Finding the Key: Examining Language Tests to Identify Predictive Factors

Sierd van den Beld – 3242609
Engelse taal en cultuur: Educatie en communicatie: Master's Thesis
Utrecht University
Supervisor: Rick de Graaff
June 2015

Abstract

For Dutch primary schools, many distinct programmes regarding teaching and learning English as a second language exist. Schools using either of two of these programmes, namely Eibo (*Engels in het basisonderwijs* / English in primary education) and vvto (*vroeg vreemdetalenonderwijs /* early foreign language education), participated in a study to find out if vvto provided pupils with a significant language skill advantage. Pupils were subjected to a battery of language tests, and this paper analyses the results of the written tests in order to identify a key predictor. Additionally, the paper analysed the reliability of the tests, and it explores the similarities and differences between test correlations per group of participants. The results show that Use of English, testing implicit grammatical knowledge and writing, is a potential key predictor for some of the groups tested, namely group 8 (age 11/12) as a whole and the vvto group 8 pupils in specific. Furthermore, it was found that items pertaining to relevant language production were amongst the most important. All tests were sufficiently reliable with the exception of one test: Reading for vvto group 5 (age 8/9). It was also found that the correlations for group 8 were similar, though with different accents for Eibo and vvto, but the correlations for group 5 were weaker. Overall, the findings herein support the overarching study, though educators should be wary of overfocusing on one aspect of language education.

Table of Contents

1. Introduction

In the past decade, Dutch schools have thoroughly embraced the use of English as a language to teach in, at all levels of education. The use of English is most prevalent in secondary education, in which 123 schools (Europees Platform) out of a total of 659 (CBS, 80/81) offer bilingual education at some level, nearly all of which offer English as the second language. Though this was originally focused at vwo-level, it later expanded to include nearly all levels of secondary education (Europees Platform). English in higher education is approximately as frequent: Wächter and Maiworm surveyed institutions across Europe, finding that at least 42 % of all Dutch institutions offer programmes taught in English, comprising at least 17 % of all programmes (25-26).[1] Lastly, approximately one in seven primary schools offer so-called vvto, or *vroeg vreemdetalenonderwijs* as of 2013 (Europees Platform), and as of the 2014/15 academic year, twelve schools have started with a pilot for bilingual primary education (Rijksoverheid).

However, there have been people or organisations that are either opposed to or wary about the rising prevalence of English in Dutch education. The Onderwijsraad, an influential education council that offers advice to the Dutch government, has come out to warn about Dutch in higher education (8), and there has been a court case regarding the state of vvto.[2] Dutch law dictates that all primary education should be taught in Dutch[3], though it also allows that there is room for experimentation if granted permission by the

[1] This is what they call a 'pessimistic scenario'. In the optimistic scenario, encompassing the maximum numbers, these numbers change to 87 % and 34 %.

[2] See rechtspraak.nl, case number 318762 / HA ZA 08-2781.

[3] Art. 9, lid 13. Wet Primair Onderwijs

Education Ministry[4]. This permission, however, is temporary, lasting for a maximum of fifteen years.

The current status of English in Dutch primary schools is that English is one of several subjects the school has to offer by law, and there is a certain level pupils are supposed to attain at the end of their primary school career, though this is not specifically tested in a standardised test, which is only done for mathematics and Dutch: the so-called CITO-toets. The norm for English in primary education, or Eibo, is that pupils in the last two years are taught for an hour a week, though many schools reported they teach less than that: the average reported teaching time is 51 minutes, and over half of surveyed schools reported an average of 45 minutes or less (Geurts and Hemker, 2013, p.40-41). For vvto, this is radically different: Earlybird, a Rotterdam-based organisation dedicated to furthering English vvto, expects schools following their curriculum to teach English for at least an hour a week, starting at age 4, or group 1. This earlier start claims to use what is commonly known as the Critical Period of language acquisition. The Critical Period Hypothesis, in its weak form, claims that the complete acquisition of an L2 “will more *likely* be complete if begun in childhood than if it does not start until a later age” (Saville-Troike, 2005, p.83). However, other researchers contend that the Critical Period Hypothesis has not been proven: Muñoz states that “cumulative exposure [...], input quality and contact with native speakers are more deterministic factors than mere starting age.” (Muñoz, 2014, p. 478). Vvto seeks to offer children a leg up with regards to English by starting to teach the language earlier and by offering more English language education over the course of the eight years of primary education.

[4] Art 2 Experimentenwet Onderwijs

Part of the reason this study is done is because the claim that vvto children are better at English than their regular Eibo counterparts has not yet been substantiated. It is assumed that more English equals better command of English, but because there is no standardised test measuring that, there is no proof either way. Moreover, even if an advantage is present, is it actually a significant advantage? There are many questions, and not nearly enough answers to argue in favour of either side.

2. Theoretical Framework

2.1 L2 Learning

This study seeks to compare two different approaches towards teaching a second or foreign language in primary school. There is a great amount of research concerning L2 learning and teaching, yet these are not always applicable to the current study, due to differing contexts. For example, Flege, Yeni-Komshian and Liu (1999) evaluated the critical age hypothesis in Korean immigrants to the United States, but this study utilised an immersion-based, natural, language learning environment in which English was a second language for the immigrants: the language was required in order "to participate in [the American] community socially, academically, politically and economically" (Saville-Troike, 2006, p.101). Other studies, such as Stæhr (2008), focused on the results of L2 acquisition in the classroom. However the age groups are different from the current study, hindering a direct comparison. .

A constant in language acquisition discussions, supported by research findings, is that input matters a great deal, and a trend is that an early start is seen as beneficial. There certainly are advantages to this early start: it takes advantage of younger children's natural language learning abilities (Genesee, 2004, via Genesee, 2014), and Newport (1990) argued that a young learner's lack of cognitive abilities, compared to an older learner, allows a child to learn less potential language structures because they do not analyse all possibilities. Paradoxically, this shallow pool of data allows for easier learning of morphology, as many incorrect outcomes, which the older learners would analyse, are dismissed. Another advantage of an earlier start, noted by Genesee (2014) is that the integrated structure required by immersion or content-based approaches, is easier to set up in the early part of school, before significant abstraction and complexity are required.

On the other hand, older learners possess advantages as well, such as a greater capacity to learn and “better memory for vocabulary” (Saville-Troike, 2006). They are also capable of utilising their increased knowledge of the world and of their L1 to effect a greater transfer of skills, such as literacy and writing (Genesee, 2014). The more abstract nature of language teaching in secondary schools can also benefit them, allowing for learning strategies that make use of their analytical skills (Genesee, 2014). Additionally, when it comes to immersion-based learning, there can be a self-selecting component involved for those who start learning an L2 relatively late, which can be a beneficial factor (Genesee, 2014).[5][6]

The ultimate attainment, or UA, of younger and older language learners differs. Though younger learners in an immigration setting had higher UA after a “long period of residence”, the older learners' increased brain capacity allowed them to acquire language more rapidly at first (Muñoz, 2014), and in Canada, late immersion program pupils were not consistently outperformed by early immersion pupils (Genesee, 2014). It was often the case, yet not always, suggesting that exposure and use are not the only factors involved. However, these situations, again, concerned second language situations, as opposed to the foreign language situation that is present in this study. In fact, research suggests that the advantage the early learners have in a naturalistic language learning context might not be around in the foreign language learning context. This is primarily due to the limit of effective input present in the FL-situation, which hinders younger

[5] Genesee mentions immersion-based Canadian schools, contrasting early immersion (from age 4 onwards) and late immersion (age 12 onwards).

[6] This contrast is also somewhat present in Dutch secondary schools, cf. Regular English, 'Versterkt Engels', and bilingual education.

learners' “implicit learning mechanisms”, which require a great deal of input (Muñoz, 2014). This is possibly due to the locative and contextual memory younger children possess (van der Plank, 2008), which means that the young learners remember specific instances of language being used. In contrast, the older learners, whose brains are more able to grasp abstractions and who are more able to recall words and phrases in lists (van der Plank, 2008), have an advantage in explicit learning (Muñoz, 2014).

2.2 Early learners (vvto) and Late learners (Eibo)

In the Netherlands, there are several guidelines for the final attainment levels of pupils in primary education. These guidelines differ somewhat from the Common European Framework of Reference for Languages, though one guideline, which states that pupils should be able to gather information from simple spoken and written texts (Ministerie OCW, 2006), is similar to the CEFR's descriptor of the A2 level for Reading for Information and Argument, which states that someone at that level “can identify specific information in simpler written material.” (Council of Europe, 2001) However, as these are guidelines, there are no consequences involved for schools if pupils do not attain these levels. In addition, the amount of time spent on English varies greatly between schools, especially when taking vvto into account. A CITO-groep survey (*PPON)* found that surveyed vvto schools spent anywhere from 216 to 507 hours on English in the eight years of primary education, compared to the average of 64 hours the Eibo school headmasters reported[7][8] (Geurts and Hemker, 2013, p.40). These 64 hours are concentrated in the last two years of primary school. Some schools offer alternate versions of Eibo, such as teaching more hours of English (*versterkt Eibo),* earlier

[7] Primary schools are required to teach at least 7520 hours in eight years. (Wet Primair Onderwijs)

[8] Group 8 teachers reported a number approximately 5 % lower.

instruction in English *(vervroegd Eibo),* or *Content and Language Integrated Learning,* or CLIL, which focuses on projects or subjects in English, as opposed to foreign language instruction (Holdinga, 2007, p. 13). These alternatives bridge the gap between usual Eibo and vvto.

In this study, the vvto pupils all attended schools that utilise Earlybird programmes. Earlybird is but one option for vvto in the Netherlands, with approximately a quarter of the thousand vvto-schools in the Netherlands following its programme (Earlybird vvto Landelijke Ontwikkelingen 2015; Rijksoverheid, 2015). Earlybird's programme aims to provide at least one hour per week of instruction in English, for all age groups, allowing for at least 320 hours of English by the metric the PPON used (Earlybird FAQ, 2015). Though this still falls short of immersion-based programmes as described in Genesee (2014), it is a great deal more than Eibo schools teach. Schools utilising the Earlybird programme focus on listening and speaking in the first few years, with reading and writing only gradually incorporated from the second half of grade 4 onwards (Earlybird Methodiek, 2015), thus ensuring that the teaching of reading and writing in Dutch, which is traditionally started in grade 3, or ages 6-7, is not hindered by introducing another language with its own idiosyncrasies. The question remains, however, if this limited amount of input is sufficient to avoid the problems that Muñoz outlined with regards to foreign language teaching. Earlybird may offer more hours of English, but “[e]ffective instruction is critical if the extra time and early start are to be advantageous” (Genesee, 2014, p.29).

2.3 Interconnectedness of Language

Learning to utilise a language is an incredibly complex task, as language is used for several skills. For example, speaking requires a learner to utilise knowledge of content, lexis, the sound system and information, amongst other skills (Hinkel, 2006, p. 114), and when speaking and listening are utilised simultaneously, as they are in a normal conversation, there is an additional need to identify potential problems “at the fast pace of a real conversational exchange” (Hinkel, 2006, p. 114). Similarly, learning to read and write draws upon several subskills of language as well. Crystal identified “phonology, graphology, vocabulary, grammar, discourse and variety” (2003, p. 442) as being essential to the language learning process. With the exception of graphology, which Crystal defined as “the (study of the) writing system of a language” (2003, p. 463), these subskills are relevant for the speaking process as well, and an understanding of them is required for obtaining the relevant listening skills. As such, the observation that language skills are used in tandem, as opposed to separately (Hinkel, 2006, p. 113) is one that is very important within the current communicative paradigm of language teaching.

However, it is hard to say if proficiency in one skill can be used to predict proficiency in another skill, moreso when taking the foreign language, relatively young learners, context of this study into account. Nation and Snowling (2004) suggested that oral language skills are able to predict a child’s reading comprehension in their L1. In the study, they tested children aged 8 to 13.5 year old with English as their L1. They claimed that “English speaking children must learn to read not only regular words [..] but also exception words” (Nation & Snowling, 2004 p. 352). Such exception words include *bow*, the pronunciation of which is contextually influenced, as the word can be both a noun and a verb. The child with better oral language skills is able to use that increased awareness of vocabulary and syntactic skills to better understand the exceptions. Feyten (1991) argued that skill in listening can be indicative of overall skill in language. Her research, following students enrolled in a “summer intensive language program”, tested for proficiency in two foreign languages, adding to the predictive value of her findings. On the other hand, the correlations between skill in listening and the overall skill, consisting of the average score on separate speaking, listening, and grammar-reading-vocabulary tests, were relatively weak.

Bozorgian (2012) found that, for Iranian subjects using the International English Language Testing System’s tests, listening on the one hand, and reading, writing, and speaking on the other hand were significantly correlated. However, Bozorgian’s findings were about adult L2 learners, which is a different age group from the tests carried out with regards to the target group in the current study. Stæhr (2008) found that 15 and 16 year olds’ scores for the vocabulary part of the Danish national school leaving examination were strongly correlated with scores in reading, writing and listening parts. The correlation between reading and vocabulary scores was highest at .83, which is congruent with other studies showing a strong correlation between reading comprehension and vocabulary (Aarnoutse & van Leeuwe, 2010; Qian, 2002; August et al, 2005).

There are several reasons why finding a correlation between language skills is important. First of all, there has been comparatively little research with regards to the interdependentness of the language skills, even though there are many examples of two skills that have an effect on each other, such as reading comprehension and oral proficiency (Mills, 2009), reading and writing (Couzijn, in Rijlaarsdam, 2005), or the aforementioned study by Nation and Snowling. The research that does investigate more than two skills often focuses on one variable as the key component, and although correlations are present, it is not possible to give an indication as to which skill is the most important as a predictor. Another caveat is that the results on a test do not necessarily translate to skill in a language: other factors, e.g. IQ, amongst others, can also make the results and the skill interdependent. Overall, the aspect of exactly how interconnected the language skills are has not been extensively documented, even without taking the age of the participants into account. Secondly, the correlation could shed light on which skill serves as the best predictor of learner proficiency. Though Stæhr and Bozorgian's studies do serve as an indication, the age groups of their learners were different from the group in this study, which consists of younger learners of a foreign language as opposed to older learners of a foreign language. The study is also different from Nation and Snowling's research in that it does not measure skill in the learner's L1, though the age group does match in this regard. The findings of this study could support conclusions drawn earlier when testing which skill best predicts foreign language proficiency in older learners, but they are relevant first and foremost for L2 teachers in primary and early secondary education, as these ages correspond to the ages tested. If a correlation is present, which is very much expected from the available literature, then the teachers could gain additional insight into which aspect of language could be deemed most important to teach. Additionally, there would be less incentive to test pupils in multiple skills if proficiency in one skill is strongly correlated with the others, thus possibly reducing the amount of tests pupils take. However, the correlation does not guarantee a skill that is important; the correlation could also come from teaching skills in conjunction: e.g. listening and speaking can be, and often are, taught together.

2.4 Research question

This paper is focused on the degree to which the five proficiency tests, measuring skill in Listening, Reading, Use of English, Speaking, and Spelling utilised by the overarching study, and the skills used by the learners, belonging to three distinct groups, can be assigned a predictive value; i.e. can learners scoring well on any test provide teachers with valid expectations for other tests? The main question, therefore, is as follows:

- Which skill can best serve as a key predictor for other skills in a language test battery for English in primary education?

For a test to qualify as a key predictor, the aggregated *coefficient of determination*, or r^2, has to be of a sufficient level.

In order to find out whether or not a test can be a key predictor, it is necessary to first examine the reliability of the tests, as insufficiently reliable tests cannot be used to give a good indication of learner skill. Additionally, as the predictive value is primarily based on the correlations between different tests, the intercorrelation between all different tests, at all tested levels, is relevant as well. As such, this paper will answer these two sub-questions:

- Are the tests designed for this study sufficiently reliable?
- What are the correlations between all the tests, for each group of participants? Are these correlations similar for all groups?

Previous research indicates that strong correlations can be present, though it is uncertain as to which skill, if one skill in particular, proves to be the keystone. Given the set-up of the Earlybird curriculum, in which listening and some speaking starts at age 4 and formal instruction in reading and writing in English tentatively starts at age 7/8 (Earlybird website), the presence of a strong correlation between the oral and written tests in the vvto group 5 participants appears less likely, as pupils might not have enough experience with reading and writing in English just yet.

3. Method

The overarching project sought to determine whether or not the Earlybird programme was effective at teaching the children a second language. Two groups were formed: 10 Earlybird, or early start, schools provided group 5 and group 8 pupils, and 9 Eibo, or late start, schools provided group 8 pupils to contrast with the Earlybird group 8 pupils. As Eibo schools generally only start teaching English in group 7, only group 8 was tested for these schools. All Earlybird schools had used the programme for a minimum of eight years.

In order to gauge the pupils' proficiency, several written tests were conducted, testing the ability of pupils in Listening, Reading, Use of English and, for group 8 pupils, Spelling. Additionally, about one in five pupils were randomly selected to take part in a Speaking exam, done in pairs.

3.1 Participants:

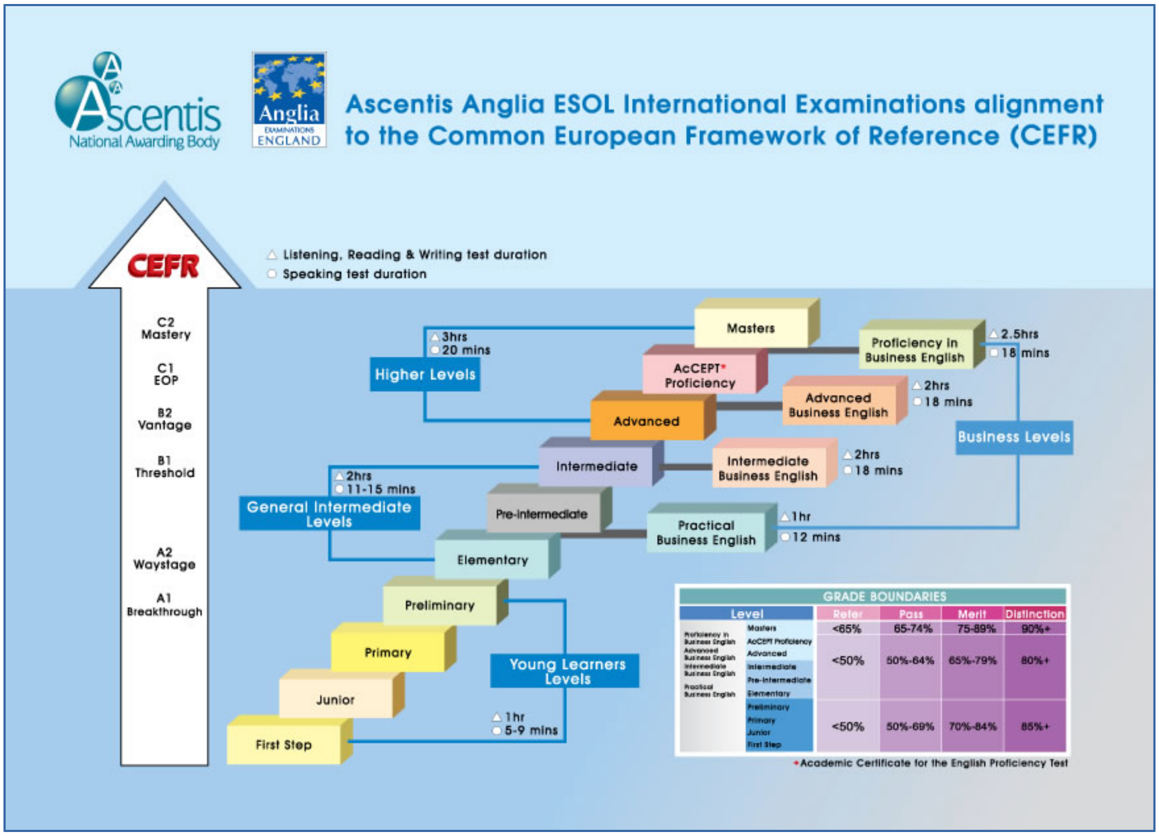
In total, 811 pupils took part in this study. Of these, 318 were group 5 Earlybird pupils, 301 were group 8 Earlybird pupils and 292 were group 8 Eibo pupils. Earlybird pupils that were selected were required to have been at an Earlybird school since group 3, or age 6-7, in order to ensure that the data was a representative reflection of the school and its teaching method. Pupils that were absent on the day that their class was tested were excluded from this data, but no other factors, such as learning disabilities and other or additional L1s than Dutch, were taken into account for the overall results. Participants received a personal registration number to ensure privacy, and parents were notified about the tests and asked for consent. Parental consent was not properly obtained in one school, leading to the discrepancy of 10 Earlybird schools, yet only 9 Eibo schools.

3.2 Instruments:

The participants had to take several tests, among which were Reading, Listening and Use of English, which tested the vocabulary, implicit grammatical competence, and knowledge of sentence construction of the participants. All these tests were tailored to the estimated skill level of the age group: group 5 did not do the same tests as group 8, though there was overlap present in some of the tests. The questions on these exams were primarily provided by Anglia Network Europe, which created the exams utilising various texts and questions of differing difficulty. Additionally, they were also responsible for overseeing the development of the tests, with some additional questions provided by Earlybird. Group 8 also had to do a written spelling exercise made by Ans van Berkel (van Berkel, Philipsen, & Feuerstake, 2013). Lastly, the Speaking test, which was done by approximately one in five pupils, was also created by Anglia Network Europe, using several of their speaking tests as the basis. Anglia also trained undergraduate students in correctly administering this Speaking test. All tests were administered on the same day, under the supervision of the undergraduate students, as well as the regular group teachers for the written tests if there were more groups than undergraduate students. The written tests were done individually, and the speaking exam was done in pairs, as per Anglia’s standards.

All tests, with the exception of the spelling exercise, were specifically designed for this exam. This was done in order to adequately survey the level of the pupils, as that level can vary greatly from pupil to pupil. All questions were taken from different Anglia tests, ranging from First Step to Proficiency on the Anglia Examinations alignment. These levels are roughly equivalent to below A1 to C1 on the scale used in the Common

European Framework of Reference, though the connections between the Anglia scale or single Anglia test items and the CEFR are estimates that were not further validated at item level in this study. For a schematic overview, please see Figure 1.



Level		Refer	Pass	Merit	Distinction
	Masters	<65%	65-74%	75-89%	90%+
Proficiency in Business English Advanced Business English Intermediate Business English Practical Business English	AcCEPT Proficiency Advanced Intermediate Pre-Intermediate Elementary	<50%	50%-64%	65%-79%	80%+
	Preliminary Primary Junior First Step	<50%	50%-69%	70%-84%	85%+

*Academic Certificate for the English Proficiency Test

Figure 1: Anglia Examinations alignment (Anglia Handbook for Teachers, 2009)

Prior to the study, a small scale pilot was run on two Earlybird and one Eibo school in order to ensure the validity and reliability of the tests. The undergraduate students made notes of all questions asked during the tests, but later marking also revealed that several questions did not properly assess the pupils' skill. As such, modifications were made, and these modifications later became the exams that were used in this study.

All texts and questions, except for van Berkel's Spelling exercise, came from Anglia tests at a certain level. Table 1 shows which level of tests the questions for the written exams came from, in percentages. Note that this does not indicate the level of the individual items, and the following table should be used as an indication, rather than an exact measurement.

Table 1: Source of the questions on the tests[9]

Test	Level	First Step Junior Primary[10]	Preliminary	Elementary	Pre-Intermediate	Intermediate	Advanced	Proficiency
Reading Group 5		26 %	53 %	21 %				
Listening Group 5		24 %	20 %	24 %		11 %		20 %
Use of English Group 5		38 %	43 %	19 %				
Reading Group 8				33 %	17 %	33 %	17 %	
Listening Group 8		8 %	15 %	4 %		35 %		39 %
Use of English Group 8		22 %	22 %	22 %	17 %	17 %		

[9] Percentages can exceed or fall short of 100 % due to rounding.

[10] These levels are judged to be below A1 on the CEFR scale. As such, they are grouped together.

The speaking exams had a range of First Step to Intermediate, with the examiner selecting questions that were thought fit for the pupil's level of understanding. The main part of the speaking exam was designed so that group 5 pupils were asked questions at the below A1-levels of First Step, Junior and Primary, and group 8 pupils were asked questions from Primary up to Pre-Intermediate. All examiners had a sheet of questions and the levels they were judged to be. This allowed the examiner to ask pupils questions that corresponded to their level of understanding and production, thus ensuring that pupils were producing enough language to grade them on. If the speaking exam was deemed too easy for the pupils, the examiner was allowed to offer them an additional exercise at the Preliminary level for group 5 and the Intermediate level for group 8.

The amalgamation of the texts and questions of different levels was something that had not been done before by Anglia, and as such, there is a slight risk of certain exercises being less effective or, for the Speaking exam, less determinate than they are in their original mono-level context.

3.3 Procedure:

All tests were conducted between April 9th and April 25th 2013 with two examiners going to every school in order to do so. In all cases, the Listening exam was the first test the children had to do, though the order following that was up to the examiner to decide, depending on several factors such as varying classroom hours from school to school. It was advised in a manual handed out to the undergraduate students, as well as the schools, that the order was to be Listening, followed by Reading, Use of English and, for group 8, Spelling in the morning, with the afternoon taken up by the Speaking Exams. In the event that there were more different classes than examiners, either the regular teacher or the

Earlybird native speaker conducted the written exams as per instructions laid out in the manual. Only the undergraduate students, who had been trained by Anglia, were allowed to conduct the Speaking test.

Prior to each written test, the examiner told the pupils about the number of questions that were in the upcoming test, the amount of different parts that the questions were divided over and the time that they had for the exam. In addition, for Listening, the pupils were told, both by the examiner and by the recording that was used, to read the questions before the relevant section started. For Reading and Use of English, pupils were told that, if they were unable to finish the exam within the allotted time, they were to underline the last question that they answered. All written tests were done individually, and though pupils were allowed to ask questions, the examiners were not allowed to answer in more than generalities.

For the Spelling, the undergraduate students or group teachers read out the full sentence twice, followed by speaking the two words that the pupils had to write down twice. At the end, all fifteen sentences were read out once more.

The pupils for the Speaking exams had been randomly selected beforehand and a list was sent to the examiners to work off. The exam was done in pairs, without regard for estimated level, which meant that it was possible for a Pre-Intermediate pupil and a Primary pupil to be in the same exam. Individual questions were asked for the first two parts of the exam, but the last part required the pupils to work together. All speaking exams were recorded using a portable sound recorder.

3.4 Analysis:

All written exams were marked by the undergraduate students, filling in every answer in an Excel sheet so that the data was readily available for SPSS analysis later. Missing values, such as unanswered questions, were given a separate code. Though it was not possible to definitively score pupils as belonging to a certain level on the CEFR or in the Anglia hierarchy for the written tests, it was possible to score the Speaking exams on the latter scale. Pupils were assigned a certain level by the examiners, with a professional moderator giving a second opinion to ensure that the examiners were grading according to official Anglia guidelines. All levels assigned for the speaking tests were according to the Anglia hierarchy in order to be able to have a better differentiation at the lower levels, primarily below A1 on the CEFR.

To determine the scores, every answer that was given was put into an Excel file, which was then converted into SPSS, allowing for several analyses to be carried out. The mean scores and standard deviations were calculated, using several variables, such as Eibo or Earlybird. Additionally, Multilevel analysis was carried out, on both school~ and pupil level. Moreover, Pearson Product-Moment Correlation Coefficients were utilised in order to express the inter-test correlations. In this, the significance of the correlation was also expressed. In order to calculate whether a correlation is significantly higher between different groups of participants, *Fisher's r to z* transformation was used. If the outcome, expressed as *Z,* was statistically significant, then it follows that the correlations differ between the groups of participants (Kenny, 1987). In order to calculate whether a test qualifies as a key predictor, the average of all correlations is squared (r^2), as this indicates which proportion of the results was predicted by other variables. If the r^2 exceeded 0.5, then a test can be a key predictor, as more than half of the score can be attributed to influence from other variables (Kenny, 1987). For group 8, the r^2 was calculated twice:

once taking van Berkel's et al.'s Spelling exercise into account and once leaving it out, as excluding this exercise allows for analysis of the tests made specifically for this study. Lastly, the reliability, as expressed by Cronbach’s Alpha, of each written and amalgamated test was calculated, on the level of the individual test as well on the level of separate questions.

4. Results

4.1 Group 8 results

The test results for group 8, as seen in Table 2, revealed that Earlybird pupils scored approximately 5-6 % higher on average than Eibo pupils in all tests, with comparable standard deviations. Earlybird pupils scored higher on Use of English in particular.

Table 2: Results group 8 per test

Group 8	**Vvto**		**Eibo**	
	Mean	Std	**Mean**	Std
Listening	**0.83**	0.11	**0.78**	0.13
Reading	**0.67**	0.16	**0.61**	0.16
Use of English	**0.68**	0.15	**0.60**	0.15
Spelling	**0.64**	0.24	**0.58**	0.24
Total	**0.70**	0.19	**0.64**	0.19

As the tests were not calibrated to conform to a single skill level, for example CEFR B1, the score differential between individual tests is not indicative of more or less skilled pupils in that area.

Multilevel analysis found that the higher mean scores by Earlybird pupils were statistically significant for Listening (t= -2.88[11], p <0.01), Reading (t= -2.70, p <0.01) and Use of English (t = -5.32. p <0.001), as can be seen in Table 3.

Table 3: Effects of vvto on tests

	t	p
listening * vvto	-2.884	0.004
reading * vvto	-2.695	0.007
use of English * vvto	-5.322	0.000
spelling * vvto	-1.956	0.062

[11] A negative t-value indicates that Eibo scored lower than vvto.

4.2 Group 5 results

Table 4: Results group 5 per test

Group 5	**Mean**	Std
Listening	**0.68**	0.13
Reading	**0.69**	0.13
Use of English	**0.60**	0.13

Several items overlapped between the group 5 and group 8 tests. 62 % of the Listening tests, 21 % of the Reading tests and 55 % of the Use of English tests were the same for both groups. The overlapping items were isolated and analysed, but no correlations consisting of just the overlapping items were calculated, as this did not fit in the scope of the study.

Table 5: Mean scores and effect sizes on overlapping test items

	Gr 5 (Mean)	Gr 8 (Mean)	Effect (d)	Gr 5 (Mean)	Gr 8 (Mean)	Effect (d)
	vvto	Eibo	vvto gr 5 Eibo gr 8	vvto	vvto	vvto gr 5 / 8
Listening	.64	.82	1.11	.64	.86	1.41
Reading	.49	.78	1.78	.49	.83	2.21
Use of English	.45	.74	1.72	.45	.81	2.23

Both regular Eibo group 8 and vvto group 8 scored higher than the group 5 pupils on the overlapping parts, with significant effect sizes. Moreover, the mean scores on the overlapping segment are all below the overall average for group 5, whereas they are above average for both Eibo group 8 and vvto group 8, indicating that this overlap consists of easier exercises in the group 8 test and harder exercises in the group 5 test.

Lastly, 188 pupils, divided roughly equally over the three groups, participated in speaking tests. Though the questions given differed between both sets of group 8 participants and the group 5 participants, the scale upon which these pupils were scored was the same, ranging from First Step (level 1) to Intermediate (level 7). As such, it is possible to compare the results of all three groups, as can be seen in Table 6. The average level of Eibo group 8 pupils was 3.33, or somewhere between Primary and Preliminary on Anglia's scale. The average level of the Earlybird pupils in group 8 was 4.35, somewhere between Preliminary and Elementary. This difference is statistically significant (p=-0.028). Group 5 pupils had a mean score of 1.68, which is between First Step and Junior.

Table 6: Mean speaking scores, all groups.

	N	min	max	**Mean**	std
Eibo group 8	63	1	6	**3.33**	1.23
Vvto group 8	65	2	7	**4.35**	1.34
Vvto group 5	60	1	5	**1.68**	0.94

To summarise: it was found that, for group 8 pupils, vvto had significant interaction effects with the written tests. Vvto was also significant for the speaking test, and both group 8 subgroups scored significantly better than group 5 on overlapping items and the speaking test.

4.3 Reliability

Cronbach's Alpha expresses the internal consistency of the tests. A high internal consistency indicates that the questions in the exam are answered in such a way that whether or not someone answers a question correctly is an indication of their overall score and skill. A Cronbach's Alpha value of .7 or higher indicates that the items are in fact "highly correlated" (Lowie & Seton, 2013). Lowie and Seton also mention that an acceptable Cronbach's Alpha is usually around .80, though it depends on how critical one wants to be.

4.3.1 Reliability Group 8

The results of several questions in the listening exam for group 8 were removed from the analysis performed. One of these questions was removed due to faulty results, whereas the others lacked variance.

The internal reliability was generally quite high. All four written tests in group 8 scored above .82, meaning that they were all acceptably reliable. Spelling had the highest Cronbach's Alpha, at .89.

Table 7: Reliability for group 8

Group 8	Listening	Reading	Use of English	Speaking
Cronbach's Alpha	.82	.87	.87	.89

Each test featured certain questions that, if they were removed, would have a negative effect on the overall Cronbach's Alpha. These questions are therefore likely the most important for the reliability of the test. Conversely, it is possible that some items in the tests hinder the overall reliability. The following section will make note of the most important items for the reliability, as well as the items that would improve reliability if they were removed.

For Listening (Alpha=.82), one of the three items with the highest internal reliability tested picking up information from a monologue, whereas the other two items tested the proper way to reply to a query. The latter is an element that also features in oral proficiency and the Speaking test, in which pupils were also asked to reply to questions the examiners asked. Understanding queries and producing replies in a routine situation falls under the Overall Spoken Interaction header in the CEFR (Council of Europe, 2011), and these questions essentially tested this, thus testing pupils' knowledge of what they should produce. Though understanding what to say is no guarantee that they are able to actually say it, there is a link with language production. However, of the two items that negatively impacted the reliability, one was an item in the same overarching exercise that tested the query.

Additionally, the exercise containing these items came from a Proficiency-level test, which was the highest level that was tested. However, these questions, while ostensibly belonging to a test Anglia deemed equivalent to C1 on the CEFR, are unlikely to be of that CEFR level. A reply to a query first occurs at A2 level for Conversation in the CEFR, stating that pupils can “use simple everyday polite forms of greeting and address” and “make and respond to invitations, invitations [sic] and apologies.” (Council of Europe, 2011). Though it cannot be definitively said at which CEFR level the queries belong, it appears unlikely that they are of C1 level. Further supporting this is the PPON report by cito, which ventures, admittedly on the basis of limited data, that Dutch listeners do not attain B1 in listening at the age of 12 (Geurts and Hemker, 2013).

For Reading (Alpha=.87), the three items with the highest internal reliability were found in the first three texts of the test. The level from which the texts were taken gradually curved upwards, starting at Elementary for texts 1 and 2, through Pre-Intermediate for text 3 and Intermediate for texts 4 and 5, to Advanced for text 6. Conversely, there were six items that negatively impacted the reliability, though with minuscule adjustments should they be removed. All of these items were found in the last two texts.

For Use of English (Alpha=.87), the two items with the highest internal reliability dealt with constructing sentences: they were given a string of words that had to be put in the right order. These items came from an Elementary-level test. Another exercise, this one fill in the gap, was also taken from an Elementary-level test. Though pupils performed better on the fill in the gap exercise (M=.65) than the word order exercise (M=.47), the latter was overall more reliable than the former, meaning the questions in the exercise are a better judge of pupil skill. Additionally, the other two sentence construction exercises, one taken from a Primary-level test and one from a Pre-Intermediate level test, were not as instrumental for the overall reliability, though the average scores in these exercises were lower than those in fill in the gap exercises of a comparable level. The third most important item dealt with choosing the right verb form in an Intermediate-level fill in the blank exercise that also yielded two of the three items that negatively impacted the Cronbach's Alpha.

4.3.2 Reliability Group 5

For group 5, Listening and Use of English were acceptably reliable, though reading was less reliable. One question in the group 5 reading exam was removed from the analysis due to lack of variance.

Table 8: Reliability for group 5

Group 5	Listening	Reading	Use of English
Cronbach's Alpha	.84	.72	.80

For Listening (Alpha=.84), the most reliable items were from Primary or Preliminary-level tests, and all three were the same kind: choose the correct answer out of two. Unlike group 8, the exercise that asked pupils to reply to a query was not as important. Two out of the five items that hindered the reliability came from this exercise, including the least reliable item in the test.

For Use of English (Alpha=.80), the most reliable items dealt with creating sentences, as they did for group 8. Both of the exercises that these items belonged to were also in group 8's Use of English, and one of the items, taken from an Elementary-level test, was also one of the most reliable items for group 8's Use of English. The other two items were lifted from a Preliminary level test. Additionally, as seen in group 8, the average scores were lower in the sentence construction exercises than on other equivalent-level exercises. Three items had a negative impact on the overall reliability.

For Reading (Alpha=.72), two of the most reliable items came from exercises with two possibilities, though the most reliable item was from a text with three options per question. There are seven items that hinder the overall reliability.

In conclusion, the internal consistency of almost all the tests is sufficient, with the exception of reading for group 5. In addition, questions that dealt with active production, such as replying to a query and constructing sentences were generally more reliable for predicting the pupils' scores, though it cannot be said that these questions were solely responsible where possible.

4.4 Correlations

Pearson Product-Moment Correlation Coefficients were used in order to determine correlation between any combination of the four or five tests that were performed. A correlation coefficient, or *r,* of more than 0.5 or less than -0.5 "signifies a moderately strong relationship." (Lowie & Seton, 2013). The closer the *r*-value is to 1 or -1, the stronger the relationship is. A positive coefficient indicates that, as one score goes up, the other goes up as well.

4.4.1 Correlations group 8

Table 9:Inter-subject correlations for group 8

		Listening	Reading	Use of English	Spelling	Speaking
Listening	Pearson Correlation	1				
	N	558				
Reading	Pearson Correlation	.746**	1			
	N	556	560			
Use of English	Pearson Correlation	.715**	.768**	1		
	N	555	557	558		
Spelling	Pearson Correlation	.583**	.567**	.557**	1	
	N	555	558	557	559	
Speaking	Pearson Correlation	.760**	.696**	.754**	.547**	1
	N	127	128	128	128	128

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

Table 9 details inter-subject correlations for group 8 as a whole. All tests predict the score on other tests, to a certain degree, with at least moderately strong relationships that are all statistically significant (p<0.01). Listening shows the highest correlation with the other tests, though Use of English is approximately equally correlated, and Speaking and Reading are not far behind. The Spelling exercise has less of a correlation with the other tests, but this follows from the different source material and aim of the test. It did not seek to gauge overall pupil skill at a rough level, which was an implied factor for Listening, Use of English and Reading, and the goal of the Speaking test. Additionally, it was the only test that was not developed by Anglia Network Europe and Earlybird.

Table 10: Inter-subject correlations for Eibo and vvto

		Listening	Reading	Use of English	Spelling	Speaking
Listening	Eibo Correlation	1				
	vvto Correlation	1				
	Z values	1				
Reading	Eibo Correlation	.726**	1			
	vvto Correlation	.751**	1			
	Z values	-0.65	1			
Use of English	Eibo Correlation	.616**	.700**	1		
	vvto Correlation	.797**	.821**	1		
	Z values	-4.35**	1.65	1		
Spelling	Eibo Correlation	.574**	.563**	.497**	1	
	vvto Correlation	.572**	.550**	.594**	1	
	Z values	0.03	0.22	-1.95	1	
Speaking	Eibo Correlation	.771**	.633**	.709**	.528**	1
	vvto Correlation	.684**	.667**	.710**	.473**	1
	Z values	1.02	-0.33	-0.01	0.41	1

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

The greatest difference between the Eibo and the vvto group is the correlation between Use of English and all the other written tests, which is higher for vvto. There are pairs that have higher correlations for Eibo, such as Listening-Speaking, and, to a lesser extent, Spelling-Speaking, but the difference in the *r-value* is less pronounced than it is for the vvto Use of English correlations. However, applying Fisher's r to z transformation to the data reveals that the only set of correlations that are significantly different between different groups of participants is Listening and Use of English[12]. A negative Z value indicates that the correlation was higher for the vvto-group.

[12] Lowry's Vassarstats website was used to calulate these numbers: see http://vassarstats.net/rdiff.html

Table 11: r^2 for group 8 with Spelling exercise

r^2	Listening	Reading	Use of English	Spelling	Speaking
Group 8	.491	.482	.488	.318	.475
Eibo	.451	.430	.398	.292	.436
vvto	.491	.486	.535	.299	.401

When taking the Spelling exercise into account, only the vvto Use of English test can be designated as a key predictor. The r^2 of the Spelling exercises are a great deal lower than the r^2 of the tests made specifically for the study, indicating that the skills tested in the Anglia-made tests might differ from the skills required for the Spelling exercise.

Table 12: r^2 for group 8 without Spelling exercise

r^2	Listening	Reading	Use of English	Speaking
Group 8	.548	.543	.556	.543
Eibo	.496	.471	.456	.496
vvto	.554	.557	.602	.472

Removing the Spelling exercise from the calculations raises the r^2 of all the tests, showing that the scores of every amalgamated test taken by the whole of group 8 can be mostly attributed to the other test scores, rather than other factors, whilst a majority of the results for Eibo stem from other variables. However, for Eibo Listening and Speaking, r^2 is nearly 0.5, meaning it is likely these are good indicators as well, albeit not as good as other tests.

4.4.2 Correlations group 5

Table 13: Inter-subject correlations for group 5

		Listening	Reading	Use of English	Speaking
Listening	Group 5 Correlation	1			
	Group 8 Correlation	1			
	Sig. (2-tailed)				
	N	315			
Reading	Group 5 Correlation	.620**	1		
	Group 8 Correlation	.746**	1		
	Sig. (2-tailed)	.000			
	N	315	317		
Use of English	Group 5 Correlation	.698**	.654**	1	
	Group 8 Correlation	.715**	.768**	1	
	Sig. (2-tailed)	.000	.000		
	N	314	315	315	
Speaking	Group 5 Correlation	.540**	.304*	.476**	1
	Group 8 Correlation	.760**	.696**	.754**	1
	Sig. (2-tailed)	.000	.018	.000	
	N	60	60	60	60

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

*. Correlation is significant at the 0.05 level (2-tailed).

All of group 5's correlations are lower than the group 8 equivalent correlations. The lower correlations for Reading are somewhat understandable, as that test was less reliable than all other tests administered. Multiple correlations fall below 0.5, and one is only significant at the 0.05 level, rather than the 0.01level of all other correlations.

Table 14: r^2 for group 5

r^2	Listening	Reading	Use of English	Speaking
Group 5	.384	.277	.371	.194

Listening and Use of English serve as the best predictors of pupil skill for group 5, though, again, these scores are lower than their group 8 equivalents. Speaking in particular is not a good predictor at all. This is possibly due to the fact that speaking, listening and participating are the key tenets of Earlybird's early curriculum (Earlybird Methodiek, 2015), and of those, it is unlikely that formalised testing features in the listening part. Moreover, the Speaking examinations were aimed at eliciting speech from the pupils, and the exam was not as formal as the other tests. It is possible that these two factors influenced the apparent disconnect between the different skills.

5. Discussion and Conclusion

5.1 Main thesis question

The overarching research project set out to research the effectiveness of vvto as compared to regular Eibo, as well as to quantify any advantages vvto pupils might have had. This thesis sought to investigate the interconnectedness of the different skills that were tested, and to answer the following question: *Is there any one test utilised in this study that can serve as a key predictor for other tests?* Based on the correlation analysis carried out, it can be concluded that there is one such test for one of the three distinct groups that were tested: Use of English for the group 8 vvto participants. For the other groups, too much of the results was left up to other factors. If van Berkel's Spelling exercise was not included in the analysis, all remaining tests serve as key predictors for the whole of group 8, and Listening and Reading joined Use of English for the vvto group.

5.2 Reliability

It was required to find out whether or not the written tests that were specifically amalgamated for this study were actually reliable enough. All the written tests, with the exception of Reading for group 5, were sufficiently reliable as per the Cronbach's Alpha. There are several possibilities as to why that particular test was not sufficiently reliable. Firstly, it might be due to how the Earlybird programme is set up, as Earlybird’s curriculum starts with the spoken skills of listening and speaking at the age of 4. Though pupils do start learning to read and write Dutch in group 3, they only do so in group 4 for English in the Earlybird programme. Use of English primarily tests knowledge of grammar and vocabulary, and as such, there is limited influence of the little experience with writing in English for those pupils. As the reading tests did ask for some longer and relatively in-depth reading, it could be possible that the pupils did not understand either the questions or the texts themselves. As such, pupils might have reverted to guessing, which decreases the internal consistency of the tests.

Another possible explanation comes from orthography. English has a very opaque orthography, where phonemes and letters are often inconsistently matched. If a language has an opaque orthography, pupils cannot simply extrapolate the knowledge of how a word sounds to what a word looks like. By comparison, Dutch has “an intermediate orthography” (Borgwaldt, Hellwig, & de Groot, 2005), making it relatively easier for pupils to transfer their knowledge of how a word sounds to how it is spelled. Pupils might have known the word, and its meaning and use, but not its spelling, and thus, they could not utilise their knowledge, thus lowering the chances of a knowledgeable pupil scoring appropriately high.

A third possible explanation is that several questions in the reading exam were not properly understood by the children, despite using a similar structure as other questions in the same exercise. Evidence supporting this can be found in the first exercise of the test. The exercise used was from a First Step-level test, asking pupils whether a statement about a picture was true or false. Most of the items in this exercise were answered correctly around 95 % of the time, yet one of the items only had a mean score of .72, which is at odds with the expectations raised by the other items in the same exercise. One possible explanation for this is that the question was not understood properly. The question asked whether a cracked egg with yolk running out of it was an egg, and the picture might not have matched with some pupils' internal concept of an egg.

A reliability analysis reveals that there are indeed several items that negatively impact Cronbach's Alpha. The items that hinder are primarily questions that are either at the lowest level, i.e. First Step, that was tested or at the highest level that was tested. The presence of the easiest text suggests that the third explanation is plausible, yet the presence of the higher level text suggests that developmental or curricular causes for this cannot be ruled out. Removal of any one of the hindering items gives an increase in the value of anywhere between .002 and .013. However, the easier text influences the total reliability more, and as such, it appears more plausible that some of the questions were not properly understood. However, there is no conclusive evidence supporting one explanation over the others and it is probable that it was a combination of factors that caused the Cronbach's Alpha value to drop below .8.

5.3 Interconnectedness

Overall, all skills tested were highly interconnected for group 8, though with different accents for the vvto and non-vvto groups. If van Berkel's Spelling was taken into account, no real predictor could be found for the non-vvto group and the whole of group 8, with all the amalgamated tests made for this study scoring approximately equally and Spelling being less of a predictor. However, for the Earlybird pupils, it appears likely that the Use of English test is the most accurate predictor, in particular for Listening and Reading. The correlations for group 5 were weaker than those of group 8, even after removing the less reliable Reading test.

5.4 Individual tests

For Listening, there is a comparable relationship between Listening and Reading to what to what Bozorgian (2012) found: .735 for Bozorgian versus .746 for group 8 in this study. However, a much stronger relationship between Listening and production-related tests is present. Though Use of English focuses primarily on form, as opposed to writing on the IELTS, there is a stronger correlation between Listening and Speaking: Bozorgian found a .654 correlation coefficient (2012) to the .76 that was found in this study. It is possible that the IELTS listening exams are only listening comprehension, whereas the inclusion of some questions that ask pupils to reply to a statement in this study might have been of influence.

Use of English was consistently one of the better predictors in the study, especially so for the group 8 Earlybird pupils. Even so, there are differences between vvto and non-vvto pupils: the former group's Use of English and Listening scores have a stronger relationship relative to the other relationships. It is possible that the extra input gives pupils a better notion of what is grammatically correct, which is what a majority of the Use of English exercises tested.

Though oral proficiency plays a crucial role for reading comprehension (Mills, 2009), it is hard to say that the two skills have a very strong relationship in this study, even if there are statistically significant correlations. The weak relationship between Reading and Speaking for group 5 is of particular note here: group 5 has comparatively little experience with reading in English, yet far more experience with speaking it. Though this experience does not necessarily translate to high scores on the Speaking component of the study, the weak relationship and comparatively low p-value ($p<0.05$) run somewhat counter to Mills' claims. However, the group 5 Reading test was less reliable than the other tests as well, being the only test that was not deemed reliable enough (Alpha=.718), which may have influenced the correlations as well.

The correlations for Spelling are weaker than those found by Stæhr regarding vocabulary size and reading comprehension. Though the Spelling exercise did not test the size of pupils' vocabulary, the size can be a factor in knowing the word that was asked for. Overall, these correlations were also weaker than the other correlations, likely due to the different origin of the exercise.

Though Speaking was not the strongest predictor overall, it did have a strong correlation with Listening for two groups: group 8 Eibo and group 8 as a whole. The vvto correlations for Speaking and Listening were weaker, even after taking the overall lower correlations for the group 5 results into account.

One reason for the relative strength of some of the Use of English relationships might lie in the Earlybird curriculum, which provides vvto-pupils with a great deal more input compared to their Eibo counterparts, and vvto schools mostly utilised either a teacher specialising in teaching English (vakleerkracht) or a native speaker for teaching English (de Haan, 2014). However, cito (Geurks and Hemker, 2013, p. 48) reported that 85 % of the teachers responsible for teaching English, for both Eibo and vvto, were the regular teachers (groepsleerkracht), making the prevalence of the subject-specific teachers or the native speakers in the Earlybird schools a slight anomaly, though it should be noted that only 6 vvto schools were surveyed. Even so, it appears likely that the vvto schools provided pupils with more input and with better input, as regular teachers usually are less proficient in English than vakleerkrachten or native speakers (Thijs, Trimbos, Tuin, Bodde, & de Graaff, 2011). This better input might have led to pupils being better at choosing the correct form, which was a part of both Listening and Use of English. However, this is at odds with the lower correlation between Speaking and Use of English, and even though the relationship between Spelling and Use of English is stronger than in the other two groups, it is still not as strong a relationship as other subskills have.

Based on the available data, Use of English appears to be a key predictor, as its correlations and resulting r^2 values are highest. Furthermore, questions that relate directly to production appear to be important, as gleaned from the Cronbach's Alpha analysis of Use of English for both groups, as well as Listening for group 8. More precisely, production that can serve an immediate, concrete, purpose, is relevant. It follows that Speaking should be of some importance as well, but the results do not support that conclusion outright.

5.5 Recommendations for testing practice

The findings of this study indicate that, if teachers of English want to assess the level of their pupils, they should do so using exercises designed to test production that has the potential to be of immediate relevance for the pupils. Given the status of Use of English as a key predictor, it follows that this be done through writing, but utilising short queries of how to reply, as used in the Listening exercises, should likely be included as well in some form. The goal is to let pupils produce in natural situations. For Use of English, Listening, and Speaking, these situations should be fairly easy to set up.

However, any focus on Use of English should not be to the point of losing sight of the other skills. Teaching languages requires a holistic approach, and its testing should reflect that. If Use of English becomes disproportionately influential in testing practice, then a real risk of teaching to the test occurs. At that point, it becomes unclear whether Use of English remains a valid predictor.

5.6 Limitations

One of the strengths, yet also one of the limitations, of this study is that many of the tests were specifically tailored for this study. The strength is that the study is likely to be more accurate, since there are no extraneous factors involved, yet the limitation therein is that it does not connect to any international tests done in foreign language education. IELTS, TOEFL, Cambridge et al. use different metrics to measure each skill, making it hard to translate a successful approach here (i.e. focus on Use of English) to a successful approach towards Cambridge exams.

A second limitation is present in the analyses. Though it was not originally in the scope of the study, an analysis of the questions that overlapped between group 5 and group 8 could have been relevant as an indication of how much pupils gain in proficiency in those three years.

Thirdly, the relative lack of intercorrelatedness (r^2) for the Eibo group 8 tests might hinder the overall validity of the study. If van Berkel's exercise is not taken into account, only Eibo lacked key predictors. As Eibo is still the standard and more prevalent mode of teaching English in primary schools in the Netherlands, the results of this study might be less relevant for Dutch primary education as a whole.

5.7 Further Research

One potential avenue of further research regarding the differences between Eibo pupils and vvto pupils would be a longitudinal study with tests in group 8 and after the second year of Dutch secondary education. This could ascertain whether vvto has a lasting advantage, but additionally, if similar tests are used, it might prove possible to be more definitive on which tests and associated skills can serve as key predictors.

Another possibility, if more comparisons between vvto and Eibo are required, is to let group 8 pupils, from both groups, write something. The results suggest that production in any form was fairly important for reliability, but pupils writing something only occurred in the Use of English exam, in a semi-closed format. A writing test could quantify the importance of production, allowing for more accurate teaching.

In an increasingly global world, teaching children to utilise more than just their native tongue is more than just beneficial. Knowing another language broadens the mind; reveals paths unseen; opens doors locked to the uninitiated. Giving learners a boost in learning about a language is essential for making sure that they will be able to use the knowledge to the best of their ability. In this, teachers and researchers are no different than pupils: we all seek to find the key that will help us unlock our full potential.

References

Aarnoutse, C. & van Leeuwe, J. (1998). Relation Between Reading Comprehension, Vocabulary, Reading Pleasure, and Reading Frequency, Educational Research and Evaluation. *An International Journal on Theory and Practice 4*(2), 143-166

Wächter, B, & Maiworm, F. (2008). *English-Taught Programmes in European Higher Education.* Bonn: Academic Cooperation Association.

Anglia Network Europe (2013). *Exams*. Retrieved from http://anglianetwork.eu/index.php/examinations.html

Berkel, A. van, Philipsen, K., & Feuerstake, M. (2013). *Spellingtoets Engels voor de groepen*

7 en 8 van vvtoE-scholen. Rotterdam: Early Bird en Europees Platform.

Borgwaldt, S. R.; Hellwig, F. M. and de Groot, A, M.B (2005). Onset entropy matters – Letter-to-phoneme mappings in seven languages. *Reading and Writing,18,* 211-229.

Bozorgian, H, (2012). The Relationship between Listening and Other Language Skills in International English Language Testing System. *Theory and Practice in Language Studies 2*(4), 657-663.

Centraal Bureau voor de Statistiek (2012). *Jaarboek Onderwijs in Cijfers.* Den Haag/Heerlen.

Geurts, B, & Hemker, B,. *Balans van het Engels aan het einde van de basisschool 4.* Uitkomsten van de vierde peiling in 2012. Arnhem: CITO..

Council of Europe (2011). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Structured overview of all CEFR scales.

Crystal, D. (2003). *The Cambridge Encyclopedia of The English Language.* Cambridge: Cambridge University Press.

EarlyBird (2013). *Methodiek.* Retrieved from http://earlybirdie.nl/index.php?page=Over_EarlyBird-Methodiek&pid=173

Europees Platform (2014). *TTO Scholen.* Retrieved from http://www.europeesplatform.nl/tto/scholen/ .

Feyten, C. M. (1991). The Power of Listening Ability: An Overlooked Dimension in Language Acquisition. *The Modern Language Journal 75*(2), 173-180.

de Haan, R. (2013). *Testing the Winds after 10 Years of EarlyBird Elementary Foreign Language Education* (unpublished MA thesis). Utrecht University, Utrecht, the Netherlands. Retrieved from Igitur.

Hinkel, E. (2006). Current Perspectives on Teaching the Four Skills. *TESOL Quarterly, 40*(1), 109-131.

Holdinga, L. (2007). *Van Engels in het basisonderwijs naar Engels in het voortgezet onderwijs* (unpublished MA thesis). Utrecht University, Utrecht, the Netherlands. Retrieved from Igitur.

Kenny, D. A. (1987). *Statistics for the Social and Behavioral Sciences.* Little, Brown and Company (Canada) Limited. Found at http://davidakenny.net/books.htm

Mills, K. (2009). Floating on a Sea of Talk: Reading Comprehension Through Speaking and Listening. *The Reading Teacher, 63*(4), 325-329.

Lowie, W. & Seton, B. (2013). *Essential Statistics for Applied Linguistics*. Chippenham and Eastbourne: CPI Antony Rowe.

Muñoz, C.(2014). Contrasting Effects of Starting Age and Input on the Oral Performance of Foreign Language Learners. *Applied Linguistics 35(*4), 463-482.

Nation, K. & Snowling, M.J. (2004). Beyond phonological skills: broader language skills contribute to the development of reading. *Journal of Research in Reading 27*(4), 342-356.

Onderwijsraad (2011). *Weloverwogen gebruik van Engels in het hoger onderwijs*. Den Haag.

Qian, D. D. (2002). Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective. *Language Learning 52*(3), 513-536.

Rijksoverheid (2015). *Wat houdt de pilot tweetalig primair onderwijs (tpo) in?* Retrieved from http://www.rijksoverheid.nl/onderwerpen/basisonderwijs/vraag-en-antwoord/wat-houdt-de-pilot-tweetalig-primair-onderwijs-tpo-in.html

Rijlaarsdam, G. (2005). Observerend Leren: Een kernactiviteit in taalvaardigheidsonderwijs. *Levende Talen Tijdschrift, 6*(4), 10-28.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal 36*(2), 139-152

Saville-Troike, M. (2006). *Introducing Second Language Acquisition.* Cambridge: Cambridge University Press.

Thijs, A.; Trimbos, B.; Tuin, D.; Bodde, M. and de Graaff, R. (2011). *Engels in het Basisonderwijs: Vakdossier.*