# Above and Beyond:

## Testing the Winds after 10 Years of EarlyBird Elementary Foreign Language Education

*On Speaking Proficiency*


Ruth de Haan

3080471


July 2013

Supervised by: prof. dr. Rick de Graaff


Universiteit Utrecht - Faculteit Geesteswetenschappen

Abstract

English speaking proficiency differences were examined at Dutch elementary schools with both a regular EIBO program (English in the last two years of elementary school) and an EarlyBird program (early English language program starting in the first year). Keeping in mind that the results are still preliminary, children attending EarlyBird schools attained, on average, higher proficiency levels than those attending EIBO schools. Also, there was a strong correlation between a student's speaking skills and their reading, listening, use of English and spelling skills. There was no interaction between a candidate's speaking proficiency results and CITO/school exit level. Schools with a native speaker attained slightly higher proficiency results, on average, than those without. Based on theories by Bachman (1990) and Bachman & Savignon (1986), the Anglia exams used were found to be reliable and the results valid. Together, these findings not only support EarlyBird's claims of more, better and earlier English, but they will also play an instrumental role in the future of (early) foreign language education in the Netherlands.

*Keywords:* EarlyBird, EIBO, Anglia, CITO, adaptive language testing, oral proficiency interview, native speaker, bilingual, early English

## Table of Contents

# 1. Introduction

## 1.1 Background

“No man is an island, entire of itself,” said John Donne. “Every man is a piece of the continent, part of the main" (1624). Although written in 1624, Donne’s words ring truer now than ever before. In this world of fading borders, interdependent markets and intercultural exchanges, no man is an island, and as a result, no country, culture or, more specifically, language is either. In a world defined by blended borders and inter-lingual communication, monolingualism will no longer suffice. It is not surprising, therefore, that the need for foreign language education is on the increase.

In a country as small as the Netherlands, the need for multilingualism is perhaps even more crucial than for its European counterparts. Although the Netherlands was once a major colonial contester in the world arena, Dutch simply never attained the status of lingua franca. In order to continue to compete economically and politically at an international level, the Dutch have set forth a tradition of foreign language learning. As early as 1863, Dutch law already mandated that all those attending secondary education must study French, German and English (www.talenexpo.nl). Now, 150 years later, the Netherlands numbers almost 1,000 elementary schools that are specialized in bilingual learning, all teaching in English, French, German or Frisian. Although these schools cannot be considered bilingual in the secondary education sense of the word, where at least 50% of the classes are taught in English, they are taking definite steps towards such a form of education. For now, it suffices to say that early bilingual education is used to describe elementary schools that teach a foreign language for a minimum of 60 minutes per week. Bilingual education, therefore, is no longer the exception.

With such an increase in numbers, however, no centralized standard could be guaranteed. The quality of bilingual education can drastically differ per school, as can the focus and level of

expertise.

**1.2 Early Bird**

In 2003, in response to the need for standardization and further development of bilingual efforts at elementary schools, EarlyBird (EB) was born. In close collaboration with the European Platform, an organization aimed at endorsing and facilitating internationalization in education, and with BOOR, a school network in Rotterdam, EarlyBird became an active player in the field of English language education. The program's primary focus was the development of English communicative competence at an early age (EarlyBird, *Handboek*, 2010). Although bilingual education is widely accepted at Dutch secondary schools, it does not yet enjoy a similar status in elementary education. EarlyBird, therefore, was an exemplary front-runner in this respect.

Now, ten years later, EarlyBird has developed into a well-established English language program. Although originally only affiliated with Rotterdam and its schools, it has expanded into a center of expertise on early bilingual education throughout the Netherlands. With 38 schools in the Rotterdam area (of which 34 are part of the BOOR network) and 217 throughout the rest of the country (K. Philipsen, personal written communication, June 5, 2013), EarlyBird's unequivocal commitment to English bilingual education has certainly paid off. After ten years of service, however, it is high time that the EarlyBird program is subjected to a thorough investigation. Not only did EarlyBird want to determine the level of English its students reached by the end of elementary school and compare it to the CEFR levels (Common European Framework of Reference for Languages), it also wanted to set down qualitative standards to be met by every EarlyBird school. Furthermore, it was interested in measuring the development in level an EarlyBird child undergoes between different stages of the program, in addition to the personal traits that may have affected their language development. In short, the qualitative standards of EarlyBird and the eventual

outcome in terms of English proficiency need to be standardized and determined in greater detail.

### 1.3 EarlyBird vs. EIBO

For the present study, a distinction must be made between EarlyBird schools and EIBO schools. As mentioned before, English became a mandatory subject in 1986 for all group 7 and 8 students. Schools who strictly follow this protocol are called *Engels in het Basisonderwijs* (EIBO) schools. These schools teach English in the last two years of elementary education (ages 11-12). As of recently, there is also a strand of EIBO (early EIBO) where students start learning English from group 5 onwards (ages 9-10). There are other schools, however, that have chosen to start teaching English from the very first day a student starts going to school (ages 4-6). These schools offer so-called *vervroegd vreemde talen onderwijs* (vvto), or early foreign language education. One program that is specialized in such early foreign language education is EarlyBird. Both EIBO and EarlyBird, however, must meet the Ministry of Education's requirements concerning the four language goals for English (see section 2.2.1). Furthermore, Dutch must remain the language of education for both types of programs. A recent development in this respect is the 15% vvto project, which is aimed at establishing English as the language of elementary education for a maximum of 15% per week.

### 1.4 Research Thus Far

Several facets of English language acquisition have already been studied in greater detail since the introduction of English as a mandatory subject at elementary schools in 1986. In 2005, for example, Goorhuis & de Bot asserted that early bilingual education did not cause a negative effect on the development of a child's L1. Tests related to reading and spelling skills have also been conducted (van Berkel, Philipsen & Feuerstake, 2013). Furthermore, assessments of the EarlyBird program in general were tested recently using EarlyBird's own

*Kijkwizer*, which is an assessment tool used to determine whether a school meets all the criteria for becoming an EarlyBird school.

Another project is the Foreign Languages in Primary school Project (FLiPP). This project concerned itself with a wide array of topics, addressing issues such as the general development of language skills in students and the potential side effects of bilingual education (Unsworth, de Bot, Persson, & Prins, 2012). The results of FLiPP, in turn, played an instrumental role in the 15% vvto project (van Loon & Setz, 2012). In this project, it was piloted that bilingual elementary education could have up to 15% of their classes in English. It was not until 2012 that a test run was conducted for the assessment of English speaking proficiency. A slightly revised version of this test is what will be used for the present study.

**1.5 Research Focus**

For this thorough EarlyBird program check-up, so to speak, EIBO and EarlyBird students were subjected to a series of Listening, Reading, Use of English, Spelling and Speaking proficiency tests. As a mere component of a larger research project, this paper will concentrate on a sub-domain thereof, namely that of the English Speaking proficiency test. More specifically, although these speaking tests were performed on both group 5 and group 8 students at EarlyBird schools, and on group 8 students at EIBO schools, this paper will focus primarily on the levels attained by group 8 students. The results for group 5 will be discussed, but they will not be the main subject of comparison. This choice was made because of a primary need to compare final EarlyBird levels with final EIBO levels. By comparing final levels of English speaking proficiency, both between schools and between the groups themselves, this paper hopes to contribute to the larger discussion at hand concerning early foreign language development. Not only will this paper test the measure of success of the EarlyBird program, but it will also help establish attainable end terms for English foreign language education at large.

**1.6 Procedure**

In order to do so, written and speaking tests were developed to measure the attained levels of English in an EB program by the end of elementary school. These tests were designed in collaboration with Anglia Examinations England, an English testing institution. The tests were conducted at ten EarlyBird schools and nine EIBO schools. Both EB and EIBO schools were tested in order to obtain a representative and realistic comparison. At EB schools, both group 5 and group 8 were tested, whereas at EIBO schools only group 8 was tested. Group 5 and group 8 students took a Listening, Reading, Use of English and Speaking test. Group 8 received an additional spelling test (Dictation). The Listening test was 45 minutes long for group 8 and 30 minutes long for group 5, and consisted of a series of questions for which a student needed to listen to and understand the spoken English on a CD. The Reading test was also 45 minutes long for group 8 and 30 minutes long for group 5, and consisted of a series texts that needed to be read and understood in order to answer the multiple choice questions. The Use of English test was 30 minutes long for group 8 and 30 minutes long for group 5, and was comprised of questions related to sentence structure, meaning and spelling. The Dictation was only conducted on group 8. This test took 20 minutes and consisted of 20 words that were assessed as either completely wrong or completely right.

**1.7 Relevance**

Measuring English speaking proficiency is a highly relevant component when assessing a student's level of English. During a child's first years in an EarlyBird program, the focus is primarily on developing English listening, understanding and speaking skills (EarlyBird, 2010, *Handboek*). According to the EarlyBird *Curriculum* (2010), "this reflects how children learn to communicate in their first language and is in harmony with the teaching philosophy of the education system in the Netherlands" (p. 2). In other words, it would be

unnatural to start teaching English literacy skills in the early EarlyBird years, since students will not even be learning these skills yet in their mother tongue. EarlyBird students will therefore have a head start in terms of speaking skills when compared to EIBO students, who do not start learning English until group 7. The level of a student's English speaking proficiency should therefore be a representative measure of the EarlyBird program's level of success, since the EarlyBird program focuses largely on the development of communicative competence (EarlyBird, 2010, *Handboek*). Furthermore, as will be explained in further detail in section 2.1.3, the speaking tests used are highly unique, since they can test language well below an A1 level.

# 2. Theoretical Framework

## 2.1 The Role of Language Tests

When measuring speaking proficiency, not the only the results must be taken into consideration, but also the method with which the proficiency was tested. An important element to consider is the effect of the testing format itself. One of the reasons schools and programs are often hesitant to incorporate speaking proficiency tests is because such tests involve skills that are difficult to define and measure. Moreover, tests of this nature can be highly subjective, thereby making it a challenge to guarantee a measure of reliability. As Front et al. (2012) point out, "While it has been widely recognized that stimulus materials impact test performance, our understanding of the way in which test takers make use of these materials in their responses, particularly in the context of listening-speaking tasks, remains predominantly intuitive" (p. 345). In other words, answers may be interpreted differently, and teachers may differ in leniency or strictness when it comes to scoring rates. Results could even differ when tested by a native speaker instead of a non-native speaker (Van den Doel, 2006).

### 2.1.1. Language testing in general.

In their critique on the American Council on the Teaching of Foreign Languages (ACTFL) *Proficiency Guidelines*, a breakthrough set of guidelines for foreign language testing and teaching in the eighties, Bachman and Savignon (1986) advocate the development of test guidelines that "provide a standard for defining and measuring language proficiency that would be independent of specific languages, contexts, and domains of discourse" (380). A system should be devised, they argue, in which language tests all over the world define language ability in a similar manner. Moreover, the results of such a test should be comparable across different languages and tests. Although an admirable aim, both Bachman and Savignon foresaw the difficulties involved in such an endeavor. Only a few years later,

however, the Central European Framework of Reference for Languages (CEFR) came into being, thereby addressing part of the problem. Speaking test results could now be compared across languages and countries, but standardizing the test itself still remained a challenge.

The problem with language testing, asserts Bachman (1990), is that "language is both the instrument and the object of measurement" (p. 2). In other words, that what is being tested, is also influenced by the language that is used during the test. Furthermore, Frost et al (2011) point out that since language testing generally focuses on speaking, listening, reading and writing as four separate language skills rather than as interdependent abilities, there is a risk involved in interpreting proficiency levels based on only one of these scores. "Real world communicative acts," say Frost et al, "rely on the integration of two or more of these skills, as well as other non-linguistic cognitive abilities" (p. 346).

Communicative language ability does not involve a simple transfer of information. Rather, it involves two language users (test-taker and the rater), a situation, and the discourse (Bachman, 1990). A test score is always the result of the interaction between different factors. Therefore not only the test format but also the concept of communicative language ability must be defined in such a way as to elicit language use that is similar to that of natural, non-test situations (p. 9). Shohamy (1999) comments on this in her reflection on the development of language testing in the 1980s, explaining that the rationale behind this need to define language ability is that "a clear identification of the construct and structure of language ability will make it possible to design tests to match such descriptions" (p. 156-157). A key measurement issue, argues Bachman (1980), "is determining the extent to which the sample of language use we obtain from a test adequately characterizes the overall potential language use of the individual" (p. 11). A closer look must be taken at test tasks, he says, in order to determine whether they generate language use that will be indicative of a test-taker's level of proficiency. Naturally, the testing method, including such elements as

time and speed, are instrumental. Furthermore, factors such as the (un)familiarity of a testing environment and the explicitness of testing instructions play a pivotal role as well. “Tasks influence the language used,” says Bygate (2009), “and so to appraise students’ language we first need to understand the linguistic demands of our task,” including its phonological, lexico-grammatical and discourse features (p. 414-415). In addition to such technical dilemmas, there are also personal factors that may have influenced the results (Figure 1). Although attempts can be made at controlling the test conditions, little can be done about, for instance, the test-taker's mental state of mind, physical ailments, fatigue or motivation (Bachman 1990).

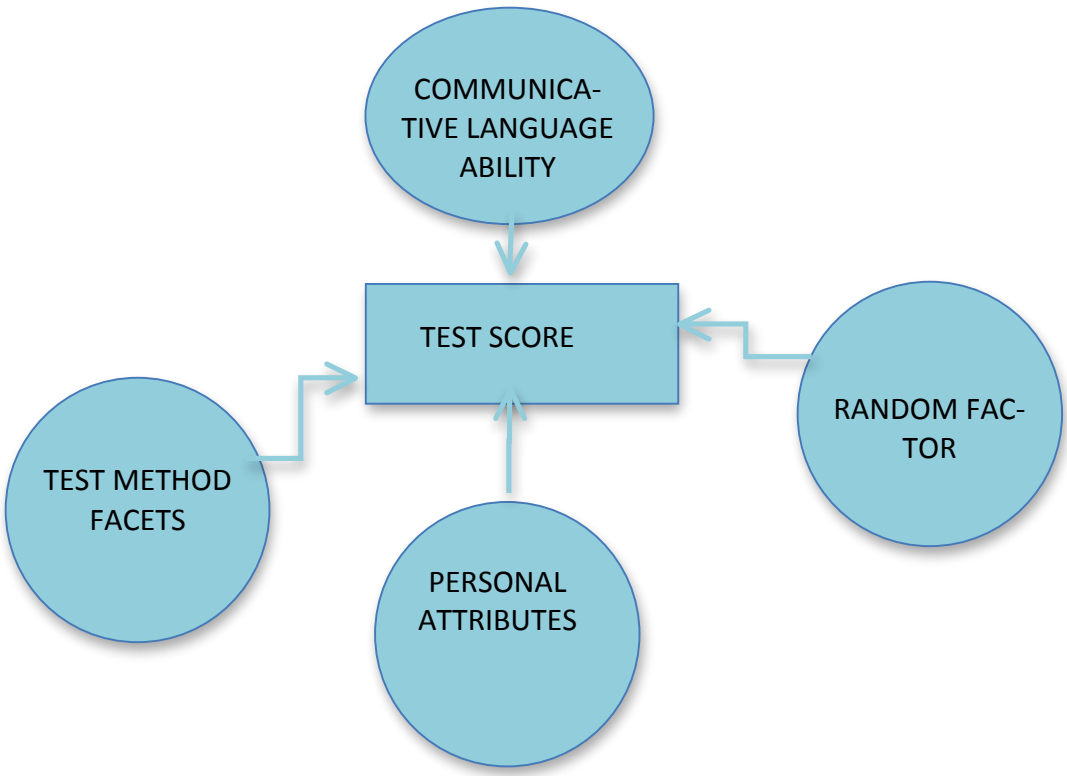


***Figure 1.*** **Factors Affecting Language Test Scores, taken from Bachman (1990)**

Bachman's theory on the matter, therefore, is that test developers must control the testing elements that *can* be controlled. By doing so, claims of reliability and validity can be met, in which reliability promises a measure of consistency rather than chance, and validity determines whether test scores are *meaningful, appropriate,* and *useful* (Bachman, 1990).

Simply put, a test score will be meaningful if it is solely a measure of the ability being tested. A test is therefore not meaningful if the possibility exists that its results have been influenced by other abilities that are not being tested for. A test score can be considered appropriate if the test is being used for the appropriate purpose and target group. Similarly, a test is useful if its results provide information that is relevant to what is being tested (Bachman, 1990). However, Bachman adds, although reliability and validity are admirable aims, they will primarily serve as mere guidelines, rather than as attainable beacons of perfection.

In terms of test development, Bachman (1990) provides the following stepping-stones to most effectively control potential side effects, and to facilitate reliability and validity:

(1) "Provide clear and unambiguous theoretical definitions of the abilities we want to measure;

(2) Specify precisely the conditions, or operations that we will follow in eliciting and observing performance, and

(3) Quantify our observations so as to assure that our measurement scales have the properties we requires" (p. 50).

### 2.1.2. Oral proficiency tests.

According to Bygate (2009), communicative language teaching and testing has undergone a development over the past decades, involving a transition from speech as a mere medium to speech as a target skill. A natural consequence of this development is that speaking tests have become a phenomenon of their own. Although much of the language testing criteria mentioned above applies to speaking tests as well, it is worth zooming in on the particulars of such a test.

As with any test, there are both benefits and setbacks involved when testing oral language proficiency. When developing such a test, both sides must be taken into consideration. In terms of benefits, a speaking test has a definite advantage over written tests

with respect to its flexibility. Whereas a written test is often standardized in terms of contents, an oral examiner can "elicit topics, illocutionary acts, and sociolinguistic registers appropriate to the context and to the candidate's needs and interests" (Bachman & Savignon, 1986, p. 388). The result of such freedom of content, continues Bachman, is that "the interviewer is not constrained to elicit a particular set of discrete grammatical structures, vocabulary items, or speech acts" (p. 388). An interviewer can therefore play around with the types of questions asked and yet still elicit the necessary data. A setback to this flexibility of content, however, is that people appear to communicate better about topics they are unfamiliar with than about topics they master. The reason, discovered Lantolf and Khanji, is that the unfamiliar task requires a test-taker to demonstrate their linguistic ability, whereas a familiar task simply involves their participation in a comfortable, communicative dialog (as cited in Lantolf & Frawley, 1985). The danger, therefore, is that an interviewer's assessment may be based on a test-taker's familiarity with the exam content, rather than on their general ability to communicate.

According to Lang (personal communication, June 4, 2013), the dynamics appear to be similar for children. In the Speaking test's third task, for example, students were asked to find one word that did not fit in with the other three. The students enjoyed this task to such an extent that they "wanted to show the assessor they knew [the answer] even if that meant making mistakes and wandering into difficult linguistic territories". Lang adds that "their confidence in finding an answer encouraged them to speak and attempt explanations even if they were not capable of doing so". For children, too, it is a case of familiarity with the exam content (since they are eager to show that they understand the type of question that is being asked) that causes them to attempt linguistic structures they do not yet master.

According to Iwashita et al (2008), a drawback in oral proficiency testing is that factors determining proficiency can also differ per proficiency level. For the lower

proficiency levels, for example, vocabulary and pronunciation were deemed most important by a panel of teachers, and fluency and grammar were least important. For the higher levels, however, vocabulary, pronunciation, fluency and grammar were equally influential in determining a student's proficiency level. De Jong and van Ginkel found that pronunciation was the determining factor at the lowest levels, and that fluency became increasingly important in the higher levels (as cited in Iwashita et al., 2008, p. 26). Naturally, an objective assessment can only be guaranteed if all interviewers agree on the same required skills per level.

Another benefit of a speaking proficiency exam is that it engages its participants in active, two-sided communication. Since much depends on a language learner's ability to interact in the foreign language, one could say that a test of speaking proficiency best represents real-life language communication skills. The drawback, however, is that such a two-sided proficiency test literally involves two sides, namely the test-taker and the examiner. A test-taker's success will therefore depend on the performance of both individuals (Lantolf & Frawley, 1985). Simply because an interlocutor is present, test results will be affected by the *reciprocity* condition and the *time-pressure* condition (Bygate, 2009). The reciprocity condition, says Bygate, may lead a speaker to adjust their talk in order to meet the examiner's interests and expectations. Similarly, the time-pressure condition refers to the immediacy with which a task needs to be answered or completed. Both conditions may therefore skew the results, simply because an interviewer is present.

In general defense of oral interviews, Bachman & Savignon (1986) argue that the results of such a test can simply be interpreted in two different ways: either 1) the test scores can be seen as a result of the test method and therefore be considered less reliable and valid, or 2) the test scores can be considered "highly reliable indicators of individuals' skill in using grammatical structures accurately *in the contexts and under the conditions that are included*

*in the testing procedure*" (p. 384). The second option implies that the very nature of a proficiency interview is perhaps a true test of a test-taker's skills, since it measures proficiency and grammatical accuracy under the most trying conditions. Nevertheless, there remains a clear need, say Frost et al (2011), to further determine the measure in which a diverse array of stimuli may affect the assessed level of proficiency.

Although originally written as a response to the ACTFL *Proficiency Guidelines*, the following pieces of advice by Bachman & Savignon (1986) are highly relevant for all speaking test development:

(1) "We must clearly define and distinguish the ability and method factors that we expect to influence ratings.

(2) We need to design and conduct empirical studies to estimate the effect of test method factors.

(3) In interpreting oral interview ratings, we must be clear about whether we want to make inferences about a general domain of CLP [Communicative Language Proficiency] or whether we want to interpret ratings in a more limited (but perhaps more meaningful) way as indicators of the ability to use language under specific conditions" (p. 384-385).

**2.1.3. EarlyBird tests.**

For the present study, EarlyBird required oral proficiency tests that met the following criteria. First, the test tasks themselves needed to be *adaptive*, meaning the test could be adapted to the different ability levels of the students. Second, the speaking tests required assessments that could be performed both at low and high proficiency levels. Third, the tasks and contents of the group 5 and group 8 tests needed to be similar, so that potential progress and development between these EarlyBird groups could be measured.

After a pilot run with a speaking test from March to June, 2012 (Lang, personal communication, June 4, 2013), EarlyBird selected Anglia Examinations England as the most suitable test developer. Not only was Anglia familiar with Dutch education in general (Philipsen, personal communication, June 5, 2013), but they were also specialized in testing so-called *Young Learners* (www.anglia.org/student-parents/young-learner). Contrary to many other proficiency interviews, Anglia is capable of testing at levels below that of A1, rather than only from A1 to C2. More specifically, their pre-A1 level spans a range of three distinct Anglia levels, namely *First Step*, *Junior* and *Primary.* All three of these levels are below the CEFR starting level of A1. This means that Anglia was also able to create tests for EarlyBird's Group 5 students, since they required testing levels between *First Step* (Pre-A1) and *Preliminary* (A1). With two speaking tests that were similar in contents but different in level, EarlyBird could compare the results of both tests in order to assess the potential growth and development in speaking proficiency between EarlyBird group 5 and EarlyBird group 8.

What is unique to these versions of Anglia's speaking tests, is that they were developed as blended tests. This means that, contrary to Anglia's tradition of testing only one specific level on a student, these tests consisted of a blend of different levels per task. "The test needed to be blended," says Denise Lang, the Research & Development Team Manager at Anglia Examinations, "to better react to the level of English proficiency used by that particular student" (personal communication, June 4, 2013). This way, an assessor could effectively test a range of levels in order to determine the level a student was most comfortable at. "Many of these questions," continues Lang, " are designed to elicit specific grammatical responses appropriate to that level . . . but they also needed to seamlessly transition up or down to other levels as needed" (D. Lang, personal communication, June 4, 2013).

From lowest to highest, Anglia speaking tests span a total of ten levels, ranging from

the aforementioned *First Step* (pre-A1) to *Masters* (C2) (see Appendix A for a description of each level). For the present study, the group 5 speaking test questions were limited to a *First Step* (pre-A1) to *Preliminary* (A1) range, whereas the group 8 blended speaking test questions ranged from *Primary* (pre-A1) to *Intermediate* (B1). In Anglia speaking tests, two students are interviewed per test, but both do receive an individual Anglia level assessment. For further details on the exact testing and assessment procedure, see section 4.3. In terms of its setup, Anglia speaking exams always consist of three sections, namely an introductory task and two other exercises (D. Lang, personal communication, June 4, 2013). In this case, the two exercises involved a student's observations concerning a picture, and second, a word game (see section 4.2 for more detail). Most of the questions used for the first task have been tried, tested and used before by Anglia Examinations England. The same applies to the questions of the second task. The word game in the third task was used in Anglia exams before as well, although new levels were added for EarlyBird (Lang, personal communication, June 4, 2013).

In addition to the general background and set up of the Anglia exam, it is also important to consider the factors defining its measure of reliability and validity (Bachman, 1990). In their critique on the ACTFL oral exam, Bachman & Savignon (1986) stipulate a number of guidelines regarding the reliability and validity of speaking tests. One is to define the test score rating scales in terms of the absence or presence of certain elements (and the level of control and range thereof), rather than simply base it on a specific individual or actual "performance norms" (p. 388). They also advise test developers to provide interviewers with a list of topics and contexts that will be suitable for eliciting speech fit for rating. In his chapter on the assessment of oral proficiency interview transcripts, Bygate (2009) argues that a "robust appraisal of a particular performance also depends on the ability to identify the presence of features likely to correlate with a given level of proficiency" (p.

414). In other words, an interviewer must also be made aware of the type of features that are inherent to each level as proficiency develops.

For the purpose of this study, however, a study primarily concerned with the post-development aspects of the conducted speaking tests, it will suffice to simply determine whether the suggested criteria by Bachman (1990) and Bachman&Savignon (1986). The results of this criteria check can be found in section 5.6.

## 2.2 The Role of Speaking Proficiency in the EB and EIBO programs

In 1986, English became a mandatory subject in Dutch elementary education (Thijs, Trimbos, Tuin, Bodde, & de Graaff, 2011). As a result, all elementary schools were expected to offer English as a foreign language in the last two years of elementary school, namely in group 7 and group 8. However, a distinction must be made between schools that adhered to this plan (EIBO schools), and schools who started teaching English from group 1 onwards, using programs such as that of EarlyBird.

In order to properly analyze and accurately interpret the speaking exam results of both the EarlyBird and EIBO program, it will be important to discuss the manner in which these programs address the development of speaking proficiency skills. The role a program attributes to English speaking proficiency will be an indication of its value and importance, and therefore indirectly also an indication of the time and effort a program spends on developing a child's speaking proficiency.

### 2.2.1. EIBO.

In 2006, the Dutch Ministry of Education (2006) formulated the third edition of the so-called *Kerndoelen* (core goals) of English in elementary education. These goals set a minimum standard for English language education at elementary schools in the Netherlands

and pertain to all types of English foreign language learning at elementary schools. These goals read as follows:

(1) Students learn how to gather information from simple spoken and written English texts.

(2) Students learn how to ask for or give information on simple subjects and they learn to feel at ease when expressing themselves in English.

(3) Students learn how to spell a few simple words about everyday subjects.

(4) Students learn how to look up the definition and spelling of English words by using the dictionary (Ministerie, 2006).

In terms of speaking proficiency development, only the first two goals are relevant. Although these goals set out admirable guidelines for English language education in general, little is specified in terms of how these goals can be achieved or what levels need to be attained. As of 2006, CITO (a central testing agency for school exit levels) has a CITO *Me2!* exam, which is designed for students who learn English in group 7 and 8. What is interesting to note, however, is that although two goals concern the development of speaking proficiency, no element in de CITO *Me2!* exam actually tests speaking proficiency. Instead, the exam is comprised of a listening, reading, and vocabulary section (CITO, 2006). The third vocabulary section holds much potential for interactive oral communication, and yet it resorts to mere multiple chance questions. As Heesters et al (2008) point out, EIBO speaking proficiency development consists primarily of the repetition of new words out loud, and of conducting standard conversations. Little is done in terms of singing, playing games, telling stories or chatting in English.

In terms of CEFR levels, the CITO *Me2!* exam tests at an A1 and A2 level, with an emphasis on listening and reading skills. At an A1 writing level, a child will be able to write a card to someone. At an A1 speaking level, a child will be able to describe the composition of

their family in terms of number of siblings, ages, etc. At an A2 speaking level, a child will also be able to discuss and describe their hobbies (EarlyBird, 2010, *Handboek*).

### 2.2.2. EarlyBird.

The EarlyBird focus is on communicative competence, namely the development of listening skills and speaking skills (EarlyBird). More specifically, in addition to the aforementioned *Kerndoelen*, the program's aims are as following:

(1) "the development of fluency in English,

(2) the enhancement of general language skills, and

(3) the use of English to learn more about the world we live in" (EarlyBird, 2010, *Curriculum*, p. 2).

Early English schools place a stronger emphasis on the development of speaking, listening and vocabulary skills (Thijs, Trimbos, Tuin, Bodde, & de Graaff, 2011). In line with this tradition, EarlyBird also concentrates its efforts on listening, understanding and speaking skills, and only starts to develop literacy skills after group 4. As can be seen in Table 1, speaking and listening are the primary components of the EarlyBird program. This focus on communication skills, as opposed to grammar and formal language learning, continues throughout all of elementary school (EarlyBird, 2010, *Handboek*). It is important to note, however, that since the EarlyBird program has two strands (early immersion: start in group 1&2 or middle-immersion: start in group 5&6), not all students will have enjoyed the same amount of exposure. With an early immersion, students will have had four to five extra years of listening and speaking.

Table 1. *EB language skills per age group*

| | Age 4-6 (group 1/2) | Age 7-8 (group 3/4) | Age 9-10 (group 5/6) | Age 11-12 (group 7/8) |
|---|---|---|---|---|
| **Listening & Understanding** | x | x | x | x |
| **Speaking** | x | x | x | x |
| **Reading & Understanding** | | x? | x | x |
| **Writing** | | | x? | x |

*Note.* From *Handboek EarlyBird* (EarlyBird, 2010)

Throughout the EarlyBird program, therefore, the focus is mostly on communication skills, rather than formal language skills such as reading and writing. With such a concentration on speaking and listening, it is only natural to assume that EarlyBird children receive much meaningful input (Krashen, 1985). Furthermore, the EarlyBird program also encourages the use of games and songs in order to elicit as much English communication as possible (EarlyBird, 2010, *Handboek* & *Curriculum*). In contrast to EIBO, where more formal learning and frontal teaching occurs (Herder & de Bot, 2008, p. 28), "structures are not presented as separate rules to be learnt but are embedded in songs, chants, rhymes and stories" (EarlyBird, 2010, *Handboek*, p. 12). Much of EarlyBird's proficiency development also occurs in an informal setting. The added benefit of the EarlyBird program is that English input is not only generated in the English classes themselves, but also in different activities throughout the day. EarlyBird teachers are provided with a plethora of ideas, games and activities aimed at encouraging children to speak, removing potential speaking hurdles, training vocabulary, stimulating drama and interaction with puppets, and differentiating per child's proficiency level (EarlyBird). Part of this is done using the Total Physical Response (TPR) technique, in which learners are encouraged to communicate by means of drama, songs and dance, to name a few (EarlyBird, 2010, *Curriculum*).

Another factor to take into consideration is that the EarlyBird program also offers intensive vocabulary training. As stipulated by the EarlyBird *Curriculum* (2010), "the programme is based on the principle of introducing high-frequency English words, *in a meaningful context for children*" (p. 3). The theory underlying this approach is that learning from meaningful input predominantly occurs when the language learner is familiar with at least 95% of the words (Herder & de Bot, 2008). Since EarlyBird children will have heard the first two thousand (highly frequent) English words by the end of group 8 (EarlyBird, 2010, *Curriculum*), it could be suggested that this, in turn, will also contribute to the development of increased speaking proficiency.

In terms of CEFR levels, EarlyBird has not formalized any end terms as of yet. They have, however, made certain predictions that have yet to be tested. As seen in Table 2, EarlyBird predicts that group 8 students will attain CEFR levels in English that are the same as those attained by the end of the second year of secondary education, which is two years after group 8. As of yet, these levels of attainment are still in their testing phase and must therefore, for the time being, be seen as mere guidelines rather than as definite end terms. What is interesting to note is that the goals for oral skills (Listening & Understanding) are not higher than for Interaction or Reading & Understanding skills, and only possibly higher than Writing. This is striking, since Table 1 shows that EarlyBird students receive oral training for a much longer period of time than any other skill. Nevertheless, the results of the present study will still help establish whether the relevant levels for speaking (Listening & Understanding) will be attainable per school exit level by the end of group 8.

Table 2. *EB's Level Aims per School Exit Level*

| | VMBO | HAVO | VWO |
|---|---|---|---|
| **Listening & Understanding** | A1 | A2 | A2/B1 |
| **Interaction** | A1 | A2 | A2/B1 |
| **Reading & Understanding** | A1 | A2 | A2/B1 |
| **Writing** | A1 | A1/A2 | A2 |

*Note*. From Platform, 2011

### 2.2.3. Role of native speaker.

A crucial factor to consider when assessing the role of speaking proficiency in a program, is the presence (or lack thereof) of a native English speaker. Regardless of whether or not a school's program encourages and facilitates an excess amount of oral communication, the measure of success will still depend on the English teacher's level of expertise. According to Arva & Medgyes (2008), there are four areas in which native speaker teachers (NS) differ from non-native speakers (NNS) (as cited in Herder & de Bot, 2008). As Table 3 demonstrates, NS's and NNS's differ in terms of their language use, general attitude, attitude towards the target language, and attitude towards teaching.

Table 3. *Differences between NS & NNS*

| | Domain | NS | NNS |
|---|---|---|---|
| **1** | Own language use | Expert use | Non-expert use |
| **2** | General attitude | Pro: more innovative, flexible<br>Con: higher standards & expectations | Pro: realistic expectations<br>Con: less innovative, more carefully planned |
| **3** | Attitude towards the target language | Focus on fluency, meaning, use of language and oral communication skills | Focus on more formal aspects of language |
| **4** | Attitude | Less traditional form of | More traditional form of teaching; teach at the |

| | | | |
|---|---|---|---|
| | towards teaching | teaching; flexible; less goal-oriented teaching | front of the class; use the given materials, methods, tests more often. Work more goal-oriented. |

Based on the fact that native speakers are more flexible, fluent and innovative with respect to both teaching and the English language (see Table 3), one could suggest that native speakers might be more suited for the task of encouraging spontaneous communication through games, songs and drama. Moreover, Heesters et al. (2008) point out that one third of the elementary school teachers of English had not received any specific training in English. Therefore, in order to properly interpret the results of the present study, the types of English teachers present at each school must be compared with the results for the respective school. As will be seen in section 5.6, a distinction was made between teachers who were *native speaker*, *vakleerkracht* or *groepsleerkracht*. The groups were defined as following: Native speakers are teachers who speak English as their mother tongue, were born in an English-speaking country or are unequivocally native in their use of the English language. For the purpose of this study, a teacher was only considered native if their mother tongue was English. The second type of teacher, a *vakleerkracht*, is a teacher specialized in teaching English, albeit not as a mother tongue speaker. These teachers may have minored in early bilingual education and are required to have a minimum level of B2. The third group, *groepsleerkracht*, is a group's classroom teacher who also teaches English. These teachers have received limited training in English, although some may have pursued a minor in English as part of their teaching degree. One of the benefits of a *groepsleerkracht* is that they will be familiar with all the didactic and pedagogical theories involved in teaching young children. An obvious setback, however, is that a *groepsleerkracht* will have received less training than a *vakleerkracht*. Also, a *groepsleerkracht* may be tempted to skip English lessons more often in favor of other subjects (Thijs, Trimbos, Tuin, Bodde, & de Graaff,

2011). As will be seen in the results in section 5, there may be a relationship between a school's average level of English speaking proficiency and the type of English teacher available.

# 3. Research Question

## 3.1 Central Research Question

As a sub-domain of a larger project, this paper will focus on the following question: *What difference in level of English speaking proficiency, if any, does the EarlyBird program yield by the last year of elementary school, when compared to that of regular EIBO schools?*

When answering this question, it will be important to discern not only between the effects of EB group 8 and EIBO group 8, but also between the different grades themselves. For example, a comparison between EarlyBird group 5 and EarlyBird group 8 will be instrumental in assessing a student's potential development within the EarlyBird program. Furthermore, a comparison between EarlyBird group 5 speaking proficiency and that of EIBO group 8 will also be highly relevant.

## 3.2 Additional Sub Questions

In addition to the main research theme, it was also highly crucial to consider other elements that may either validate or refute the findings of the aforementioned question. The first section of the theoretical framework therefore addressed the potential complications involved in language testing. This section attempted to ascertain whether the speaking test method can be considered reliable and valid. The second section zoomed in on the EarlyBird and EIBO programs in order to discern the role of speaking proficiency in both.

The results section will focus on the following questions:

*(1) How do students perform on the speaking test as compared to the other written tests?*

This is a relevant question to ask because it will help determine whether a student's speaking skills as measured in the central question are related to their performance in other language skills. The relationship between the two will help establish whether the answer to the central question is also automatically indicative of a student's other language skills.

(2) *How do speaking proficiency results relate to CITO scores/school exit* level?

If there appears to be a relationship between proficiency and school exit level, then this can help explain the proficiency scores per level.

*(3) What influence can the school have had on the results in terms of its English teacher type and the length of its involvement in the EarlyBird program?*

This sub question will be essential in determining the potential influences on speaking proficiency that are independent of a student's language skills.

*(4) Does the Anglia speaking exam meet the criteria for reliability and validity?*

Addressing this question will help determine whether the speaking test used is reliable and, in turn, whether the results can be considered useful or valid for all testing situations. If the speaking test is reliable and useful, then it can be concluded that the results as measured by the central question are valid.

# 4. Method

## 4.1 Participants

As mentioned in the general introduction, 19 elementary schools took part in this project; 10 of these were EarlyBird schools and 9 were EIBO schools. In the interest of time, only a select group of students per school took part in the speaking test. In both group 5 and in group 8, a random selection of students was generated by selecting every fourth child out of the list of names as presented by each school. Although the selection was random, some criteria did have to be met in order for the child to be selected. At EarlyBird schools, group 5 children needed a minimum attendance of three years and group 8 children a minimum of five years. In other words, all students must have attended their EB school since group 3 or earlier in order to participate in the speaking test. Naturally, this only applied to EB schools, since the years of attendance at an EIBO school was not relevant for the comparative analysis. Other criteria that did concern both EB and EIBO schools was that students on the randomly selected list had to be tested in that exact order. Candidate names could only be omitted if they were either absent or suffered from an impairment that might hinder their performance on a speaking test. If a student's name was omitted, the reason was always properly documented by every test-taker. Once omitted, the next on the list was included. A total of 63 EIBO group 8 students, 65 EarlyBird group 8 students and 60 EarlyBird group 5 students were assessed. More candidates had participated, but only the assessments that required no second moderation were included in the final analysis.

## 4.2 Instruments

The primary component required is naturally the speaking test itself (see Appendix B and Appendix C for examples). This test consisted of four separate tasks. The first task was an introductory task in which the test-taker asked candidates about their families, age, pets and other areas of personal interest. The second task involved questions concerning a large

picture of either a coffee place (for group 5) or the beach (group 8). The third task consisted of an odd-one-out word game in which students were given four words per turn. They were then asked by their interlocutor to work together with their test partner in order to determine the word that did not fit in with the other three words. The fourth task was optional and was only used if the candidate's level was deemed appropriate for the task.

The speaking test was developed by Anglia Examinations England and was designed as a blended exam. This means that each task included questions at several different Anglia levels, as opposed to questions at only one level. This allowed the examiners to ask questions of gradually increasing difficulty in order to gauge the speaking proficiency level at which a student was most comfortable. Per question that was asked, the examiner could tick a box to indicate that a question had been asked. Each question was also color coded to indicate its level. For group 5, the given Anglia levels ranged from *First Step* (pre-A1) to *Preliminary* (A1), with another extra optional task at *Preliminary* level (A1). The speaking test for group 8 ranged from *Primary* (pre-A1) to *Pre-intermediate* (Pre-B1), with an extra optional task at *Intermediate* (B1).

Table 4. *Anglia Speaking Test Levels Per Group*

| Grp. | Range | | | | | | Optional Task |
|---|---|---|---|---|---|---|---|
| 5 | First Step | Junior | Primary | Pre-liminary | | | Pre-liminary |
| | (Pre-A1) | (Pre-A1) | (Pre-A1) | (A1) | | | (A1) |
| 8 | | | Primary | Pre-liminary | Elemen-tary | Pre-interme-diate | Inter-mediate |
| | | | (Pre-A1) | (A1) | (A2) | (Pre-B1) | (B1) |

### 4.3 Procedure

Six English education graduate students from Utrecht University conducted all of the Anglia tests. These students first took a written English proficiency test themselves to determine whether their own level of English was high enough. An Anglia training session was also completed beforehand, after which the interviewers were given a chance to practice during a trial run that was conducted at two EB schools and one EIBO school. Another training and feedback session was held after the first trial run specifically for the speaking test.

During the official testing days, two graduate students were sent to each of the 19 schools to conduct the speaking tests. Since both group 5 and group 8 needed to be tested at EarlyBird schools, each graduate student was made responsible for testing the selected students from one group. At EIBO schools, only group 8 was tested and so both graduate students could be present at the same speaking tests. When this was the case, only one would be the interviewer while the other took additional notes for clarification.

Per group, the randomly selected students were tested in pairs in order to create a more natural and conversational interaction. The proficiency levels, however, were assigned on an individual basis. For EarlyBird schools, a total of three pairs (six students) needed to be tested per group. For EIBO schools, a total of four pairs (eight students) needed to be tested in group 8. The allotted time for group 5 interviews was 15-20 minutes and the allotted time for group 8 interviews was 20-25 minutes. For each school, all tests were conducted on the same day. All interviews were recorded and saved for further analysis using the Olympus VN-5500 PC DNS.

### 4.4 Speaking Test Selection

When comparing speaking test results between EarlyBird and EIBO and between the different groups themselves, it is important to determine whether the random selection of

students that was chosen for the speaking tests is truly a representative selection. In order to determine whether this is the case, the average of the CITO scores was calculated for both the students who took the speaking test and for those who did not.

Table 5. *Average CITO Scores of (Non-)selected Group 8 Students*

| | EIBO grp 8 | EB grp 8 |
|---|---|---|
| **Parameter** | M | M |
| **Selected students** | 536,159 | 538,492 |
| **Non-selected students** | 533,628 | 533,203 |

Differences in CITO scores were compared between selected and non-selected pupils for the speaking test. Analyses show a main effect for selection ($F=12.99$; $p< .001$), but no main effect for EB/EIBO ($F=.772$; $p=.38$). In other words, selected pupils had significantly higher CITO scores than non-selected students, but EB and EIBO groups did not differ on CITO scores. Furthermore, no interaction effect was found between selection and EB/EIBO: the selected EB pupils did not differ from the selected EIBO pupils with respect to their CITO scores ($F=1.616$; $p=.20$). Therefore, we can consider the selected EB and EIBO pupils to be comparable with respect to average CITO scores. Both selections equally represent the same difference from the non-selected group.

**4.5 Scoring and Assessment**

After the testing days had been completed, all participating graduate students assigned scores and levels to the speaking tests they themselves had conducted. This was done by replaying all the recorded tests and assessing them according to an Anglia marking scheme. In order to make a first, general estimate of a candidate's proficiency level, the graduate students could refer to Anglia's *Handbook for Teachers*, in which proficiency descriptors were provided per level. These descriptors gave a general indication of both the vocabulary inherent to each proficiency level, in addition to the types of tasks and sentences each

proficiency level may require. Based on these descriptors, a general estimate of a proficiency level could be made.

When a general proficiency level estimate was made, a second step could be taken. This second step involved assigning scores to separate categories within the level. For each Anglia level, a test could be assessed in terms of five categories, namely Communication, Content, Pronunciation, Range of Vocabulary and Grammatical Accuracy (see Appendix D and Appendix E). Within these categories, a student's level could be rated as Distinction, Merit, Pass, Refer or Unmarked. If a student scored sufficiently high within a level, i.e. their Communication, Content, Pronunciation, Vocabulary and Grammar were all within the Merit or Distinction range, then a candidate's level could be assessed at a higher Anglia proficiency level. If a candidate's scores were primarily within the Unmarked to Pass range, then the candidate might need to be assessed at a lower proficiency level.

Although still pending, it is certainly worth mentioning that the scoring procedure involves a third step as well. As a safeguard against potential subjectivity or inconsistent scoring, a second assessment would naturally be desired, if not required. This assessment is currently being conducted by an official Anglia speaking test examiner. Of the speaking tests conducted, two pairs (four students) are reassessed by the Anglia moderator. If one or more of the moderator's four assessments per graduate students differs, then extra moderation will be required for the student's other proficiency assessments. As already mentioned, however, the moderator's assessments are still pending and her conclusions must therefore be taken into consideration at a later date. For now it must be made clear that the current results are based on assessments as conducted by the graduate students, and that an interrater reliability analysis is yet to be performed. Nevertheless, these results will still provide a representative first comparison between the two different English programs.

**4.6 Analysis**

As mentioned before, this paper will address a total of six questions. For each of the questions, the method for analysis is described below.

*(1) What difference in level of English speaking proficiency, if any, does the EarlyBird program yield by the last year of elementary school, when compared to that of regular EIBO schools?*

To answer this question, the speaking test results of EB group 8 were compared with those of EIBO group 8. This was done by totaling the number of students per proficiency level, and then by comparing these numbers per school type. The same was done for a comparison between EB group 5 and EB group 8, and also for EB group 5 and EIBO group 8.

*(2) How do students perform on the speaking test as compared to the other written tests?*

In order to perform such a comparison, the correlation was calculated between the results of the speaking test and each of the Anglia written tests.

*(3) How do speaking proficiency results relate to CITO scores/school exit level?*

The same test as above was performed on the relationship between the speaking score and a candidate's CITO score. A T-Test was performed to measure whether a candidate's CITO score was a good predictor of their English speaking proficiency level. In other words, is the effect of a candidate's CITO score larger or smaller on their speaking score than on the scores for their English listening, reading, use of English, and spelling skills.

*(4) What influence can the school have had on the results in terms of its English teacher type and the length of its involvement in the EarlyBird program?*

To answer this question, EarlyBird and EIBO curricula, handbooks and methods were studied at great length. Furthermore, the director of EarlyBird provided information concerning the use of native speakers and the active years per EarlyBird school.

*(5) Does the Anglia speaking exam meet the criteria for reliability and validity?*

For this component, the Anglia speaking exam will be held up against the suggested criteria as given by Bachman (1990) and Bachman & Savignon (1986).

# 5. Results

## 5.1. Inter-program Analyses

A total of 63 EIBO group 8 students, 65 EB group 8 students and 60 EB group 5 students participated in the speaking test. 6 students were interviewed per EarlyBird class and 8 students per EIBO class. For some classes there were results for less than either 6 (EB) or 8 (EIBO) students. This is the case for tests where interviewers were uncertain about their assessment. These assessments therefore required a second moderation and were taken out of consideration for the present study. All students were interviewed once, and were awarded one Anglia proficiency level based on their performance during the test. All analyses are based on the data as scored by the graduate students. These analyses do not yet include the data as scored by the Anglia moderator.

Table 5 below shows the average speaking proficiency scores per program and group. As is evident, the average score of EarlyBird group 8 is highest. EIBO group 8 comes in second, and EarlyBird group 5 comes in last.

Table 6. *Speaking Test Score Averages for EB and EIBO*

| | N | Minimum Score | Maximum Score | Mean | Std. Dev. |
|---|---|---|---|---|---|
| **EB grp 5** | 60 | 1 | 5 | 1,683 | 0,948 |
| **EB grp 8** | 65 | 1 | 7 | 4,354 | 1,340 |
| **EIBO grp 8** | 63 | 1 | 6 | 3,333 | 1,231 |

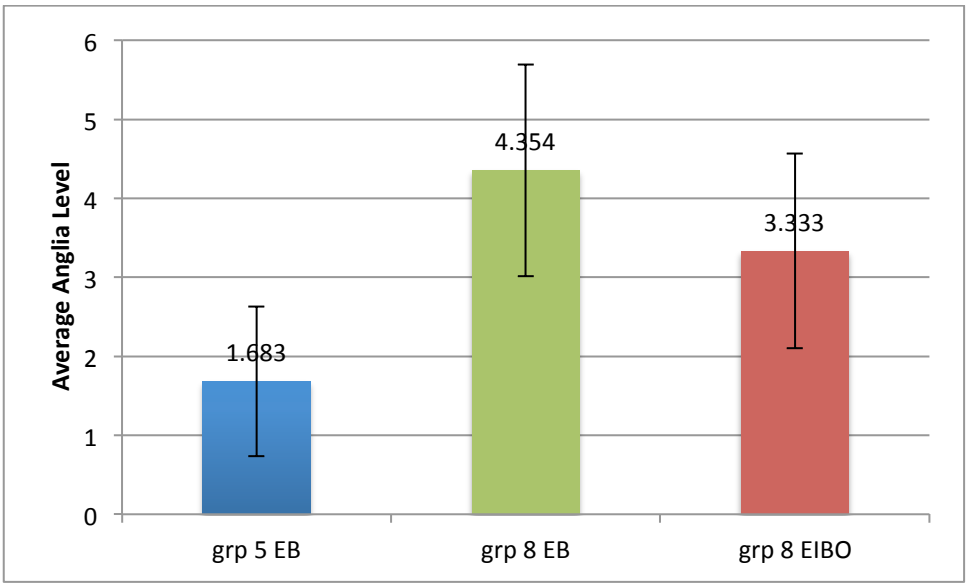


*Figure 2.* Average Anglia Level Per Program Type

### 5.1.1. EB group 8 and EIBO group 8.

Figure 3 below is a first comparison between the final speaking proficiency levels of both programs. The red columns represent the number of group 8 EIBO students in a given level and the green columns indicate the number of group 8 EB students per level. The numbers on the horizontal axis represent the ordinal numbers assigned to each Anglia proficiency level, where 1: First Step (pre-A1), 2: Junior (pre-A1), 3: Primary (pre-A1), 4: Preliminary (A1), 5: Elementary (A2), 6: Pre-intermediate (pre-B1), and 7: Intermediate (B1). The numbers on the vertical axis denote the number of students per level.

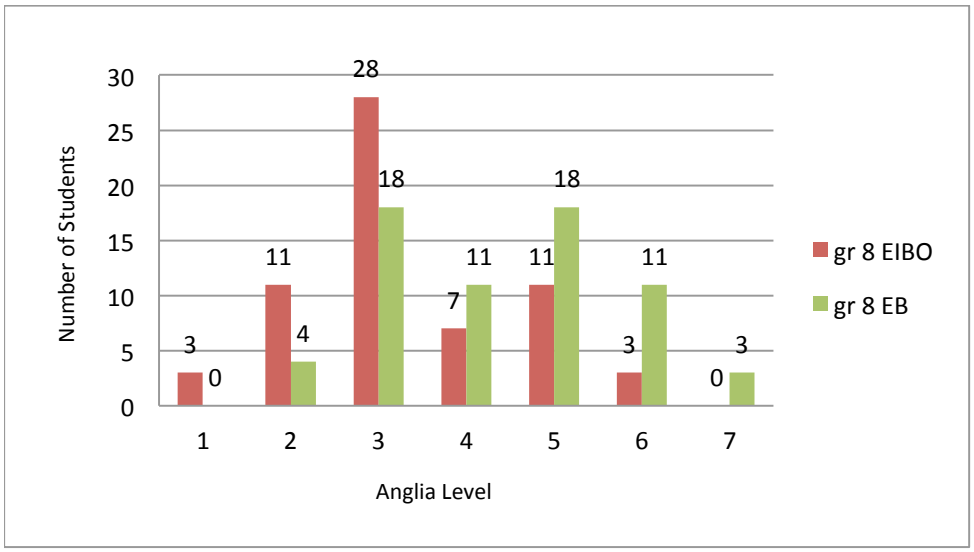


*Figure 3.* EB group 8 vs. EIBO group 8 speaking proficiency

It is clear that most EB/EIBO group 8 students score within the 3 to 5 range (*Primary – Elementary*), and that only a few outliers score well below average (1: *First Step*), or well above average (7: *Intermediate*). None of the group 8 EIBO students scored as high as *Intermediate* (7), and none of the group 8 EB students scored as low as *First Step* (1). Furthermore, the majority of the group 8 EIBO students were at a *Primary* level (3), whereas the majority of the group 8 EB students scored between *Primary* (3) and *Elementary* (5).

### 5.1.2. EB group 5 and EIBO group 8.

In addition to a comparison between the final inter-program levels, a comparison between EB group 5 and EIBO group 8 will also be highly useful. This comparison will indicate how far students will already have come halfway through the EB program as compared to EIBO students by the end of their program. A potential hypothesis was that EB group 5 students may be scoring as well as EIBO group 8 students.

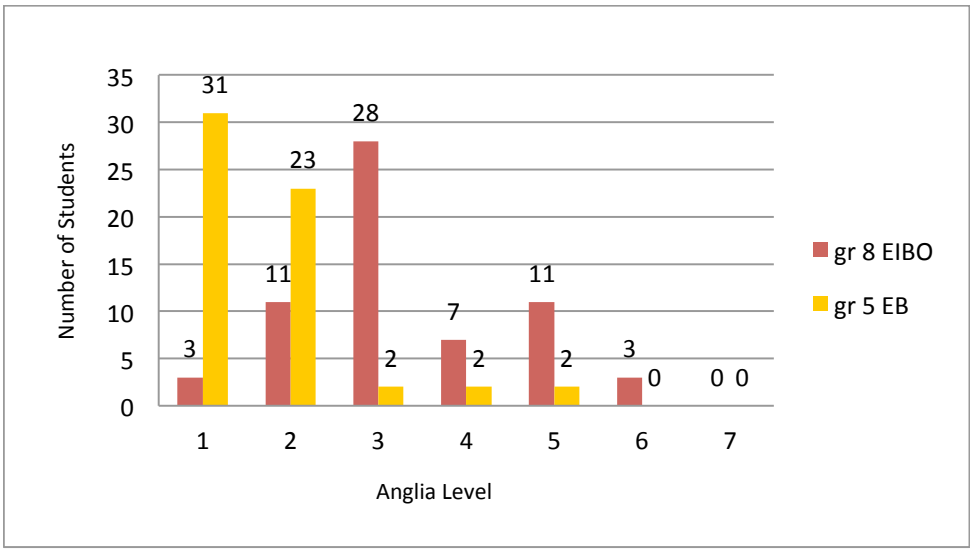


*Figure 4.* EB group 5 vs. EIBO group 8 speaking proficiency

As was already evident in Figure 3, the majority of EIBO group 8 students are at a *Primary* (3) level of proficiency. The majority of the EB group 5 students score within the *First Step* (1) to *Junior* (2) range. Although the majority of both groups score in different ranges, it is interesting to note that there is a slight overlap, though very minor, for five of the six Anglia levels. Nevertheless, it must be concluded from both Table 5 and Figure 4 above that EB group 5 does not yet score as well as EIBO group 8.

**5.2 Intra-program Analyses**

An intra-program analysis was also conducted between EB group 5 and EB group 8. This test was useful for determining the potential development within the program, in addition to possible end terms for the program as well.

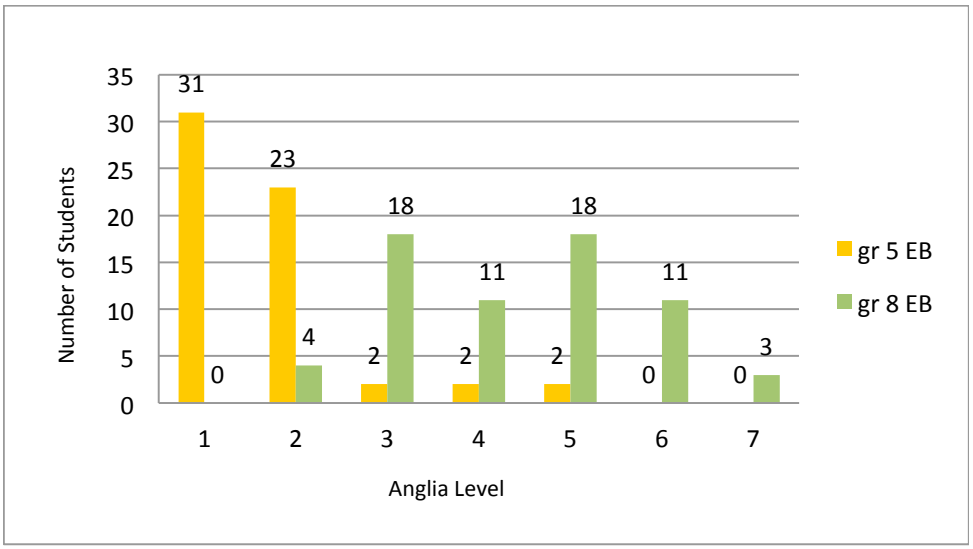


*Figure 5.* EB group 5 vs. EB group 8 speaking proficiency

As can be seen in the figure above, the majority of EB group 5 is at a *First Step* (1) or *Junior* (2) level. A few EB group 5 students score just as well as the majority of the EB group 8 students, but most of the EB group 8 scores are well above those of B group 5. Much progress is therefore made between EB group 5 and EB group 8.

### 5.3 Speaking vs. Other Test Skills

Table 7. *Correlation Between Speaking and Listening, Reading, Use of English, Dictation Scores*

| | | Tspeaking | Dtot | Utot | Rtot | Ltot |
|---|---|---|---|---|---|---|
| Tspeaking | Pearson Correlation | 1 | ,547** | ,754** | ,696** | ,689** |
| | N | 128 | 128 | 128 | 128 | 128 |
| Dtot | Pearson Correlation | ,547** | 1 | ,761** | ,769** | ,745** |
| | N | 128 | 606 | 606 | 606 | 606 |
| Utot | Pearson Correlation | ,754** | ,761** | 1 | ,867** | ,828** |
| | N | 128 | 606 | 606 | 606 | 606 |
| Rtot | Pearson Correlation | ,696** | ,769** | ,867** | 1 | ,844** |
| | N | 128 | 606 | 606 | 606 | 606 |
| Ltot | Pearson Correlation | ,689** | ,745** | ,828** | ,844** | 1 |
| | N | 128 | 606 | 606 | 606 | 606 |

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

For this comparison, a Pearson product-moment correlation coefficient was calculated

in order to determine the type and strength of the relationship between a student's speaking score and their other Anglia test scores, namely those for dictation (D), use of English (U), reading (R), and Listening (L). For the purpose of this paper only the top row is relevant. The other four rows are included, however, as a means for comparison.

For this analysis, $p < 0.01$. In other words, all correlation coefficients with a $p \leq 0.01$ are considered to be statistically significant for this analysis. As is evident in Table 6, all correlations are significant and are therefore not a mere reflection of chance. Furthermore, there appears to be a strong correlation between a student's speaking test performance and their other Anglia tests. The strongest effect can be seen between speaking and use of English. This means that the higher a student's score was on the speaking test, the higher their score was on the use of English test. The smallest correlation is between speaking and a student's dictation score.

**5.4 CITO vs. Speaking**

Figure 6 below shows the relationship between a student’s CITO score and their speaking test score. For each CITO advice level, the average speaking test score is given, both for EB group 8 and EIBO group 8.

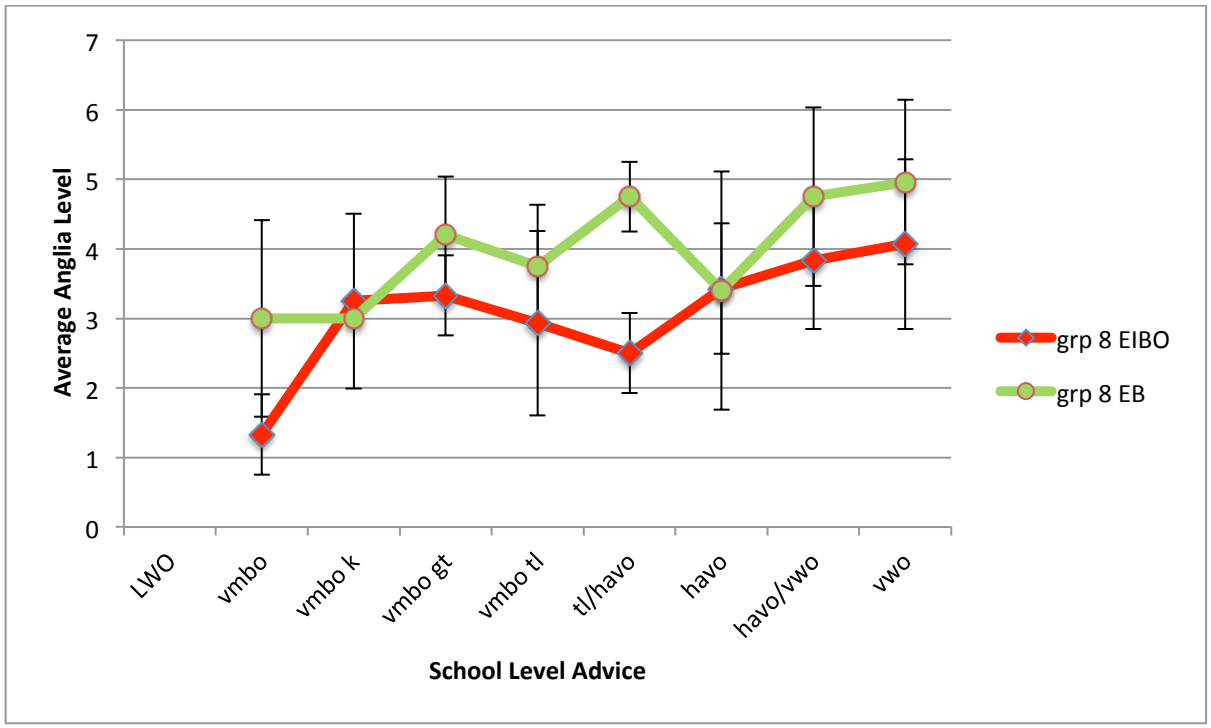


*Figure 6.* School exit levels compared to speaking test score

At first glance, it appears as if a student's CITO might in some ways be also a predictor of a student's speaking test score. Further tests showed that this did not appear to be the case. First of all, the strength of the interaction between the CITO score and the type of test was measured. As can be seen in Table 7, F=20,102; p < .001. This indicates that a student's CITO score may have more effect on some test scores than on others, and additionally, that this interaction is significant.

Table 8. *Interaction Between CITO Score and Anglia Test Scores*

| Source | F | p |
|---|---|---|
| **Test** | 22,876 | ,000 |
| **Cito** | 210,666 | ,000 |
| **vvto** | 13,374 | ,001 |
| | | |
| **Test * cito** | 20,102 | ,000 |
| **Test * vvto** | 7,560 | ,000 |

As Table 8 demonstrates, a student's CITO score is a good predictor of a student's written test scores, but not of their speaking test score. The interaction is not significant for speaking (p = .37), whereas it is for the Listening, Reading and Use of English tests, and slightly less so for Dictation.

Table 9. *Interaction Between CITO Score and Test Type*

| Parameter | t | p |
|---|---|---|
| **Listening * cito** | 2,261 | ,024 |
| **Reading * cito** | 7,079 | ,000 |
| **Use of English * cito** | 4,036 | ,000 |
| **Dictation * cito** | 9,710 | ,000 |
| **Speaking * cito** | -,899 | ,370 |

### 5.5 Interaction School Characteristics and Speaking Test Scores

Another important factor to take into account is the influence a student's school may have on their speaking proficiency test level. A distinction was made between two defining aspects, namely that of the presence of a native speaker (or absence thereof), and number of years a school has been an EarlyBird school (K. Philipsen, personal communication, June 5, 2013). The question at hand is whether either factor appears to interact with the school's average speaking test scores.

For the first defining aspect, only the results of EB group 5 and EB group 8 were taken into consideration, since only these groups can demonstrate a potential difference in speaking test scores as a result of different types of English teachers. For this comparison, a distinction was made between three types of elementary school English teachers. The first group consists of native English speakers. These teachers speak English as their mother tongue, were born in an English-speaking country and are unequivocally native in the use of the English language. The second type of English teacher is a so-called *vakleerkracht*, a teacher specialized in teaching English, albeit not as a mother tongue speaker. The third

group is comprised of a group's classroom teacher who also happens to teach English. These teachers have received little training in English at the primary teacher educational institute, although some may have pursued a minor in English as part of their teaching degree.

Figure 7 below demonstrates the relationship between the type of English teacher a school had and the school's corresponding average speaking test score. As might be expected, schools with a native speaker had the highest average speaking test scores. This appears to be the case for both EB group 5 and EB group 8.

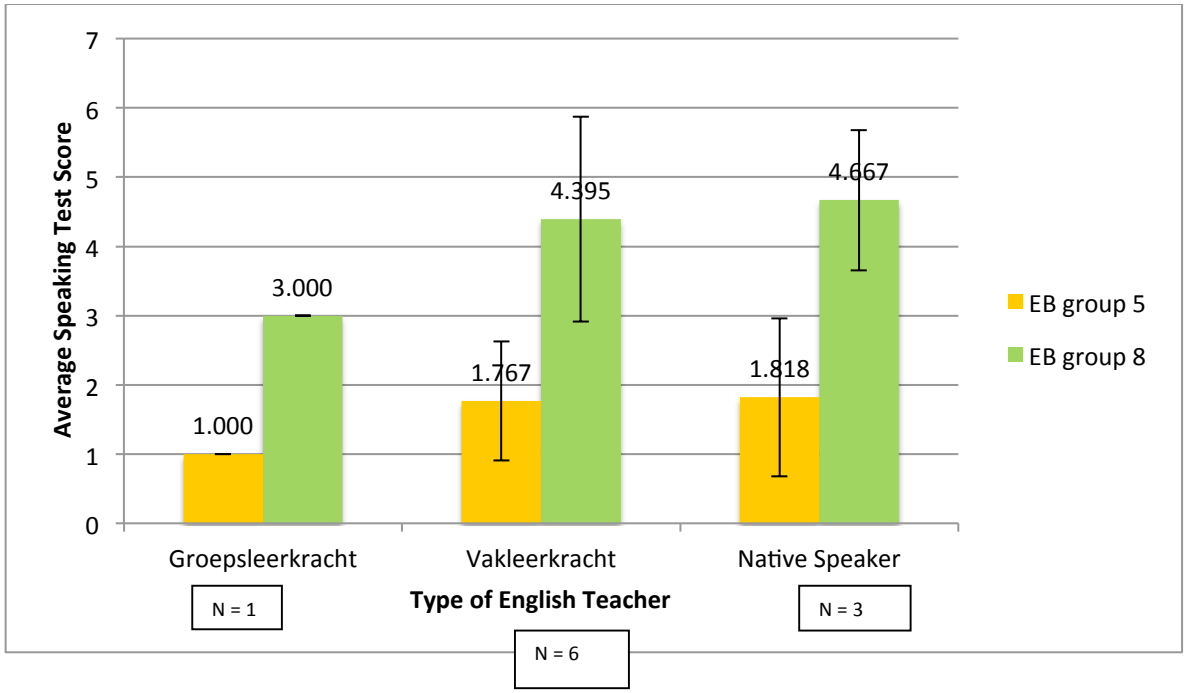


*Figure 7.* Effect of English Teacher Type on Average Speaking Test Score

A second school characteristic to take into consideration is the effect of the number of years the school has been an EarlyBird school. In line with Krashen's theories on exposure and sufficient comprehensible input (1985), the underlying assumption is that the more (i.e. longer) exposure there has been to English, the higher the results will be.

When looking at Figure 8 below, it is difficult to tell whether this is the case. It appears to be true for half of the EB group 8 students, since there is a positive relationship between EarlyBird years and speaking test scores. At the same time there are two other EarlyBird schools that have only had an EB program for five years and yet, on average, score

in a similar fashion to that of the EarlyBird group 8 schools with eight or more years to their credit. The lowest average score was at a school with six years of the EarlyBird program.

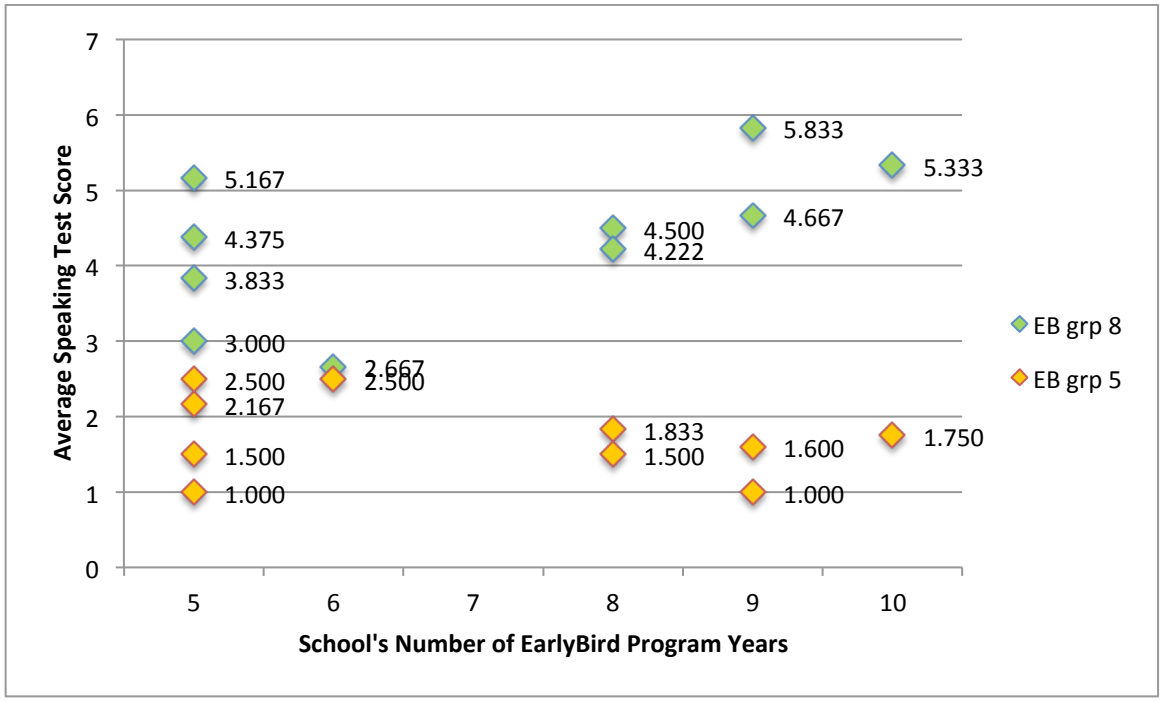


*Figure 8.* Interaction between School's Number of EB Years and Average Speaking Scores

### 5.6 Reliability and Validity of Anglia's Speaking Test

Table 10. *Criteria Checklist Based on Bachman (1990) and Bachman & Savignon (1986)*

| Cited from | Criteria | Criteria satisfied? |
|---|---|---|
| **Bachman (1990), p. 50** | Provide clear and unambiguous theoretical definitions of the abilities we want to measure | Yes. Each interviewer received the Anglia Examinations *Handbook for Teachers* in which the criteria and characteristics of each proficiency level were described at length, including both performance descriptors and vocabulary that were unique for each level. |
| | Specify precisely the conditions, or operations that we will follow in eliciting and observing performance | Yes. All interviewers received both a pre-exam brief (see Appendix F), in which the exact instructions for the test-takers were specified, and a speaking exam itself, in which the questions per exam task were |

| | | |
|---|---|---|
| | | exactly the same for all interviewers. Likewise, the criteria for assessment were also identical. |
| | Quantify our observations so as to assure that our measurement scales have the properties we require | Yes. Each level was split into five separate categories that needed to be tested for: Communication, Content, Pronunciation, Grammar and Vocabulary. Each ability was specified in terms of the elements that needed be present or absent. Each ability was also quantified and assigned a score between 0 (Unmarked) and 4 (Distinction) (see Appendix D and Appendix E) |
| **Bachman & Savignon (1986), p. 384-385** | We must clearly define and distinguish the ability and method factors that we expect to influence ratings. | Yes and no. Yes, because the interviewers were given explicit instructions on the types of language/questions/prompts that were not to be used during this test, due to their potentially (dis)advantageous influence. No, because not all potential factors of influence were discussed, such as an interviewer's attitude, pace, language, form, etc. |
| | We need to design and conduct empirical studies to estimate the effect of test method factors. | Undecided. No contact with Anglia Examinations headquarters. |
| | In interpreting oral interview ratings, we must be clear about whether we want to make inferences about a general domain of CLP [Communicative Language Proficiency] or whether we want to interpret ratings in a more limited (but | No. While conducting these tests, it was assumed that inferences were being made with respect to a test-taker's CLP in general, rather than as a measure of their abilities under certain testing conditions. |

| | |
|---|---|
| | perhaps more meaningful) way<br>as indicators of the ability to use<br>language under specific<br>conditions |

# 6. Discussion & Conclusion

## 6.1 Answer to the Central Research Question

This sub-project of the larger EarlyBird project at hand was set up in order to answer the following main question:

> *What difference in level of English speaking proficiency, if any, does the EarlyBird program yield by the last year of elementary school, when compared to that of regular EIBO schools?*

Based on the findings in sections 5.1 and 5.2, it can be seen that EarlyBird group 8 students scored higher, on average, than EIBO group 8 students. In other words, EarlyBird group 8 students had a higher level of English speaking proficiency. When comparing EarlyBird group 5 students with the same EIBO group 8 students, it appears that the average EB group 5 proficiency level is still well below that of EIBO group 8 students. Therefore, contrary to what was expected, EarlyBird group 5 students do not outperform EIBO group 8 students. Lastly, after comparing EarlyBird group 5 speaking scores with those of EarlyBird group 8, it became evident that there is a significantly large difference in level between the two groups. Much progress is therefore made between group 5 and group 8 in the EarlyBird program. However, as mentioned before, a second assessment has yet to take place, as does an analysis of significance.

## 6.2. Answer to the Sub Questions

In addition to the central question, there was also the following sub questions:

(1) *How do students perform on the speaking test as compared to the other written Anglia tests?*

Based on the correlation analysis between the average speaking test score and the average Listening, Reading, Use of English and Dictation scores, it can be concluded that these scores interact strongly. The highest correlation is between Speaking and Use of English. This can

be explained, since Use of English tested skills that were most closely related to those of productive competence. The lowest correlation was between speaking and dictation. This, too, is a logical finding, since a student's spelling and speaking skills are not necessarily interdependent.

(2) *How do speaking proficiency results relate to CITO scores/school exit level?*

In contrast to the other Anglia (written) tests, a student's CITO score was not a good predictor of their speaking test score. The correlation was not significant, whereas for Listening, Reading, Use of English and Spelling it was. The assumption, therefore, that a student's CITO score is an indication of their level of school achievement appears irrelevant when it comes to a student's L2 speaking skills. This can be attributed to several reasons, one of which is that speaking skills can also largely be a result of a student's exposure to English outside of the classroom. A student can therefore perform poorly on the CITO test and yet demonstrate a high level of English speaking proficiency. Naturally, other aspects such as motivation and natural aptitude can also play a role. These, however, would need to be researched in greater detail.

(3) *What influence can the school have had on the results?*

For this question, a comparison was made based on two factors, namely 1) a school's English teacher type, and 2) a school's number of EarlyBird years. For the first question, a distinction was made between three types of teachers: a. *groepsleerkracht:* own classroom teacher, b. *vakleerkracht:* a non-native teacher specialized in English, and c. a native speaker teacher of English. When comparing the average speaking scores for the three types of English teachers, it was found that schools with a native speaker scored slightly higher than schools with a non-native teacher who was specialized in English (*vakleerkracht)*. At schools with only a classroom teacher of English, the speaking test scores did not exceed a *Primary* (3) level. Schools with a type b teacher scored between *Junior* (2) and *Intermediate* (7). Schools with a

native speaker scored between *Primary* (3) and *Preintermediate* (6). A limitation to this comparison, however, is that there was only a small number of EarlyBird schools involved.

(4) *Does the Anglia speaking exam meet the criteria for reliability and validity?*

For the purposes of this paper, only a theoretical assessment could be made of the speaking test's reliability and validity. As explained before, a test's measure of reliability will indicate the extent to which the test scores are replicable, or whether they are simply based on chance. The validity of a test demonstrates whether the rest results are *meaningful*, *appropriate* and *useful* (see section 2.1.1). It goes without saying that in order to assess reliability and validity, more tests must be done in that respect. However, for the limited scope of this paper, it suffices to say that the speaking test meets the criteria as stipulated by Bachman (1990), although more research would need to be done concerning the required characteristics of a setting, test rubric, input and expected response, to name a few (p. 48-56).

Although further research is required before meaningful claims can be made concerning the results of this project, it is already highly commendable that EarlyBird initiated such an investigation to begin with. With this project, EarlyBird is taking progressive steps towards the fine-tuning and standardization of its program. Not only will the results be relevant for the future of EarlyBird, but they may also serve as a catalyst for change towards the future of bilingual education in the Netherlands.

Before that happens, however, a few points of discussion must be taken into consideration. First, it is important to realize that the results as discussed in this paper are preliminary. An official Anglia assessor will be moderating a quarter of each graduate student's interviews. Although these students received extensive training on how to conduct an Anglia speaking test, there may still be a margin of error involved. Nevertheless, the moderations that have already taken place indicate that there is a sufficient overlap between the assessments of the graduate students and those of the official Anglia assessor. However,

additional analyses will have to be performed once the required moderations have been completed.

Another point for discussion is that the present study did not take into account the two strands of EarlyBird education (early immersion and middle immersion), nor did it take note of which EIBO schools may have already been transitioning into an EarlyBird program. Given the fact that some EIBO schools also have their own version of a middle immersion program (English from group 5 onwards), further research will be needed to determine the potential effect these different program strands may have had on the school's results.

A third point for discussion concerns the use of one type of test for both the EarlyBird and EIBO schools. In his discussion of appropriateness and usefulness with respect to test score validity, Bachman (1990) asserts that "a score derived from a test developed to measure the language abilities of monolingual elementary school children, for example, might not be appropriate for determining the second language proficiency of bilingual children of the same ages and grade levels" (p. 25). In other words, designing one language tests for two very different audiences may give one group an advantage over the other. Although this project does not necessarily deal with groups of monolingual and bilingual children (as in Bachman's example), it does use an Anglia test for both target audiences. As part of the EarlyBird curriculum, however, schools may choose to administer Anglia exams in the final year of elementary school. Students at the EarlyBird schools may therefore have had an unfair advantage in terms of their familiarity with the exam contents. Furthermore, students at EIBO schools have had little to no training in English reading skills, for example, and the one third of the teachers who do train these skills, only make use of Dutch questions instead of English ones (Heesters, Feddema, van der Schoot, & Hemker, 2008). A test score may therefore also be a result of test familiarity rather than of actual language skills. Similarly, the same can be said of English speaking tests, a phenomenon unheard of at EIBO schools. With the

EarlyBird focus on communicative competence, however, EarlyBird students will have received much more training when it comes to demonstrating their proficiency skills under specific conditions. Granted, this project is a test of EarlyBird school performance. That being said, students tested at EarlyBird schools may still have had more of an advantage than simply their extra years of English immersion.

When looking back at the speaking proficiency results, it becomes clear that EarlyBird group 8 students score higher than EIBO group 8 students. It would have been alarming had this not been the case, considering all of the extra training, immersion and exposure that students enjoy at EarlyBird schools. The EarlyBird slogan of ‘more, better and earlier English’ could therefore not be more true.

The implications of these results extend beyond the scope of this study. Not only will more attention be paid to the need for native speakers, but English foreign language learning as a whole may be brought to new dimensions. Combined with the knowledge of a future secured of 15% vvto, early birdies will continue to soar above and beyond.

**References**

Ascentis Anglia. (2009). *Handbook for Teachers: Full Examination Syllabus and Specifications.* Anglia Examinations Syndicate Limited.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford UP.

Bachman, L. F., & Savignon, S. J. (1986). The Evaluation of Communicative Language Proficiency: A Critique of the ACTFL Oral Interview. *The Modern Language Journal, 70* (4), 380-390. Retrieved from http://www.jstor.org/stable/326817

Bachman, L. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford UP.

Berkel, A. van, K. Philipsen & M. Feuerstake (2013).*Spellingtoets Engels voor de groepen 7 en 8 van vvtoE-scholen*. Rotterdam: Early Bird en Europees Platform.

Bygate, M. (2009). Teaching and testing speaking. In M.H. Long & C.J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 412-440). West Sussex: Wiley-Blackwell.

CITO (2006) *Me2!* Retrieved from http://www.cito.nl/nl/onderwijs/primair%20onderwijs/alle_producten/3227a460241f4f2da733ecbcab69dbff.aspx

Doel, R. van den (2006). *How friendly are the natives?* Utrecht: LOT.

Donne, J. (1624). *Meditation XVII.* Retrieved June 18, 2013, from The Literature Network: http://www.online-literature.com/donne/409/

EarlyBird (2010). *EarlyBird Curriculum.* Rotterdam: Boor.

EarlyBird (2010). *Handboek EarlyBird.* Rotterdam: Boor.

EarlyBird (n.d.). *Methodiek*. Retrieved from EarlyBird: meer, beter en vroeger Engels: www.earlybirdie.nl

Frost, K., Elder, C., Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369. doi: 10.1177/0265532211424479

Goorhuis, S., & Bot, K. de (2005). *De ontwikkeling van de Nederlandse taalvaardigheid van kleuters met vroeg vreemde-taal onderwijs.* Groningen: UMCG/RUG.

Goorhuis, S., & Bot, K. de (2005). Heeft vroeg vreemdetalenonderwijs een negatief effect op de Nederlandse taalontwikkeling van kinderen? *Levende Talen Tijdschrift*, 6(3), 3-7.

Heesters, K., Feddema, M., van der Schoot, F., & Hemker, B. (2008). *Balans van het Engels aan het einde van de basisschool 3.* Cito. Arnhem: Stichting Cito Instituut voor Toetsontwikkeling.

Herder, A. A., & Bot, C. L. de (2008). *Vroeg Engels in het Nederlandse Taalcurriculum.* Groningen: Rijksuniversiteit Groningen.

Iwashita, N., Brown, A., McNamara, T., O'Hagan, S. (2008). Assessed levels of second language acquisition: How distinct? *Applied Linguistics*, *29*(1), 24-49. doi: 10.1093/applin/amm017

Krashen, S. D. (1985). *The Input Hypothesis: Issues and Implications.* London: Longman.

Lang, D. Written communication. (June 4, 2013).

Lantolf, J.L., & Frawley, W. (1985). Oral-proficiency testing: A critical analysis. *The Modern Language Journal*, *69*(4), 337-345. Retrieved from http://www.jstor.org/stable/328404

Loon, D. van, & Setz, W. (2012). *Ervaringsrapport Proefproject 15% vvto.* Universiteit Utrecht. Utrecht: Europees Platform.

Ministerie van OCW (2006). *Kerndoelen Primair Onderwijs*. Den Haag: Ministerie van OCW.

Philipsen, K. Written communication. (June 5, 2013).

Platform vvto Nederland (2011). Naar eindtermen vvto Engels, een eerste verkenning. *Europees Platform.*

Shohamy, E. (1999). How can language testing and SLA benefit from each other? The case of discourse. In L.F. Bachman & A.D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 156-176). [Cambridge Books Online version]. doi: http://dx.doi.org/10.1017/CBO9781139524711.009

TalenExpo (2013). Retrieved from www.talenexpo.nl

Thijs, A., Trimbos, B., Tuin, D., Bodde, M., & Graaff, R. de (2011). *Engels in het basisonderwijs: Vakdossier.* Retrieved from www.slo.nl: http://www.slo.nl/downloads/2011/engels-in-het-basisonderwijs-vakdossier.pdf/

Unsworth, S., de Bot, K., Persson, L., & Prins, T. (2012). *Eindverslag FLiPP November 2012.* Universiteit Utrecht, Utrecht.

## Appendix A. Anglia Level Descriptors

| Anglia | CEFR | Group | | Basic Descriptors |
|---|---|---|---|---|
| **First Step** | pre-A1 | 5 | x | - has a basic vocabulary recognition of about 100 words<br>- can read and follow simple instructions<br>- can understand the language of basic identification |
| **Junior** | pre-A1 | 5 | x | - has a basic vocabulary recognition of about 200 words<br>- knows how to describe present actions<br>- can identify and describe basic location and position<br>- can follow a short, simple written text |
| **Primary** | pre-A1 | 5 | 8 | - has a basic active vocabulary of about 300 words<br>- can tell the time<br>- can describe present actions, give personal and family information, describe habits, routines, and everyday activities<br>- can communicate when and how often an action or event takes place<br>- can form questions and negatives |
| **Preliminary** | A1 | 5 | 8 | - has a basic active vocabulary of about 400 words<br>- can communicate present and past events, recently completed actions and life experiences<br>- can communicate where things are and when things happen<br>- can express opposites, comparisons and ownership<br>- can ask questions, answer questions, and write full sentences |
| **Elementary** | A2 | x | 8 | The student has sufficient active vocabulary and structural understanding<br>- to write a short connected text on descriptive or narrative topics<br>- read and understand a text from a familiar range of topics<br>- ask and answer questions about past or present events<br>- distinguish between and use a variety of tenses in familiar contexts: past, present and future<br>- express basic intention, purpose, obligation, preference and advice |
| **Pre-intermediate** | pre-B1 | x | 8 | The student has sufficient active vocabulary and structural understanding to<br>- write a short connected text on descriptive, narrative or imaginary topics<br>- read and understand a text from a familiar range of topics<br>- distinguish between and use a variety of tenses: past, present and future<br>- ask and answer questions about past or present or future events<br>- express basic intention, purpose, obligation, preference, advice, agreement and disagreement, hypothesis and process |
| **Intermediate** | | x | 8 | The student has sufficient active vocabulary and structural understanding to<br>- write clear connected text on descriptive, narrative or |

| | | | | |
|---|---|---|---|---|
| | | | | imaginary topics<br>- read and understand texts from both concrete or abstract topics<br>- distinguish between and use a variety of tenses: past, present and future<br>- ask and answer questions about past or present or future events<br>- express basic intention, purpose, obligation, preference, advice, agreement and disagreement process and hypothesis including regret and consequence<br>- repeat messages, pass on information, check facts |

*Note*. From *Handbook for Teachers* (Ascentis England, 2009)

## Appendix B. Excerpt from Group 5 Speaking Test

## Appendix C. Excerpt from Group 8 Speaking Test

## Appendix D. Performance Descriptors Group 5

Circle the descriptors that match the candidates performance

| MARKING CRITERIA ANGLIA ASCENTIS SPEAKING TEST: FIRST STEP LEVEL | | | |
|---|---|---|---|
| | COMMUNICATION/ CONTENT | PRONUNCIATION | RANGE OF VOCABULARY/ GRAMMATICAL ACCURACY |
| D | The students can comfortably participate in the activities. | Clearly understandable throughout. | The student is clearly at ease with most of the basic words and minimal structures of the level. |
| M | The student can participate in the activities with significant prompting. | Sufficiently adequate to be understandable. | The student knows a few of the most basic words and grammatical structures of the level. |
| P | The student can only participate in the activity with a lot of help and prompting. | Poor, but understandable at least some of the time. | The student knows a few of the most basic words and grammar needed for the level. |
| R | The student cannot get going in the activity despite seeming to try. | The student cannot be understood most of the time. | The student knows too few words to participate in the test. |
| U | Student says *nothing* or virtually nothing in English. | | |

| MARKING CRITERIA ANGLIA ASCENTIS SPEAKING TEST: JUNIOR, PRIMARY AND PRELIMINARY LEVELS | | | |
|---|---|---|---|
| | COMMUNICATION/ CONTENT | PRONUNCIATION | RANGE OF VOCABULARY GRAMMATICAL ACCURACY |
| D | The student can comfortably respond to the examiner's questions. | Clearly understandable throughout. | The student is clearly at ease with most of the basic words and minimal structures of the level. |
| M | The student understands the examiner most of the time and gives a correct answer to at least half of the questions. | Sufficiently adequate to be understandable. | The student knows the basic words and grammatical structures of the level. There may be a few errors. |
| P | The student understands a good proportion of the questions, and gives some right answers. | Poor, but understandable at least some of the time. | The student knows the most basic words and grammar needed for the level although there are obvious errors/omissions. |
| R | A combination of not answering and answering wrongly, making communication impossible. | The student cannot be understood most of the time. | The student knows insufficient words or grammar to participate in the test. |
| U | No communication in English taking place at all. | | |

# Appendix E. Performance Descriptors Group 8

Circle the descriptors that match the candidates performance

| MARKING CRITERIA ANGLIA ASCENTIS SPEAKING TEST: JUNIOR, PRIMARY AND PRELIMINARY LEVELS | | | |
|---|---|---|---|
| | COMMUNICATION/ CONTENT | PRONUNCIATION | RANGE OF VOCABULARY GRAMMATICAL ACCURACY |
| D | The student can comfortably respond to the examiner's questions. | Clearly understandable throughout. | The student is clearly at ease with most of the basic words and minimal structures of the level. |
| M | The student understands the examiner most of the time and gives a correct answer to at least half of the questions. | Sufficiently adequate to be understandable. | The student knows the basic words and grammatical structures of the level. There may be a few errors. |
| P | The student understands a good proportion of the questions, and gives some right answers. | Poor, but understandable at least some of the time. | The student knows the most basic words and grammar needed for the level although there are obvious errors/omissions. |
| R | A combination of not answering and answering wrongly, making communication impossible. | The student cannot be understood most of the time. | The student knows insufficient words or grammar to participate in the test. |
| U | No communication in English taking place at all. | | |

| MARKING CRITERIA ANGLIA ASCENTIS SPEAKING TEST: ELEMENTARY | | | | | |
|---|---|---|---|---|---|
| | COMMUNICATION | CONTENT | PRONUNCIATION | VOCABULARY | GRAMMAR |
| D | Communication is effective for the situation even though answers may be short and hesitation may be noticable. | Shows the ability to speak adequately about the subjects. | Words are very well articulated and can be easily understood. | A good range of vocabulary appropriate for the tasks at this level. | The grammatical forms of the level are confidently used for most of the test. There will be inaccuracies and inappropriate uses when the candidate attempts grammatical forms outside the level. |
| M | There is active participation during the conversation, even if many prompts are needed and there is a lot of hesitation. | Has the ability to speak sufficiently about the subject and can react adequately. | Good articulation, but there may be some mistakes. | An adequate range of vocabulary is used to cover all the subjects discussed, though help may have to be given. | The candidate's use of the grammatical forms of the level is sufficient for all the tasks, although there may be errors. |
| P | Some communication with the examiner takes place, but it tends to be only on repeated prompts, with short answers and with limited scope for active participation. | Can speak about the subjects in a basic way, but no more than that. | Words are sufficiently pronounced to be understood even if there are many mistakes. | Vocabulary is very limited for the level, but is just sufficient to cover most of the subjects discussed. | There may be obvious or even basic mistakes, but the use of grammatical forms appropriate to the level is adequate for understandable exchanges to take place. |
| R | Poor communication with the examiner. | Cannot speak intelligibly about the subjects. | Very poor articulation, virtually impossible to understand. | Vocabulary is not at all adequate for the situation. | The grammatical structures available to the candidate are insufficient. There are very few accurate structures observed at all. |
| U | Little or no communication in English takes place at all. | | | | |

| MARKING CRITERIA ANGLIA ASCENTIS SPEAKING TEST: PRE-INTERMEDIATE | | | | | |
|---|---|---|---|---|---|
| | COMMUNICATION | CONTENT | PRONUNCIATION | VOCABULARY | GRAMMAR |
| D | Communication is effective and comprehensible for the level. It may mostly be short answers, but a reasonable attempt at more extended responses is made too. | Shows the ability to speak more than adequately about the subjects, is clear, and can add personal views. | Words are very well articulated and can be easily understood. | A wide range of vocabulary appropriate for the level is well used. | The grammatical forms of the level are used with reasonable confidence for most of the test. |
| M | There is active participation during the conversation, even if many prompts are needed. | Has the ability to speak sufficiently about the subject and can react adequately. | Good articulation, but there may be some mistakes. | An adequate range of vocabulary is used to cover all the subjects discussed. | The candidate's use of the grammatical forms of the level is sufficient for all the tasks at this level, although there may be errors. |
| P | Some communication with the examiner takes place with prompting and assistance. | Can speak about the subjects but in a very limited way. | Words are sufficiently pronounced to be understood even if there are many mistakes. | Vocabulary is very limited for the level, but is just sufficient to cover most of the subjects discussed. | There may be obvious or even basic mistakes, but the use of grammatical forms appropriate to the level is still adequate. |
| R | Poor communication with the examiner. | Cannot speak intelligibly about the subjects. | Very poor articulation, virtually impossible to understand. | Vocabulary is not at all adequate for the situation. | The grammatical structures available to the candidate are insufficient. There are very few accurate structures observed. |
| U | Little or no communication in English takes place at all. | | | | |

| MARKING CRITERIA ANGLIA ASCENTIS SPEAKING TEST: INTERMEDIATE | | | | | |
|---|---|---|---|---|---|
| | COMMUNICATION | CONTENT | PRONUNCIATION | VOCABULARY | GRAMMAR |
| D | Can keep going comprehensively and express most of what s/he wants to say. There may be pausing for grammatical and lexical planning. | Covers the subjects of discussion satisfacorily. | Clear pronunciation and stress/intonation. | Uses appropriate words and idiom for the tasks at this level. | Inaccuracies and inappropriate uses, but generally confident with the structures demanded by the tasks. |
| M | Candidate is reasonably fluent, but has false starts and repairs. | Covers the subject adequately. | Reasonable pronunciation and stress/intonation. | Adequate words and idiom for all the tasks at this level. | Mistakes are made, but do not seriously break up the flow. |
| P | Candidate can manage tasks and contribute effectively to the discussion, but needs obvious prompting and help to keep going. | Can cover the subject adequately, but needs help and prompting. | Mother tongue easily detected, leading to oddities in stress and intonation, but not generally interferring with understanding. | Just about adequate words and idiom for the tasks, with prompting and help. | Mistakes are made, but the candidate can keep going and make him/herself understood. |
| R | Pauses and hesitation indicating lack of adequate range in candidate's spoken English to cope with the tasks at this level. | Does not cover the subject, is very hesitant about what to say, even with prompting and help. | Flow of pronunciation and intonation not inspiring confidence in the speaker having an intermediate level of spoken English. | Vocabulary too limited to be called functional at this level. | Mistakes indicating intermediate grammar in spoken English not quite achieved. |
| U | Little or no communication in English takes place at all. | | | | |

## Appendix F. Pre-Exam Brief for Group 5 and Group 8 Speaking Tests

## Appendix G. Interview with Denise Lang, Research and Development Team Manager at Anglia

June 4, 2013

**Confidential.**
**For any information, contact the author at R.Knotdehaan@gmail.com**

**Confidential.**
**For any information, contact the author at R.Knotdehaan@gmail.com**

**Confidential.**
**For any information, contact the author at R.Knotdehaan@gmail.com**

**Confidential.**
**For any information, contact the author at R.Knotdehaan@gmail.com**

## Appendix H. Interview with Karel Philipsen, Director of EarlyBird

June 5, 2013

**1. How many EarlyBird schools are there in the Netherlands?**

Ruth: Boor: 34; Buiten bestuur: 219 (deze scholen zijn buiten Rotterdam?)
Karel: Nu 221 scholen (gisteren twee nieuwe) buiten BOOR. Daarvan nu vier in Rotterdam, de rest dus buiten Rotterdam

**2. How many of these schools have:**

**a. A native speaker?**
**b. A near-native speaker?**
**c. A non-native speaker?**

Dat is een onmogelijk te beantwoorden vraag. Daar ging de voor de xxxx niet erg relevante discussie over. Je moet dat definiëren en dan kom je in de problemen. Gaf xxxx ook aan. **(Confidential. For more information, contact the author at R.Knotdehaan@gmail.com).** Als nativeness lastig te definiëren is, dan is near-nativeness nog moeilijker enz. Als ik onderscheid maak binnen de 10 EB-scholen van onderzoek, dan hebben we:

**(Confidential. For more information, contact the author at R.Knotdehaan@gmail.com)**

**3. What, according to EB, qualifies someone as a native speaker?**

Zie hierboven dus..... nog toelichting. In FLiPP-onderzoek werd opgemerkt dat de meeste vooruitgang was geboekt op een school met xxxx op xxxx-niveau. Dat was xxxx op de xxxx **(Confidential. For more information, contact the author at R.Knotdehaan@gmail.com)**

**4. How long has each participating school been an EB school?**

**(Confidential. For more information, contact the author at R.Knotdehaan@gmail.com)**

**5. What were some reasons for why schools chose not to implement the EB method?**

Als ze wel vvtoE gaan doen..... kosten (doen ze liever zelf), ambities programma (gaan ze voor EIBO met een plusje+) of aansluiting bij ander netwerk (TalenT van CHE bijvoorbeeld)

**6. Are EB schools already a certain type of school to begin with? Are these schools, for example, exclusive with respect to the type of student or level of intelligence they require?**

Nee, EB-scholen kom je overal tegen. Gaat meer om initiatief van scholen (dan zijn ze in voor verandering) dan wel besturen (als die denken dat er voldoende draagvlak is).

**(Confidential. For more information, contact the author at R.Knotdehaan@gmail.com)**

**7. Could this result in there being more havo/vwo-type students at an EB school?**

Denk ik niet, hebben we nooit zo onderzocht, was ook niet het doe. Wel weten we dat EB-scholen zich in belangstelling ouders mogen verheugen. Aantal is fors gegroeid, mede maar lang niet uitsluitend door EB-programma. Heeft te maken met die ambities, kwaliteit enz. Neem BZB: had 120 leerlingen bij start EB. Nieuwe directeur, forse inzet EB en veel verbeteringen in programma. Nu > 275 leerlingen en prognose naar 400.

**8. Why did EB choose Anglia and its testing materials, and not, for example, a Cambridge test?**

Anglia, omdat die NL. onderwijs kent, omdat ze in hun systeem een zekere opbouw hebben met behoorlijk aandacht voor Young Learners. Hun database met (veronderstelde) niveaus A1-B1 van ERK werd door xxxx e.a. gebruikt om toetsen te ontwikkelen die er voor zorgden dat van alle kinderen een minimaal niveau kon worden vastgesteld, dat er voldoende differentiatie was. Verder belangrijk dat we de speaking test al geprobeerd hadden en dat we vooral daar zochten naar een eerste ijking van die deelvaardigheid. En tot slot: Anglia gebruikte dit onderzoek ook om zelf meer inzicht te krijgen in hun examens. Subsidie van OCW was bij lange na niet toereikend voor alle inzet van xxxx. Maar Anglia is veel wijzer geworden van dit onderzoek. EB ook, en ook wij hebben veel werkdagen in dit onderzoek - heel veel werkdagen - gestopt in deze exercities. Time well spent!

**9. Has EB conducted similar tests/investigations before?**

- **a. Yes:**
  - **i. What did they test?**
  - **ii. What did they conclude?**
  - **iii. Did these tests include speaking exams?**
- **b. No:**
  - **i. Why not?**

Effectmetingen hebben we gedaan met Mobiel Engels Leren. Beide onderzoeken. SOPA/Peabody. We kunnen je de gegevens toesturen. Geen spreekvaardigheid en in methodiek EB zie je hoe belangrijk we die vinden. Verder niet te veel, onderzoek is duur.

**10. What is at the top of EB's priority list?**

- **a. More EB schools? Different methods? Development of more class material? More native speakers?**

De hamvraag, Ruth. We streven naar groei van netwerk, maar vooral naar opschaling van de werkzaamheden: ontwikkeling methodes, materialen (WordsandBirds, CLIL, teacher training). De discussie over native speakers hierboven nog beetje aangevuld: oud voorstel van Onderwijsraad was die 15 % lestijd, en dan voorkeur voor native

speakers. Onbetaalbaar in huidige condities, onwenselijk (wat gebeurt er op Clipper als de twee vakleerkrachten stoppen.....) en niet nodig bovendien: je moet jezelf gewoon een decennium de tijd geven om vvtoE netjes te laten verzorgen door de groepsleerkracht. Doen ze in andere landen ook. EB prijst zich gelukkig met dat netwerk van natives, zowel op kantoor EB als op aantal BOOR-scholen. Maar dat zal nooit het algemene beeld in NL worden. NL. kan wel van die natives leren, want ze hebben veel ervaring, ideeén enz. EB draagt dat dan graag over!

**11. What does EarlyBird hope to achieve within the *next* ten years?**

Dat in 2020 Engels een verplicht deel van de CITO is, dat kinderen van 12 jaar significant beter Engels beheersen dan nu en dat dat significant beter is dan dat van hun mede-Europeaantjes.