future research, the effects of different languages of instruction on the learning of content subject knowledge need to be examined more closely. In Dutch society, the English language is popular and omnipresent, but outside school Dutch is the language of communication for students. In case the language of instruction is less popular and students are less familiar with it, this may have a negative impact on the learning of subject content knowledge. This is in contrast with bilingual societies in which the languages of instruction in BE are also languages of communication. Within that bilingual context the language of instruction will probably not impede the learning of subject content knowledge.

Against the background of Dutch BE, the present study can be seen as a confrontation between two distinct groups of MS and BE students, who learned subject content knowledge through different languages of instruction. Based on the overall outcome, we believe Dutch BE has proven its strength in this showdown by catching up in grade 7, and coming out on top in grade 9.

Disclosure statement

ORCID

Huub Oattes 🄳 http://orcid.org/0000-0001-7615-5613
Ruben Fukkink 🄳 http://orcid.org/0000-0001-6212-9553
Ron Oostdam 🄳 http://orcid.org/0000-0003-4701-0153
Rick de Graaff 🄳 http://orcid.org/0000-0001-6154-7286
Arie Wilschut 🄳 http://orcid.org/0000-0001-9864-651X

References

Admiraal, W., G. Westhoff, and K. De Bot. 2006. "Evaluation of Bilingual Secondary Education in the Netherlands: Students' Language Proficiency in English." *Educational Research and Evaluation* 12 (1): 75–93.

Bertaux, P., C. Coonan, M. Frigols-Martin, and P. Mehisto. 2010. *The CLIL Teacher's Competences Grid*. http://www.ccn-clil.eu.

Bruton, A. 2011. "Is CLIL so Beneficial, or Just Selective? Re-Evaluating Some of the Research." *System* 39 (4): 523–532.

Cito. 2019. Examenverslagen 2015–2019, 12 October. https://www.cito.nl/onderwijs/voortgezet-onderwijs/centrale-examens-voortgezet-onderwijs/tools-en-informatie-voor-docenten/examenverslagen.

Coetzee-Lachmann, D. 2009. "Assessment of Subject Specific Task Performance of Bilingual Geography Learners: Analysing Aspects of Subject Specific Written Discourse." Unpublished doctoral dissertation. https://repositorium.ub.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-2009030617.

Coffin, C. 2006. "Learning the Language of School History: the Role of Linguistics in Mapping the Writing Demands of the Secondary School Curriculum." *Journal of Curriculum Studies* 38 (4): 413–429.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Coyle, D., P. Hood, and D. Marsh. 2010. *CLIL: Content and Language Integrated Learning*. Cambridge: Cambridge University Press.

Dallinger, S., K. Jonkmann, J. Hollm, and C. Fiege. 2016. "The Effect of Content and Language Integrated Learning on Students' English and History Competences; Killing two Birds with one Stone?" *Learning and Instruction* 41: 23–31.

Dalton-Puffer, C. 2011. "Content and Language Integrated Learning: From Practice to Principles?" *Annual Review of Applied Linguistics* 31: 182–204.

De Oliveira, L. 2011. *Knowing and Writing School History. The Language of Student's Expository Writing and Teachers' Expectations*. Charlotte, NC: Information Age.

Dunlosky, J., and K. Rawson. 2012. "Overconfidence Produces Underachievement: Inaccurate Self-Evaluations Undermine Students' Learning and Retention." *Learning and Instruction* 22 (4): 271–280.

Eurydice/European Commission/EACEA. 2017. *Key Data on Teaching Languages at School in Europe – 2017 Edition*. Eurydice report. Luxembourg: Publications Office of the European Union.

Huibregtse, I. 2001. "Effecten en Didactiek van Tweetalig Voortgezet Onderwijs in Nederland [Effects and teaching methods in bilingual high school education in the Netherlands]." Unpublished doctoral dissertation, University of Utrecht, Utrecht, Netherlands.

Jappinen, A.-K. 2005. "Thinking and Content Learning of Mathematics and Science as Cognitional Development in Content and Language Integrated Learning (CLIL): Teaching Through a Foreign Language in Finland." *Language and Education* 19 (2): 148–169.

Lublin, J. 2003. *Deep, Surface and Strategic Approaches to Learning*. Belfield: Centre for Teaching and Learning, University College Dublin.

Martinot, M., H. Kuhlemeier, and H. Feenstra. 1988. "Het meten van affectieve doelen: de validering en normering van de belevingsschaal voor wiskunde (BSW) [Measuring Affective Goals: the Validation and Standardization of the Math Experience Scale]." *Tijdschrift Voor Onderwijsresearch* 13: 65–76.

Mearns, T., R. De Graaff, and D. Coyle. 2017. "Motivation for or from Bilingual Education? A Comparative Study of Learner Views in the Netherlands." *International Journal of Bilingual Education and Bilingualism*, doi:10.1080/13670050.2017.1405906.

Midgley, C., M. Maehr, L. Hruda, E. Anderman, L. Anderman, and K. Freeman. 2000. *Manual for the Patterns of Adaptive Learning Scales*. Ann Arbor: University of Michigan.

Oattes, H., R. Oostdam, R. De Graaff, R. Fukkink, and A. Wilschut. 2018a. "CLIL in Dutch Bilingual Education: How Dutch History Teachers Focus on Second Language Learning." *Dutch Journal of Applied Linguistics* 7 (2): 156–176.

Oattes, H., R. Oostdam, R. De Graaff, and A. Wilschut. 2018b. "The Challenge of Balancing Content and Language: Perceptions of Dutch Bilingual Education History Teachers." *Teaching and Teacher Education* 70: 165–174.

Otten, R., and M. Boekaerts. 1990. "Schoolvakbeleving bij geschiedenis, Nederlands en wiskunde bij leerlingen in de brugklas [Attitude Towards History, Dutch and Maths in the First Year of Secondary Education]." In *Onderwijsleerprocessen*, edited by M. Boekaerts and E. De Corte, 101–116. Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Perez-Canado, M. 2017. "CLIL and Educational Level: A Longitudinal Study on the Impact of CLIL on Language Outcomes." *Porta Linguarum* 29: 51–70.

Pintrich, P., and E. De Groot. 1990. "Motivational and Self-Regulated Learning Components of Classroom Academic Performance." *Journal of Educational Psychology* 82 (1): 33–40.

San Isidro, X., and D. Lasagabaster. 2018. "The Impact of CLIL on Pluriliteracy Development and Content Learning in a Rural Multilingual Setting: A Longitudinal Study." *Language Teaching Research* 22. Online Publication. doi:10.1177/1362168817754103.

Seegers, G., C. Van Putten, and C. De Brabander. 2002. "Goal Orientation, Perceived Task Outcome and Task Demands in Mathematic Tasks: Effects on Students Attitude in Actual Task Settings." *British Journal of Educational Psychology* 72: 365–384.

Stohler, U. 2006. "The Acquisition of Knowledge in Bilingual Learning: an Empirical Study on the Role of Learning in Content Learning." *Vienna English Working Papers* 15 (3): 41–46.

Sturgis, P., C. Roberts, and P. Smith. 2012. "Middle Alternatives Revisited: How the Neither/nor Response Acts as a Way of Saying 'I Don't Know?'" *Sociological Methods and Research* 43 (1): 15–38.

Sylvén, L., and A. Thompson. 2015. "Language Learning Motivation and CLIL: Is There a Connection?" *Journal of Immersion and Content-Based Language Education* 3 (1): 28–50.

Taber, K. 2016. "The use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education." *Research in Science Education* 48 (6): 1273–1296.

Van der Kaap, A., and A. Visser. 2016. *Geschiedenis. Vakspecifieke Trendanalyse 2016*. Enschede: SLO.

Van Drie, J., M. Braaksma, and C. Van Boxtel. 2015. "Writing in History Effects of Writing Instruction on Historical Reasoning and Text Quality." *Journal of Writing Research* 7 (1): 123–156.

Verspoor, M., C. de Bot, and X. Xu. 2015. "The Effects of English Bilingual Education in the Netherlands." *Journal of Immersion and Content-Based Language Education* 3 (1): 4–27.

Wilschut, A. 2015. "Testing Frame of Reference Knowledge in National Examinations. Report on an Experiment in the Netherlands." In *Joined-Up History: New Directions in History Education Research* (International Review of History Education, Vol. 8), edited by A. Chapman and A. Wilschut, 85–112. Charlotte, NC: Information Age.

Zydatiss, W. 2007. *Deutsch-Englische Zuge in Berlin (DEZIBEL). Eine evaluation des bilingualen Sachfachsunterrichts in Gymnasien: Kontext, Kompetenzen, Konsequenzen* [German-English Tracks in Berlin (DEZIBEL). An Evaluation of Bilingual Subject Teaching in Secondary Schools: Context, Competencies, Consequences]. Frankfurt am Main: Pieter Lang.