Educational Research and Evaluation
Vol. 12, No. 1, February 2006, pp. 75–93

Routledge
Taylor & Francis Group

Evaluation of Bilingual Secondary Education in The Netherlands: Students' language proficiency in English[1]

Wilfried Admiraal[a]*, Gerard Westhoff[a] and Kees de Bot[b]
[a]*Utrecht University, The Netherlands, and* [b]*University of Groningen, The Netherlands*

(Received 9 July 2004; accepted 14 June 2005)

In this longitudinal study, we examined the effects of the use of English as the language of instruction in the first 4 years of secondary education in The Netherlands on the students' language proficiency in English and Dutch, and achievement in subject matters taught through English. Compared to a control group in regular secondary education, students attending this form of bilingual education showed higher scores for their English language proficiency in terms of oral proficiency and reading comprehension. No effects have been found for receptive word knowledge and no negative effects have been found with respect to the results of their school leaving exams at the end of secondary education for Dutch and subject matters taught through English.

Introduction

Although Dutch is a medium-sized language in Europe (some 20,000,000 mother tongue speakers in The Netherlands, Belgium, and France), foreign language skills have been important in Dutch society for a long time. More than 90% of its members live in less than 100 miles distance from a language border. Given the increased possibilities of mobility and communication, the radius of action of most inhabitants is potentially greater than the Dutch speaking region. Utilizing this potential has become much easier since the opening of the European internal border in 1992. Traffic of

*Corresponding author. IVLOS Institute of Education, Utrecht University, P.O. Box 80127, NL-3508 TC Utrecht, The Netherlands. E-mail: w.f.admiraal@ivlos.uu.nl

ISSN 1380-3611 (print)/ISSN 1744-4187 (online)/06/010075–19

DOI: 10.1080/13803610500392160

people and goods is free inside the EU, vocational and educational qualifications are under certain conditions acknowledged in all member states and everybody is free to live, study, work, or trade in other member states with few restrictions. In order to profit from these possibilities, proficiency in at least one European language of wider communication is a prerequisite. But acquiring functional knowledge (e.g., B1 in the Common European Framework; Council of Europe, 2001) in foreign languages using the current teaching methods would ask for a disproportional share in the curriculum time available. In order to solve this dilemma, policy-makers had to look for more efficient ways to foster foreign language skills in the Dutch educational system.

In 1989, the Ministry of Education commissioned a task force to devise effective measures to help solve the foreign language problems of the country. The task force published a National Action Programme on Foreign Languages in 1990, of which an English translation became available in 1992 (Van Els & Van Hest, 1992). This National Action Programme contains a number of recommendations that ''should contribute to the formulation of a coherent government policy for foreign language teaching in the Netherlands'' (Van Els & Van Hest, 1992, p. 2). One of the suggestions in the National Action Programme is to investigate the possibilities of bilingual secondary education.

Bilingual Education and Immersion

The meaning of the term ''bilingual education'' (BE) may be ambiguous. It can refer to a situation in which language minority children are taught in the language of the majority group, a situation often referred to as ''submersion'' and aiming at the development of their skills in that language that may be foreign to them but which is dominating daily life outside school. It can also refer to ''immersion'', in which a foreign language, for example a language that is not the language of the larger society, is the medium of instruction. In the Dutch context, BE has the characteristics of late immersion programmes (cf. Baker, 2002). Discussions at the European level have led to forms of immersion that are specific to the European setting and the label used for this type of education that has emerged is Content and Language Integrated Learning (CLIL; for discussion, see Marsh, Maljers, & Hartiala, 2001).

The predominant type of BE in Dutch secondary education is partial immersion with about 50% of the curriculum being taught in English. The curriculum content in some subject areas is taught through the medium of a second language, whereby the main focus is on that content rather than on the foreign language in which it is communicated. In accordance with current theory on Second Language Acquisition (for an overview, see Brown, 2000; Ellis, 2003; Lightbown & Spada, 1999; Mitchell & Myles, 1998; Richards & Rodgers, 2001), meaningful and authentic communication that is functional and relevant to the students' needs, and not contrived, tightly controlled or repetitive, is important for second language acquisition. Continuous insistence on correct communication is avoided. The assumption is that using the language for content learning will enhance implicit learning of that language leading to higher levels of proficiency than can normally be achieved.

English-Dutch Bilingual Education in The Netherlands

By the time that the National Action Programme was published, plans for bilingual education (BE) already existed in a few schools for secondary education. BE in The Netherlands originates from a few schools for International Education that provide secondary education exclusively accessible for students who are only temporarily in The Netherlands (e.g., children of (foreign) diplomats and employees of international businesses). These schools appeared to attract also children who did not meet these criteria. Then, BE started as a ''grass-roots'' movement in secondary education: A number of highly motivated teachers and parents convinced their schools to start a new line of teaching, in which the foreign language should be used as a medium of instruction in a substantial part of the curriculum. The first school started to use English as the medium of instruction in September 1989; in September 2005 there will be about 70 schools for secondary education using English as the medium of instruction in a large part of the curriculum, with more than 5,000 students.

Within certain restrictions, Dutch schools for secondary education are relatively free to design their bilingual curricula. The main criteria from the Dutch Ministry of Education are that 50% of the total number of lessons may be taught in English, that the Dutch pre-university curriculum is followed, and that it does not affect students' language proficiency in Dutch. It is usually offered in lower secondary education (grades 1 – 4, age group 12 – 15). The subject matters taught through English are mostly History, Geography, and Biology. In addition (to achieve the 50% English taught lessons) schools may offer Drama, Drawing, Handicrafts, Physical Education, Mathematics in English, and, in higher grades, Economics, Physics, and Chemistry. In the 2 years of upper secondary education, Dutch is in most cases medium of instruction throughout the curriculum again. There are several, mostly practical reasons for the restriction of BE to the lower secondary level. One is that at the time of our investigation, after 4 years of common curriculum, students chose a personal set of seven subject matters to specialize in for the final school leaving exams[2]. This requires a substantial regrouping of the students making it difficult to organize BE for groups that are sufficiently large. Moreover, the final school leaving examinations are in the Dutch language. Schools and parents tend to prefer a preparation in Dutch for that reason.

Several evaluation studies (e.g., Baetens Beardsmore & Kohls, 1988; Christian, 1996; DeJong, 2002; Hart, Lapkin, & Turnbull, 2003; Housen, 2002) confirm positive expectations of BE regarding the proficiency in foreign language, while no negative effects for other subjects have been documented. But in these cases BE is implemented under conditions that are different from the Dutch educational context. Therefore, the Dutch Ministry of Education decided to finance a project to evaluate BE in the Dutch setting.

On the basis of various pilot studies, a longitudinal evaluation study has been set up. Although we also looked at the effect of BE on the results achieved in the subject matters taught through English and proficiency in Dutch, the emphasis of this investigation was on the effects of the use of English as the language of instruction on the students' language proficiency in English.

Method

We investigated the achievements of the students in five schools offering English-Dutch BE in the first phase of pre-university education, the highest level of lower secondary education. In the settings investigated, the subjects English, History, and Geography were taught as usual in the schools for 4 years. English was the medium of communication in all communicative functions during the lessons (content, instruction, interpersonal communication, discipline interventions, etc.). Subject matter teachers did not give explicit attention to fostering second language acquisition, such as form-focused corrective feedback (for a full report, see Huibregtse, 2001).

In this study, the students' English language proficiency is at the centre of the evaluation of the outcomes of BE in The Netherlands: Receptive vocabulary knowledge and reading comprehension have been measured as well as oral proficiency. Another set of variables consists of student characteristics that could be related to the level of English language proficiency and language acquisition: gender, entry ability-level (score on the Cito final test of primary education), information about the home language and language contact outside school, and motivation to learn English.

Data

At the time of the start of the project in September 1993, there were six secondary schools in The Netherlands offering BE. Four of these schools have been included in the study. In September 1994, a fifth school has been added. Three of these were schools for International Education; the other two were schools with a mix of a bilingual programme and regular secondary education. Four cohorts with a total of 1,305 students have been studied: 584 students participating in BE and 721 following the regular programme for secondary education. The latter is the control group in this study coming from the two schools offering BE and two other schools (that were sampled randomly). Data have been gathered for 6 school years (see Table 1 for information on the valid data per year).

Measures

Receptive word knowledge. Receptive word knowledge in English was measured by means of the ''English as a Foreign Language Vocabulary Test'' (EFL Vocabulary Test), developed by Meara (1992) for estimation of the receptive English word knowledge of foreign language learners. The EFL Vocabulary Test is a yes/no test that consists of two different kinds of items: real words and pseudowords. Pseudowords are words that fulfil the phonological constraints of the language but do not bear meaning. The items each consist of one word and are presented visually. Students are asked to indicate whether or not they know the meaning of the word and to answer this with ''yes'' or ''no''. Students know that the test contains non-existing words, not how

Table 1. Number of students from the four cohorts with valid data on the measures of language proficiency in English (584 and 721 students in the BE and control group, respectively)

	EFL Vocabulary Test		Reading comprehension		Oral proficiency	
	BE	Control	BE	Control	BE	Control
School year 1						
September	329	124				
June	353	174				
School year 2						
September	271	121				
December	60	0				
June	315	145	383	137	348	235
School year 3						
September	161	127				
December	143	0				
June	281	202				
School year 4						
September	97	68				
December	189	27				
June	108	241				
School year 5						
September	20	25				
December	111	26				
June	32	24				
School year 6						
September	15	23				
December	38	23				

many nor their location in the test. The EFL Vocabulary Test consists of a number of tests on six different levels that are based on word frequency in written texts. The first test level is based on the 1,000 most frequently occurring English words, the second level on the next 1,000 words, and so on. For each frequency range of 1,000 words there are 10 parallel tests each of which consists of 60 items: 40 real words (a random selection from the frequency range concerned) and 20 pseudowords. The pseudowords consist of syllables of words from the frequency range involved that are put together at random. The resulting pseudowords are judged by native speakers on their consistency with the phonological rules of English. Pseudowords that appeared to meet the phonological rules are included in the test, taking into account the number of syllables. On the basis of the number of hits (''yes''-responses to real words) and false alarms (''yes''-responses to pseudowords) a score can be calculated. This score ranges from 0 to 100 and provides an indication of the size of the student's vocabulary in a given frequency range. The calculation of the test scores takes into account the fact that there are two types of correct and incorrect answers, and includes a correction for sophisticated guessing and students' response style (for more information on the

calculation of the test score, see Huibregtse, Admiraal, & Meara, 2002). The homogeneity of the various test versions appeared to be satisfactory (KR-20 > 0.70). For more information on the (construct) validity of this kind of word recognition test, we refer to the work of Meara and Buxton (1987).

The EFL Vocabulary Test was administered twice a year during a period of 5 years. The test was completed in class, generally lasting 20 – 30 min. Each time, students completed two parallel tests from different frequency levels, corresponding to the ability level of the cohort. This means that students started with a test on level 1 in their 1st year of the programme and concluded on level 5 (for cohort 95) or 6 (the three other cohorts). In Table 1, information on valid test scores for receptive word knowledge is summarized.

Reading comprehension. Reading comprehension was measured by means of the national final examinations for English for intermediate general secondary education (MAVO). It consisted of a number of texts in English with in total 50 multiple-choice questions. The questions were partly in English and partly in Dutch and the number of response alternatives ranged from three to five. The final test score equals the proportion of correct answers, with a range of 0 to 100. The reliability of the test for the four cohorts, in terms of homogeneity, ranged from KR-20 = 0.77 (for cohort 92) to KR-20 = 0.90 (for cohort 93).

This test for reading comprehension was administered once, when students reached the end of their 2nd school year. The test was completed in class, which generally last 2 hr. In Table 1, information on valid test scores for reading comprehension is summarized.

Oral proficiency. Oral proficiency in English has been assessed by means of the ''Cito Oral Proficiency Test for English'', which was designed for the final year of intermediate general secondary education. In this test, students have to use language in various real-life situations. They get their cues on audiotape and via a booklet with text and pictures. General oral proficiency (GOP) and pronunciation (P) were assessed separately. These scores were corrected by the observer for the time students needed and transformed into a range of 0 to 100. The test results of 55 students were evaluated by two observers. The correlations between the rates given by these observers appeared to be quite strong for both parts of the test (Pearson's r for GOP = 0.83 and for P = 0.78). The Cito group has conducted a study on the criterion validity of the test (Steerenberg, 1996). The scores on the test were compared to the students' scores on the oral part of the final examinations of intermediate general secondary education. These correlations appeared to be quite strong (Pearson's r for GOP and P is 0.81 and 0.66, respectively).

For each cohort, this test for oral proficiency was completed once, when students reached the end of their 2nd school year. It was administered individually by an observer (native speaker), generally lasting 20 min. Students from the first cohort (1992) did not complete this test. In Table 1, information on valid test scores for oral proficiency is summarized.

Achievements in Dutch, History, and Geography. For the subject matters being taught through English, we restricted ourselves to the results in Geography and History, since these subject matters were conducted in English in the BE schools studied and were chosen by relatively many students as a graduation subject at upper secondary level. We used the school leaving examination test scores, gathered at the end of upper secondary education. As Dutch schools show considerable differences in teaching materials during the lower secondary stage and there are no exams that are comparable over schools, it would be pointless to administer tests on these subjects at the end of the 3rd or 4th year. The school leaving exams after upper secondary level consisted for more than 50% of nationally administered, semistandardized tests (MC-items and open questions with fixed judgement criteria per question), which are constructed by the Dutch National Institute for Test Development (Cito).

Due to the research period of this study, scores on the final exams (in Dutch and in English) were only available for students of the first cohort (1992): 24 students who attended the BE programme (for 4 years) and 33 students from the regular programme for secondary education. From this group of students, only 29 students (16 students attending the BE programme and 13 students from the regular programme) had History and/or Geography in their school leaving exams.

Student characteristics. The student characteristics refer to student background information that might affect students' language proficiency in English. A questionnaire has been used for gathering language background information of students. It was developed for a study on the role of the media in the life of teenagers in Europe and on their contact with English in and outside school (Berns, Hasebrink, & Skinner, 1996). This questionnaire was handed to the students by their teachers in their 1st school year. It includes 10 open questions on the home language of the students (e.g., place of birth, the family mother tongue, and the language which is spoken at home), 6 nonstructured items on language contact (e.g., experiences with the English media and music, and holidays in English-speaking countries), and 3 open questions about the motivation to learn English. Each of these three aspects (home language, language contact, and motivation to learn English) were transformed to a scale from 0 to 10, with higher scores referring to an English language background, frequent contact with the English language, and a high motivation to learn English. This information was available for 326 BE students and 218 students from the control group. Students from the schools with a BE programme differed significantly (with $\alpha = 0.05$) from students in the control group on home language (3.0 vs. 2.4; $t = 4.8$, $df = 536.4$), language contact (7.0 vs. 6.1; $t = 6.1$, $df = 464.2$), and motivation to learn English (8.6 vs. 8.0; $t = 3.2$, $df = 392.4$).

Moreover, information about gender was available for 558 BE students and 691 students from the control group, showing that 50% of the BE students and 53% of the students in the control group were female.

In addition, information on students' general cognitive abilities at the start of the school programme has been gathered with the Cito final test for primary education, a

standardized test used in most schools for primary education in The Netherlands to determine students' entry ability-level for secondary education. These Cito scores ranged from 533 to 550, with a mean of 547.7 in the BE group and 546.3 in the control group ($t = 5.0$, $df = 459.5$, $p \leqslant 0.001$). A score on entry ability-level was available for only 464 students (280 BE students and 184 control students). This low number of students was caused by the fact that some schools for primary education did not use this test. Additionally, some schools in our study did not record these scores for their students.

Method of Data Analysis

The test results of students in the BE programme have been compared with test results of students in the regular school programme (control group). The data gathered in this study have a hierarchical structure: Data were gathered on students' performances on various occasions located in four cohorts in nine schools. The data from the EFL Vocabulary Test, for example, consisted of scores from test administrations of individual students on different occasions (reflecting the number of months students attended the school programme), who were part of one of the four cohorts of one of the nine schools. In order to take this structure into account in answering the research questions, the data have been analysed by means of multilevel regression analyses. Due to the low number of schools in this study, we used the model of Restricted Iterative Generalised Least Squares (RIGLS; see Goldstein, 1999).

Receptive word knowledge. The scores on the EFL Vocabulary Test were equated using a single-group design and an equipercentile method (Kolen & Brennan, 2004), resulting in comparable scores (from 0 to 100) for all six frequency levels and 16 measurements. For this equating procedure 1,215 (level 1 and 2), 633 (level 2 and 3), 1,420 (level 3 and 4), 1,366 (level 4 and 5), and 580 (level 5 and 6) valid student scores were used. The hierarchical structure of the data on receptive word knowledge includes four levels: occasion, student, cohort, and school. Multilevel analyses were performed using a multilevel repeated measures design. The EFL Vocabulary Test scores are the dependent variables, school programme (BE or regular education) and time (in terms of the number of months attending the school programme), the independent variables, and student characteristics the covariates. The answer to the question whether BE has an effect on the (growth of) receptive word knowledge of the students lies in the value and the significance of the parameter that indicates the interaction effect of the time variable with the variable school programme, that is, the difference between the BE group and the control group in increase of the score per occasion. As the results could be different for the six frequency levels of the test, separate analyses have been performed for each level.

Due to limited data available for student characteristics, the analyses with the covariates are conducted separately for students' gender, their entry ability-level, and language background information (home language, language contact, and motivation to learn English), respectively. In these analyses, both the main effect of the

covariate(s) and the interaction effect of the covariate(s) with the number of months students attended the school programme have been added to the model.

Reading comprehension. The hierarchical structure of the data on reading comprehension consists of three levels: student, cohort, and school. The multilevel model used for the analysis is more straightforward than the one for receptive word knowledge, due to the fact that there are no growth curves involved. As the reading comprehension test was administered only once, the score on the EFL Vocabulary Test at the first occasion has been added to the model at the student level. In this way, it was possible to control to some extent for the English language proficiency level at the start of secondary school. Again, separate analyses were carried out for the three (subsets of) covariates.

Oral proficiency. For oral proficiency, there are two dependent variables: one score for general oral proficiency (GOP) and one score for pronunciation (P). In order to investigate differences between the BE group and the control group, a multivariate multilevel analysis was conducted. There are four levels involved. The data of dependent variables (level 1) are nested within students (level 2) who belong to a cohort (level 3) in a school (level 4).

In addition, the initial score on the EFL Vocabulary Test was included in the analysis at the student level in order to control for the student's English language proficiency at the start of the school programme. The analyses with the covariates have been conducted separately for the three (subsets of) covariates.

Results

Receptive Word Knowledge

The growth curves of receptive word knowledge are similar for the six frequency levels of the EFL Vocabulary Test. In Figure 1, we present the growth curve of the student test scores of level 4, that is words with a frequency of 4,000 – 5,000. The growth curves of the test levels 1 to 3 start and end some 5 points higher; this is some 10 points for test levels 5 and 6. Note that the number of months refers to the average period of a school programme per year (10 months), indicating the start of that particular month.

The nonlinear shape of the development curve of the score is noticeable. The data correspond quite well with the logarithmic trend curve added to the graph. The increase of the scores was the largest during the first part of the educational programme, actually during the 1st school year. Throughout the following school years, the students' receptive word knowledge still grew but at a lower rate. So, a logarithmic curve instead of a linear growth curve has been fitted in the analysis of the development of receptive word knowledge.

The graph in Figure 1 shows a difference between the BE group and the control group at all test administrations. The results in Table 2 reveal that the mean score of

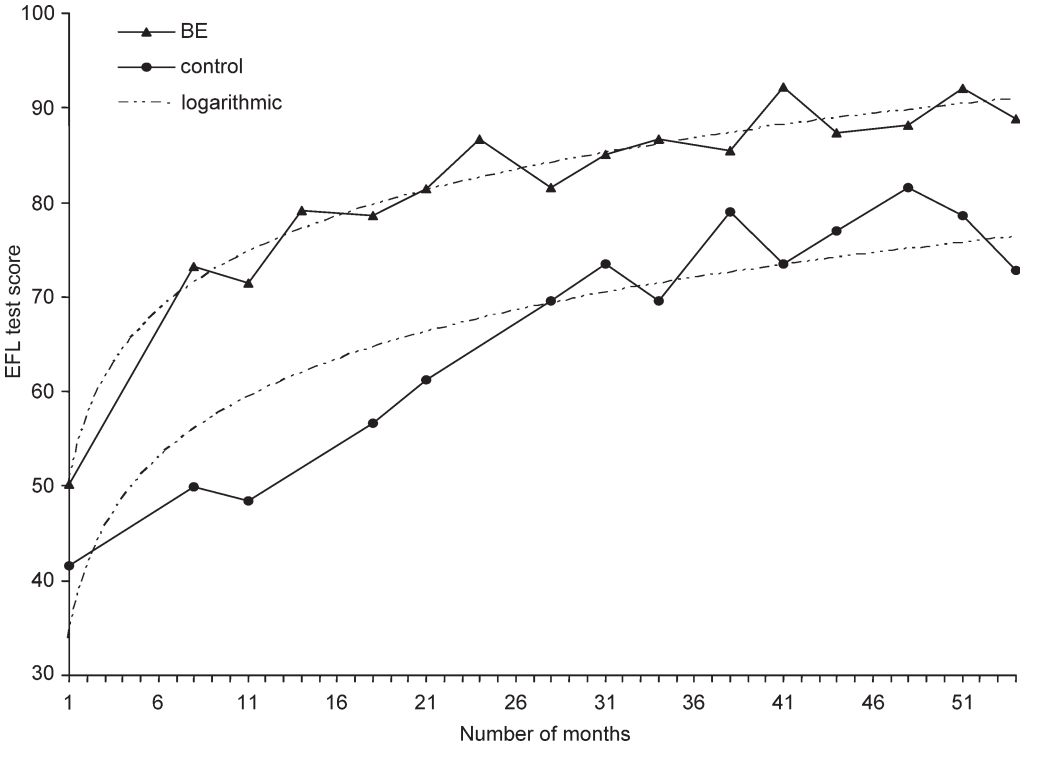

Fig. 1. Mean scores on the fourth frequency level of the EFL Vocabulary Test for the BE group and the control group plotted against the number of months in school, with an added logarithmic trend curve

the BE group at the start of the programme was significantly higher than the mean score of the control group, for all frequency levels. This effect remained significant when the student characteristics (gender, general ability, home language, language contact, and motivation to learn English) were taken into account. Students from the BE programme differed significantly on the student characteristics, which were included as covariates. For example, for test level 2, boys scored significantly higher than girls ($\beta = 5.5$; SE = 1.32), and students with high scores on general ability scored higher than the other students ($\beta = 5.2$; SE = 1.27) as was the case for students with a home language background which is related to the English language ($\beta = 1.9$; SE = 0.55). In general, the main effect of the BE programme was large in terms of proportion explained variance on school level (R^2; for the calculation of R^2, see Snijders & Bosker, 1999): It ranges from 0.48 (level 1) to 0.97 (level 6).

The results in Table 2 also show that there was no significant difference between the BE group and the control group regarding the development of the mean score on the EFL Vocabulary Test. This means that, in general, students in the BE group did not acquire English words at a faster rate than students of the control group. It also means that the BE group maintained the advantage it had at the beginning of the school programme throughout the measurement period.

Table 2. Multilevel repeated measures analyses for the EFL Vocabulary Test

	Frequency level 1		Frequency level 2		Frequency level 3		Frequency level 4		Frequency level 5		Frequency level 6	
Fixed												
Intercept	60.7	(3.52)	31.1	(6.03)	31.0	(5.68)	22.3	(6.37)	7.4	(1.69)	13.1	(3.53)
Time	**7.5**	(1.01)	**11.9**	(1.75)	**12.4**	(1.60)	**13.3**	(1.92)	**15.1**	(1.69)	**8.6**	(1.17)
School programme	**16.1**	(4.85)	**26.1**	(8.41)	**26.9**	(7.85)	**28.1**	(8.86)	**32.5**	(7.40)	**14.1**	(4.67)
Time* school programme	−2.2	(1.39)	−3.3	(2.44)	−3.6	(2.19)	−3.5	(2.67)	−4.1	(2.34)	0.9	(1.56)
Random												
Level 4 (school)												
$\sigma^2 f_0$	24.3	(28.0)	94.2	(83.37)	73.0	(73.5)	104.2	(91.89)	24.5	(11.75)	0	0
$\sigma^2 f_1$	1.0	(2.36)	5.7	(7.17)	2.8	(5.87)	7.3	(8.61)	0	0	0	0
Level 3 (cohort)												
$\sigma^2 v_0$	100.0	(33.69)	257.1	(81.88)	251.3	(81.2)	277.3	(89.5)	352.7	(94.42)	151.1	(43.47)
$\sigma^2 v_1$	11.2	(3.67)	29.1	(9.05)	28.6	(8.99)	33.7	(10.48)	45.7	(11.75)	17.9	(5.08)
Level 2 (student)												
$\sigma^2 u_0$	119.8	(9.60)	153.5	(13.15)	183.3	(15.11)	201.5	(16.53)	164.3	(14.75)	48.8	(3.02)
$\sigma^2 u_1$	7.7	(0.85)	8.1	(1.14)	9.9	(1.28)	9.37	(1.31)	5.1	(1.12)	0	0
Level 1 (occasion)												
$\sigma^2 e$	34.0	(1.01)	56.7	(1.70)	59.4	(1.77)	66.3	(1.97)	75.0	(2.22)	59.9	(1.63)

Note: Standard errors between brackets. Fixed parameters printed bold are significant for $\alpha = 0.008$ (0.05/6). Significance level has been corrected using the Bonferroni-method (Pedhadzur, 1982). * means interaction effect.

It is important to check whether the results of this analysis were affected by the differences between the BE group and the control group at the beginning of the school programme. Potential differences between the groups in the development of the scores might have become invisible because of a negative relationship between the initial test score and the increase of the score. This might be problematic as a high initial score entails less potential for growth.

In order to check whether the differences in the initial scores between the two groups had any influence on the outcomes of the analyses, analyses have been conducted for two subgroups of students: one group with an initial score below (or equal to) the overall median and one group with an initial score higher than the median. This means that for each of these subgroups there was no significant difference in the mean score between the BE group and the control group at the first measurement. However, again we do not find significant effects of the school programme on the development of students' receptive word knowledge: Matching according to the initial scores did not result in students attending the BE programme showing a different growth curve than students from the regular programme, if controlled for relevant variables. The same outcomes have been found for the group of students with an initial score above the overall median.

Reading Comprehension

The results of the multilevel analysis for reading comprehension are summarized in Table 3. It becomes clear that the students in BE group had a significantly higher mean score (77.2) than the students in the control group (55.1). The variance at the school level was greatly reduced by the addition of the variable school programme to the empty model, which means that differences between the schools in students' scores on reading comprehension are largely explained by the fact that they are BE schools or control schools (R^2 on school level = 0.94).

Several student characteristics (general cognitive ability, home language, and motivation to learn English) and the initial score on the EFL Vocabulary Test appeared to have a significant effect on the reading comprehension test score. Still, when these effects were taken into account, the difference between the BE group and the control group in their reading comprehension scores remained significant and quite large. This means that students attending the BE programme had higher scores on reading comprehension in English than students from the regular programme, if controlled for relevant variables.

Oral Proficiency

The results of the multivariate regression analysis of the scores on the oral proficiency (general oral proficiency and pronunciation as measured by the Cito Oral Proficiency test) are summarized in Table 4. These results show that the score of the BE group was significantly higher than the score of the control group for both general oral proficiency (78.0 vs. 54.6) and pronunciation (76.4 vs. 62.3).

Table 3. Multilevel regression analyses for reading comprehension

	Empty model		School programme		Initial score EFL		Gender		Cito score		Language background information	
Fixed												
Overall mean	67.8	(4.08)	55.1	(2.55)	39.4	(5.79)	54.5	(2.64)	56.3	(3.39)	42.8	(4.84)
EFL initial score					**0.3**	(0.05)						
Gender							1.7	(1.06)				
Cito score									**1.5**	(0.30)		
Home language											**1.2**	(0.42)
Language contact											−0.6	(0.39)
Motivation											**0.8**	(0.33)
School programme			**22.1**	(3.13)	**18.1**	(5.69)	**21.7**	(3.15)	**20.4**	(3.97)	**29.1**	(4.32)
Random												
Level 3 (school)												
$\sigma^2 f_0$	136.1	(70.7)	7.6	(9.75)	3.3	(2.61)	7.3	(10.01)	0	0	0	0
Level 2 (cohort)												
$\sigma^2 v_0$	20.4	(10.63)	20.3	(10.29)	41.3	(29.06)	21.0	(10.81)	35.8	(16.8)	25.9	(12.09)
Level 1 (student)												
$\sigma^2 u_0$	138.3	(8.77)	138.3	(8.77)	115.6	(10.12)	137.0	(8.82)	113.8	(12.29)	110.9	(9.31)
Sample size	520		520		272		503		186		298	

Note: Standard errors between brackets. Fixed parameters printed bold are significant for $\alpha = 0.05$.

Table 4. Multivariate multilevel regression analyses for oral proficiency

	Empty model		School programme		Initial score EFL		Gender		Cito score		Language background information	
Fixed												
Overall mean GOP	67.5	(4.47)	54.6	(4.03)	34.4	(6.12)	54.5	(4.04)	54.4	(5.89)	40.8	(7.94)
Overall mean P	70.1	(2.62)	62.3	(2.23)	45.5	(3.86)	61.6	(2.33)	59.2	(3.16)	48.7	(4.78)
EFL initial score GOP					**0.3**	(0.04)						
EFL initial score P					**0.3**	(0.04)						
Gender GOP							0.2	(1.01)				
Gender P							1.3	(0.95)				
Cito-score GOP									0.6	(0.29)		
Cito-score P									0.2	(0.27)		
Home language GOP											**1.8**	(0.41)
Home language P											**1.4**	(0.37)
Language contact GOP											0.7	(0.39)
Language contact P											**0.7**	(0.34)
Motivation GOP											**0.8**	(0.33)
Motivation P											0.4	(0.29)
School programme GOP			**23.4**	(5.38)	**26.0**	(6.85)	**23.5**	(5.34)	**22.4**	(7.14)	**19.8**	(8.29)
School programme P			**14.1**	(2.97)	**14.7**	(4.08)	**14.2**	(3.02)	**15.1**	(3.88)	**14.5**	(4.59)
Random												
Level 4 (school)												
GOP $\sigma^2 f_1$	102.3	(88.75)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
P $\sigma^2 f_2$	37.5	(30.30)	0	(0)	130.4	(49.20)	168.3	(49.39)	159.8	(61.74)	160.7	59.31
Level 3 (cohort)												
GOP $\sigma^2 v_1$	206.4	(75.24)	171.4	(50.16)	90.7	(7.32)	136.7	(8.36)	140.6	(14.7)	116.2	(9.54)
P $\sigma^2 v_2$	60.0	(23.49)	48.3	(15.31)	0	(0)	0	(0)	0	(0)	0	(0)
Level 2 (student)												
GOP $\sigma^2 u_1$	137.1	(8.25)	137.1	(8.25)	42.7	(17.33)	49.9	(15.83)	39.3	(17.84)	46.0	(18.3)
P $\sigma^2 u_2$	121.3	(7.30)	121.3	(7.30)	85.3	(6.87)	121.7	(7.44)	131.0	(13.63)	94.6	(7.77)
Sample size	579		579		323		561		201		314	

Note: Standard errors between brackets. Fixed parameters printed bold are significant for $\alpha = 0.05$.

The school programme appeared to explain the variance in the scores at the school level. At both the cohort level and the student level there is a strong positive correlation between the scores for general oral proficiency and pronunciation (0.92 and 0.54, respectively).

Analyses with the student characteristics as covariates show that there was no significant difference in the mean scores between boys and girls. Moreover, there appeared to be no significant relationship between the standard score on the Cito Final test of primary education and the scores for oral proficiency. The initial score on the EFL Vocabulary Test (representing the English language proficiency at the start of the school programme), home language, language contact (for general oral proficiency), and motivation to learn English (for pronunciation) appeared to explain part of the differences between student scores on oral proficiency. In general, there were small or moderate effects (with a $R^2 \leqslant 0.30$). In all cases, the difference between the BE group and the control group remained significant, showing strong effects for general oral proficiency and moderate effects for pronunciation. In sum, this means that students attending the BE programme had higher scores on oral proficiency in English than students from the regular school programme, if controlled for relevant variables.

Achievements in Dutch, History, and Geography

As to the results in Dutch and in the subject matter areas taught through English (History and Geography), only relatively few data were available. Firstly, because 40% of the schools decided to take part in the international baccalaureate programme and exams, this made their outcomes incomparable to the control group. Secondly, because from the other 60% only a part of the BE population chose History and Geography as an exam subject. The results of the multivariate regression analysis are summarized in Table 5. These results show that the score of the BE group was significantly higher than the score of the control group for achievement on the final exams in English (81.3 vs. 67.6). This was a strong effect. No other significant effects were found. As only a limited dataset was available, no analyses with student characteristics as covariates were performed.

Conclusions and Discussion

With respect to students' receptive vocabulary in English, assessed by means of the EFL Vocabulary Test, the BE group and the control group showed a similar growth curve. The BE group included relatively many students with a rather high initial test score combined with a negative relationship between the value of the initial score and the increase of the score. When the BE group and the control group were (if only roughly) matched by initial score, there still was no difference between the growth curves of receptive vocabulary knowledge of the BE group and the control group.

For reading comprehension and oral proficiency, as measured in this study, it can be concluded that BE led to better results in language proficiency in the target

Table 5. Multivariate multilevel regression analyses for final exams Dutch and English, and History and Geography

	Dutch and English				History and Geography			
	Empty model		School programme		Empty model		School programme	
Fixed								
Overall mean Dutch	74.3	(1.39)	74.7	(2.36)				
Overall mean English	74.4	(4.05)	67.6	(1.49)				
Overall mean History					72.5	(1.30)	71.3	(2.14)
Overall mean Geography					67.3	(1.85)	69.9	(2.18)
School programme Dutch			−1.1	(3.36)				
School programme English			**13.7**	(2.28)				
School programme History							1.9	(2.71)
School programme Geography							−4.6	(2.94)
Random								
Level 3 (school)								
Dutch $\sigma^2 v_1$	3.8	(5.34)	7.3	(7.72)				
English $\sigma^2 v_2$	59.5	(46.42)	0	(0)				
History $\sigma^2 v_1$					49.6	(15.9)	49.9	(14.95)
Geography $\sigma^2 v_2$					3.7	(8.22)	0	(0)
Level 2 (student)								
Dutch $\sigma^2 u_1$	47.8	(9.45)	47.7	(9.42)				
English $\sigma^2 u_2$	74.1	(14.53)	71.6	(13.5)				
History $\sigma^2 u_1$					38.3	(11.52)	39.7	(11.9)
Geography $\sigma^2 u_2$					0	(0)	0	(0)
Sample size	57		57		29		29	

Note: Standard errors between brackets. Fixed parameters printed bold are significant for $\alpha = 0.05$.

language than regular education. The differences that exist between the BE group and the control group in the test scores at the end of the 2nd school year were rather large. Also, when the initial score on the EFL Vocabulary Test and student characteristics (gender, general cognitive ability, home language background, language contact, and motivation to learn English) were taken into account, the BE group appeared to have a significantly higher score for reading comprehension, general oral proficiency, and pronunciation. We did not find any (negative) effects on the subjects that were taught through English (History and Geography) and for Dutch.

These results are based on a limited set of data. This is especially the case for the effects for Dutch and the subjects that were taught in English, which were History and Geography in this study. The results with respect to the students' language proficiency in English are based on a more comprehensive dataset. Yet, these results seem to be somewhat ambiguous. The bilingual school programme had strong positive effects on students' reading comprehension and oral proficiency in English, but did not appear to affect the growth curve in the receptive word knowledge of students. From the perspective of stimulating BE school programmes, this might be a disappointing result. However, the meaning of the development of students' vocabulary could be different on different ability levels: It might be more and more difficult to learn new words from a higher frequency level (see Vermeer, 2001). This would mean that the lack of a significant difference between the student growth curves of both school programmes might be interpreted as a positive effect of the BE programme, although we could not definitely demonstrate this was actually an effect of the bilingual programme.

As far as the generalizability of the conclusions is concerned, it needs to be kept in mind that the schools of this study offering bilingual education were all pioneer schools. The investigation describes the effects of BE "under construction", that is, BE in The Netherlands during the years in which bilingual education started to develop from a small-scale experiment into a full-grown educational programme with close to 70 schools throughout The Netherlands providing this type of education. It is not clear whether this type of programme will have similar outcomes in the schools that started their programmes at a later date. After all, these schools received more support and could make use of the knowledge and experience of the pioneer BE schools. However, the pioneering spirit in these early BE schools may have contributed to the results found. The conclusions of this investigation can therefore not be simply generalized to effects of bilingual education in The Netherlands at this moment.

A related issue on the generalizability refers to the possible specific position of the English language in the Dutch society. As in Scandinavia, students are motivated to learn English as it is an important trading language, and it has a prominent place in Dutch society, educational system, and scientific world. This means that the students' exposure to the English language is already quite high in a context outside school. At the time of writing, there is only one school for secondary education in The Netherlands offering bilingual education in German (and some schools for primary education that offer BE-like programmes in German or French). In order to generalize

the conclusions about bilingual education in English to BE in other languages, more (and different) research has to be set up.

Finally, more insight is needed into the quality of teaching in bilingual education. Although there are some teacher education programmes for bilingual education—using BE themselves—there seems to be room for improvement of the proficiency in English of many Dutch teachers as well as of producing comprehensible input and corrective feedback (cf. Huibregtse, 2001). Class observations and the analysis of BE teaching in terms of, for example, grammatical accuracy, spoken fluency, or corrective feedback, are necessary to explain the effects of BE reported on in this study.

Notes

1. This research was funded by Netherlands Organisation for Scientific Research (NWO; project number 95.020).
2. In the meantime, secondary education in The Netherlands has changed as a result of an educational reform.

References

Baetens Beardsmore, H., & Kohls, J. (1988). Immediate pertinence in the acquisition of multilingual proficiency: The European schools. *Canadian Modern Language Review*, *44*, 240–260.

Baker, C. (2002). *Foundations of bilingual education and bilingualism* (3rd ed.). Clevedon, UK: Multilingual Matters.

Berns, M., Hasebrink, U., & Skinner, E. (1996). *The English language and the media world of young Germans.* Unpublished manuscript.

Brown, H. D. (2000). *Principles of language learning and teaching* (4th ed.). White Plains, NY: Pearson.

Christian, D. (1996). Two-way immersion education: Students learning through two languages. *Modern Language Journal*, *80*, 66–76.

Council of Europe. (2001). *Common European framework of reference for languages, learning, assessment.* Cambridge, UK: Cambridge University Press.

DeJong, E. J. (2002). Effective bilingual education: From theory to academic achievement in a two-way bilingual program. *Bilingual Research Journal*, *26*, 65–84.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford, UK: Oxford University Press.

Goldstein, H. (1999). *Multilevel statistical models* (2nd ed.). London: Arnold.

Hart, D., Lapkin, S., & Turnbull, M. (2003). Grade 6 French immersion students' performance on large-scale reading, writing, and mathematics tests: Building explanations. *Alberta Journal of Educational Research*, *49*, 6–23.

Housen, A. (2002). Processes and outcomes in the European schools model of multilingual education. *Bilingual Research Journal*, *26*, 45–64.

Huibregtse, I. (2001). *Effecten en didactiek van tweetalig voortgezet onderwijs in Nederland* [Effects and the pedagogy of secondary bilingual education in The Netherlands]. Utrecht, The Netherlands: IVLOS Universiteit Utrecht.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*, 227–245.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). St. Paul, MN: Assessment Systems Corporation.

Lightbown, P. M., & Spada, N. (1999). *How languages are learned* (Rev. ed.). Oxford, UK: Oxford University Press.

Marsh, D., Maljers, A., & Hartiala, A. (2001). *Profiling European CLIL classrooms.* Jyväskylä, Finland: Centre for Applied Language Studies.

Meara, P. (1992). *EFL vocabulary test.* Swansea, UK: Centre for Applied Language Studies.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142–154.

Mitchell, R., & Myles, F. (1998). *Second language learning theories.* London/New York: Arnold.

Pedhadzur, E. J. (1982). *Multiple regression in behavioural research* (2nd ed.). New York: Holt, Rinehart and Winston.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge, UK: Cambridge University Press.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* London: Sage.

Steerenberg, L. (1996). *The Cito test of oral proficiency. Validity, reliability and practicality.* Arnhem, The Netherlands: Cito.

Van Els, T., & Van Hest, E. (1992). *The Dutch National Action Programme on Foreign Languages: Recommendations and policy reactions.* Zoetermeer, The Netherlands: Ministry of Education and Science.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217–234.