BA Thesis English Language and Culture, Utrecht University

A Study into the Effects of Sentence Context in English Grammar Tests in Dutch Secondary Schools

Stein Voskuil

3468046

Supervisors:

Roselinde Supheert & Merel Keijzer

June 2013

Acknowledgements

I would like to thank Hanneke Banine of the Van Maerlant Lyceum for allowing me to test her class' students and for helping me with the correction and marking of the test. Gretha Glasmacher of the Sint Joris College is also gratefully acknowledged. Even though in hindsight, the results I obtained from her students were unusable, she welcomed me very warmly and wholeheartedly to conduct this experiment. Additionally, I would like to thank Roselinde Supheert and Merel Keijzer of the University Utrecht for their helpful comments and advices on earlier drafts of this paper. Their guidance and their feedback have been of unprecedented value to me. I hope that their directions have resulted in a greatly improved paper.

Table of Contents

Acknowledgements 2

Table of Contents 3

Chapter 1: Introduction 4

Chapter 2: Theoretical Framework 7

2.1 Teaching Methods 7

2.2 Washback 10

2.3 Language Testing 11

Chapter 3: Purpose of this Study 14

3.1 Research Question 14

3.2 Hypothesis 14

Chapter 4: Method 16

4.1 Subjects 16

4.2 Test and Questionnaire 17

4.3 Procedure 19

Chapter 5: Results 20

5.1 Test 20

5.2 Questionnaire 22

Chapter 6: Discussion 23

Chapter 7: Conclusion 26

Works Cited 28

Appendix 1 30

Appendix 2 32

Appendix 3 34

Appendix 4 35

Chapter 1: Introduction

From the moment governments adopted national curriculums for their schools, or perhaps well before that time, teachers, researchers and policy makers have taken a strong interest in discovering how the art of teaching could be improved (Woods 3-4). So much has been written on how best to teach a subject that it has become tiring to book reviewers like Linda Scott, who opens her review of Jim Scrivener's *Teaching English Grammar: What to Teach and How to Teach It* exclaiming: "Oh no. Not yet another book telling us how we should teach grammar, I can hear you say? So many books have been published over the last 20 years on the subject that we might reasonably conclude that all that could be said about it has already been said" (1). This conclusion, however, has not taken the inevitability of language change into account. As historical linguist Robert McColl Millar states: "language change is ceaseless and remorseless. Every language that is spoken continues to change, not century by century, but day by day" (14). Teaching methods and research on how best to teach should always strive to keep up with this change and, therefore, no one can ever "conclude that all that could be said about it has already been said" (Scott 1). With regard to language change and language teaching, Glenn Fulcher and Fred Davidson used architecture as a metaphor for language test development and stated: "both buildings and tests grow and change over time as the needs of their users change. Sometimes, they are both used for purposes other than those intended in the original designs" (123).

However, central to the discipline of English language teaching (ELT) remains the interaction between students and teachers (Celce-Murcia and Hilles 7-8). A considerable amount of literature has been published on ELT, ESL (English as a second language) and EFL (English as a foreign language). These studies tend to focus primarily on teaching methods and on how to involve students in the teaching process, as is done by Jo McDonough and

Christopher Shaw’s *Materials and Methods in ELT: A Teacher’s Guide* or Marianne Celce-Murcia and Sharon Hilles’ *Techniques and Resources in Teaching Grammar*, who suggested that a more involving teaching method helps teachers to convey their message to the students (51).

It has also been suggested that language testing has a significantly great impact on language teaching and learning. A phenomenon called *washback* (see chapter 2.2) is used to describe this influence, and extensive research has been done into this corner of ELT (Taylor; Cheng, Watanabe and Curtis). Although Liying Cheng, Yoshinori Watanabe and Andy Curtis have found both positive and negative effects of language testing on language teaching and learning, they could not find reasons for the occurrence of these effects (7-11). Taking *washback* into account, the quest for the discovery of the best teaching methods could also begin with the improvement of language testing.

The objective of this study is to examine the effect on the test scores of a grammar test which is presented with sentences that try to involve the students and the students’ interests, as opposed to a grammar test with general sentences that are not designed to involve students. Celce-Murcia and Hilles claim “teaching points” are more easily remembered by students when they are included in a storytelling process, which invigorates the language learning process (51-52). This study will examine if an inclusion of students’ interests in English grammar tests has any effect on the test scores. Several studies have produced findings concerning this relationship between test scores and test design (Jafarpur; Perkins; Clapham; Freedle and Kostin; Xie and Andrews), but there is still insufficient data for grammar tests specifically. The article by Abdoljahad Jafarpur, and that of Caroline Clapham, as well as that of Kyle Perkins, have focussed on reading comprehension tests, whereas Roy Freedle and Irene Kostin focussed on so called “minitalk” (2) assignments. Additionally, Qin Xie and Stephen Andrews suggest that a test taker’s perception of the test is also a factor which can affect the test scores and state: "the more test takers agree that they need to use language skills in order to answer test items correctly, the more likely it is that they will endorse the test design" (54).

Researchers into ELT have analysed many different language tests, grammar tests included. However, most ELT researchers have focussed on ESL, that is to say foreign students learning English as a second language in a predominantly English-speaking country. In contrast, this study tries to extend the coverage on EFL, particularly with regard to Dutch secondary school students following English lessons on a mandatory basis.

The above-mentioned studies will be discussed in the following chapter and will serve as the foundation on which this current experimental study bases its assumptions. These assumptions will be presented in chapter 3, which includes the following research question:

> *Does altered sentence context intended to involve students in a grammar test influence the test scores?*

The hypothesis of this study is also addressed in chapter 3. In the method section, chapter 4, the experiment conducted for this study and the procedures that were followed will be explained. The results of this experiment will be presented in chapter 5 and discussed and interpreted in chapter 6. The final chapter, chapter 7, will give an overview of this paper's main findings and conclusion and will provide suggestions for further research.

Chapter 2: Theoretical Framework

During the past 25 years, much more information has become available on ELT and applied linguistics in general. A large and growing body of literature has investigated different teaching methods and techniques. Language testing has also been academically scrutinised over the last two decades. In recent years, the amount of literature on *washback* has been increasing rapidly. The term is used to describe the connection between, on the one hand, language testing and on the other hand, language teaching and learning. *Washback* denotes either a positive or negative influence (Cheng, Watanabe and Curtis 8). Some studies that might help to establish a framework on which this study will base its assumptions are discussed below.

2.1 Teaching Methods

Through the years, optimal ways to teach English have been sought by many researchers in the ELT field. Among those researchers, however, there is still some disagreement on what the best teaching methods are and how these methods should be analysed and validated. The high number of variables that can play a role in teaching methods make it difficult for researchers to account for all of these variables, draw conclusions and assign ELT pitfalls to a single cause.

Devon Woods has accounted for several of these difficulties that might also play a role in this paper. In his dissertation *Processes in ESL Teaching: A Study of the Role of Planning and Interpretive Processes in the Practice of Teaching English as a Second Language*, an extensive introduction is dedicated to the question: why study the teacher? (1-21). In it, Woods states: “Through the years, decades, and even centuries according to Kelly (1969) and Howatt (1984) in the literature on teaching a second language, there has been a large number of articles, papers and theses devoted to comparing approaches to and methods for teaching a second language" (3). Many of these studies have been criticised by researchers who questioned the validity of their conclusions. Importantly, Woods ascribes this criticism to a common difficulty which all these ESL teachers and researchers have to cope with: "The number of interrelated variables makes it extremely difficult to attribute the results to the method variables in question. None of the findings of these studies have been accepted in the field as unambiguous" (5).

Woods also emphasises the necessity of carefully and thoughtfully defining parameters, for instance, because it is both incorrect and confusing to assume that everyone has the same idea of what language is. However, he does suggest that previous ELT research has yielded useful results and contributed to the understanding of classroom teaching and the learning of English as a second language. Nevertheless, Woods recognises "three important gaps which are relevant to the theory and practice of ESL teaching" (11). Most importantly, the third gap that Woods discusses is the gap in knowledge of the perception and interpretation of language tests by participants themselves, a topic which the present study will also address. Woods claims far too little research has been done into what the teachers and learners think is happening inside the classroom: "This perspective, which may be quite different from what an outside analyst concludes is happening, has not been taken into account in any systematic way in research related to second language teaching" (13). Moreover, Woods emphasises that the way a course is given sprouts from a mixture of instructions and procedures provided by the institution, the teachers and the materials that represent the curriculum (204). Woods' dissertation is relevant to this study because it highlights the many difficulties researchers may come across when assessing the effectiveness of certain teaching methods, as it takes a laborious effort to account for all the different variables that can play a role. However, Woods notes, any study in the field will serve the understanding of language teaching and learning as a whole.

One of the studies contributing to the understanding of language teaching is *Techniques and Resources in Teaching Grammar* (1987) by Celce-Murcia and Hilles, in which they make claims that are essential for the construction of this paper's hypothesis. For instance, Celce-Murcia and Hilles stress the need to be enthusiastic as a teacher: "Grammar points can be contextualized in stories that are absorbing and just plain fun if they are selected with the interest of the class in mind and are told with a high degree of energy, and involve the students" (51-52). This enthusiasm can help to get the grammar across:

> Students will certainly appreciate and respond to your efforts to include them in the storytelling process, but they will also, we have found, enjoy learning about you through stories. […] [A]necdotes about you, your family, or you friends, as long as they are relevant and used in moderation, can be very effective. Stories should last from one to five minutes, and the more exaggerated and bizarre they are, the more likely students will remember the teaching points they illustrate. (52)

Lynda Taylor espouses Woods' theory on perception and interpretation by stating that "we should not ignore or override the attitudes and perceptions of learners themselves, or indeed those of the many teachers worldwide whose own English proficiency is based upon exposure to a particular native-speaker model" (52). However, where Celce-Murcia and Hilles, and Woods focus on teaching and learning methods, Taylor investigates the correlation between, on the one hand, language testing and, on the other hand, language teaching and learning. She says that "it is commonly assumed that tests exert a powerful influence on what happens in the classroom. […] The traditional view is that a simple linear relationship exists between teaching and testing, i.e. change the test and changes in teaching and learning will automatically follow" (54). She mitigates the simplicity of the relationship by stating that more recent research has revealed the complexity of this phenomenon called *washback*.

2.2 Washback

In addition to Taylor's paper, *washback* has been thoroughly examined by many others (Cheng, Wantanabe and Curtis; Xie and Andrews; Bailey). The effects of language tests on language teaching and learning were hypothesised as early as 1956 (Vernon) but the existence of this correlation was not proven until 1993 in a study by Charles J. Alderson and Dianne Wall, which investigated the influence of language assessment on language teaching in four different countries (123-27). Since then, much more information has become available on *washback*. Cheng, Wantanabe and Curtis have found that *washback* can have either positive or negative effects but, like Woods, they stressed the difficulty of drawing clear-cut conclusions in ELT. They explain that "[w]hether the washback effect is positive or negative will largely depend on where and how it exists and manifests itself within a particular educational context" (11).

Even more recently (2012), Xie and Andrews declared "it is still not clear precisely how testing influences teaching and learning" (50). However, they did find that the students' perception of an assigned assessment influences the result. If the students consider language skills more important to pass the test than other skills, they are more likely to support the test format (54). Although this discovery might also affect the present study, Xie and Andrews stress that their research only investigated "high-stakes, instrumental test uses" (54), which this study does not.

Nevertheless, the current study should not consider any notion of *washback* to be irrelevant or unimportant as the effects of low-stake tests have never been ruled out. Since the students in the present study were informed their test was part of a study beforehand, their attitudes towards it may have changed, which may have influenced the test scores, as Xie and Andrews pointed out.

2.3 Language Testing

With regard to *washback,* language testing itself has also been subject to a vast number of studies (Fulcher and Davidson; Freedle and Kostin; Jafarpur; Perkins; Clapham; Bachman and Palmer). Moreover, the scholarly journal *Language Testing* is entirely devoted to the subject. However, most of the studies suffer from serious weaknesses, as they only focussed, and conducted experiments, on any specific language test type but asserted the conclusion drawn from those studies might be applied to a much larger area of the ELT field. For instance, Qin Xie tried to assess if perceived test-taking skills are related to the target language skills (382) but did so by examining only two different test methods (342). On the other hand, by increasing the coverage of language tests being investigated, these studies do contribute to the understanding of language assessment in some way. Therefore, the findings which are most important and relevant to this study will be discussed here.

For instance, Lyle F. Bachman and Adrian S. Palmer have stressed the importance of using guidelines or blueprints when constructing a language test. According to them, these guidelines should describe “how actual test tasks are to be constructed, and how these tasks are to be arranged to form the test” (90). In addition to this concept of language test writing, Alan Davies has suggested that it does not matter whether writing the test is done by one or more authors (12-13).

The latter point has been devastatingly criticised by Jafarpur, who ordered six experienced teachers to create a reading comprehension test on the same subject (61-64). Although the teachers all had to abide by the same rules, Jafarpur found that there is “inadvertent variation in the kinds of items constructed by different test-developers” (70).

Jafarpur stresses that his conclusion is merely “suggestive” (71) but he does point out that the results of his research could have been less divergent “if the item constructors involved in the present investigation had been given clear specifications as well as adequate feedback on their work” (71). He adds recommendations such as peer reviewing and checks to ensure that the questions to a text are not created on the basis of the interpretation of one individual (73). Jafarpur’s study becomes particularly relevant to the current paper when he includes the notion that “the very low-stakes settings where specifications may not have to be very detailed or formalized,” should also embrace stricter test development rules (73).

Whereas Jafarpur focussed on the test constructor, others shifted their attention to the content of the test itself. For instance, different authors have measured the grammatical ability of students in a large variety of ways. These include the Modern Language Aptitude Test (Carroll et. al.), the Test for the Reception of Grammar (Bishop) and many more. In addition, the influence of context material on test scores has also been an area of interest to ELT researchers like Perkins, who found that the topical structure of ESL reading comprehension tests, or the way the topics of the text are relating to each other, could affect the difficulty, and therefore the students’ performance (164). Furthermore, when dealing with reading comprehension tests, Clapham found that:

> When the modules included *general* passages, the level of language proficiency had markedly more effect on the students’ scores than did background knowledge. However, once the modules contained only *specific* passages, background knowledge became proportionally more important. It might be hypothesized that, if all the subtests had been ‘highly specific’, background knowledge might have been made an equal or greater contribution to comprehension than language ability. (205)

This notion is in accordance with Xie and Andrews in that context can affect the validity of a test. There may be a difference between what the test claims to be assessing and what it actually does. Clapham's findings might also influence the hypothesis of this study, as they show the downsides of the use of more suggestive, involving and perhaps embellished language testing material, as opposed to objective material. Where Celce-Murcia and Hilles might support the use of a different teaching and testing method by advocating more enthusiastic and exaggerated teaching, Clapham's results seem to question the advantages of such a change.

Moreover, Freedle and Kostin found that "pure item variables play a minor role in determining […] item difficulty" (21). By examining the results of a so-called TOEFL (Teaching of English as a Foreign Language) minitalk test, a test that assesses both listening and reading comprehension skills, the researchers found that "pure item variables appear to play a minor role, while text and text associated (text/item overlap) variables play by far the major role in accounting for minitalk item difficulty" (22), by which they mean to say that the topics and context of a text can contribute tremendously to the difficulty of a test. These findings concur with those of Clapman and those of Jafarpur. They might therefore be extended to form a hypothesis on the outcome of the current study.

Chapter 3: Purpose of this Study

This study aims to enhance the overall understanding of ELT and in particular that of English language testing. The primary goal of this paper is to examine the differences between the test scores of, on the one hand, a test where the context of the sentences is adapted to the interests and experiences of Dutch secondary school students, and on the other hand an unaltered test where the context of the sentences is more neutral.

3.1 Research Question

The present study seeks to address to following question:

> *Does altered sentence context intended to involve students in a grammar test influence the test scores?*

Before this research question can yield an answer, opaqueness about which sentence context involves students in a grammar test, needs to be clarified. Moreover, it needs to be stressed that the findings of this study may only be applicable to the test format used, which is a cloze test or cloze deletion test. Woods has found that it is rather optimistic to suggest that findings of one particular test format can be applied to other test formats as well (10). Although this has to be kept in mind when formulating a hypothesis, the lack of research into cloze grammar tests necessitates generalisation of findings on other test formats.

3.2 Hypothesis

Some previous studies might suggest that altered sentence context intended to involve students will influence the test scores positively (Celce-Murcia and Hilles; Perkins), whereas other findings could lead to believe these alterations will have a negative influence on the test scores (Freedle and Kostin; Clapham; Jafarpur). However, most studies are unable to answer the current research question one way or the other (Woods; Cheng, Watanabe and Curtis; Xie and Andrews; Taylor).

Nevertheless, there is some consensus among ELT researchers concerning this subject as all agree that the test scores will change if the test design is altered. Therefore, the following hypothesis has been formulated:

It is expected that altered sentence context intended to involve students in a grammar test influences the test scores. However, because the majority of previous studies focused on high-stake tests, the low-stake tests used in this study might not have a significant influence.

Whether this influence is either positive or negative seems to be harder to predict. However, where Celce-Murcia and Hilles showed involving students yields better results, they did so by changing the teaching method as well as the testing method. On the other hand, Freedle and Kostin, and Clapham examined what happened if students were presented with a non-standard test format. The teaching of these students was unaltered during the study.

As the current paper also deals with an unaltered teaching method and only a difference in the testing format, it may be more accurate to support the findings opposing the use of such test formats. Therefore, this paper hypothesises that:

An increase of student involving sentence context will lead to lower test scores.

Chapter 4: Method

By conducting an empirical case-control experiment among Dutch secondary school students, this study seeks to address an aspect of ELT which has been poorly investigated in the past. The design of the current test was based on the test format with which the students were most familiar. This chapter will discuss all of this study's methodology, including the grading and data processing. It will also be explained how the attached questionnaire was created and conducted.

4.1 Subjects

In total, 73 Dutch secondary school's so-called *havo-vwo brugklas* students participated in this study, which means all students were probably 12 or 13 years old, although actual ages were not recorded (an overview of the Dutch educational system can be found in appendix 4). 52 of these students attended Van Maerlant Lyceum (henceforth VML), Eindhoven, and the other 21 students attended Sint-Joriscollege (henceforth SJC), Eindhoven. The initial sample consisted of 43 male students and 30 female students distributed over three classes, two at VML and another at SJC.

Both versions of the test were distributed equally with 37 students being given version X and 36 being given version Y. The majority of the students were sitting in pairs, so to minimise the chance of cheating, one of the pair was handed version X, while the other was handed version Y. To ascertain that the class had been assigned the correct lessons and had in fact learnt them, a class' minimum overall score had to be at least of 2.0 out of 8 points.

4.2 Test and Questionnaire

As the purpose of this test was to identify differences between a standard Dutch secondary school grammar test and one with a different sentence context, two tests were created for the current study, One test with a different sentence context, test X, and one unaltered control group test, test Y, were created. This was done by using an older written exam, which had already been used, as a model. However, this model consisted of twenty words and five sentences and as the present research only focused on the sentences, a greater number of sentences was desirable. Therefore, the tests in the current study consisted of ten sentences and, to make the exam more familiar to the students, fifteen words were added, but these were to be excluded from the final score comparison. Tests X and Y can be found in Appendix 1 and 2 respectively.

As part A (the words) is irrelevant for this study, only part B (the sentences) will be discussed from here. The students were assigned to study Unit 8, lesson 36, 37 and 38, from the *New Interface Coursebook 1 (t)hv Blue Label* (Bosschaart et. al. 154-56). The words and sentences that the students were assigned to learn can be found in Appendix 5. *New Interface Coursebook 1 (t)hv Blue Label* is a course book which is extensively used throughout Dutch secondary schools and represents a conventional way in which English is currently taught in the Netherlands. Two of the sentences that were used were exactly the same for both test X (5 and 9) and test Y (3 and 8) and were therefore not included in comparative calculations.

As a result, eight sentence sets could be used for the comparison. From these sentences, three were chosen from lesson 36, three from lesson 37 and two from lesson 38. The sentences for test Y were copied directly from the course book. For test X, the context of these sentences was changed to involve the students. Sentences 1Y and 1X show that the blanks that were to be filled by the students, were the same for both tests, as the students were told only to translate the bold printed part of each sentence (the correct answer is given in italics).

1Y: **Daarom heb ik nou zo’n hekel aan** spelletjes.

(…1… games.) *That’s what I hate about games.*

1X: **Daarom heb ik nou zo’n hekel aan** Feyenoord.

(…2… Feyenoord.) *That’s what I hate about Feyenoord.*

As can be seen from these sentences, only the context of the sentence, the text which is not in bold, differed for the test two versions. In the test design with which the students were most familiar, however, the students had to translate the sentence, rather than completing them. As shown in the appendices, the sentences used for the tests were in a different order for each version. This was done to prevent cheating.

To establish which sentence context would be closest to the students’ experiences, a variety of methods has been used. First of all, search engine Google was used to choose from the top trending sought items both in the Netherlands and worldwide. This has yielded the sentence context concerning the Eurovision song festival, Harry Styles, Justin Bieber, Candy Crush, Messi and Twilight. Secondly, through informal enquiry among these students, it was found that regional differences were a recurrent topic of conversation. This resulted in the creation of sentence context about Aalst, Waalre and Feyenoord.

Along with the test, a questionnaire was devised to measure the students’ perception of both versions of the grammar test. This questionnaire, which can be found in Appendix 3, consisted of two questions. The students were asked whether they enjoyed the test more than the written tests with which they were already familiar. They had to indicate their liking on a 1 to 5 scale, where 3 denoted that the test was as much fun as the tests they were used to. For the second question, students had to express how much easier or more difficult they found the test compared to their customary tests. Once again, this was done on a 1 to 5 scale, where 1 denoted the current test was much easier and 5 denoted that it was much harder. Ideally, test Y would score 3 on both questions, as it was created to serve as a control test.

4.3 Procedure

Prior to the test, all students were told that the result would modestly contribute to their overall English grade. The students were asked to remove everything but a pen and a blank piece of paper from their tables before the tests were handed over. As stated above, both versions of the test were distributed equally among the subjects. In theory, each student was allowed 45 minutes to complete the test but the actual duration was estimated at roughly 15 minutes. The students were asked to fill in the questionnaire directly after they had finished their test. At VML, one class was tested by the researcher after the first afternoon break at approximately 11:25 while the other class was tested after the second afternoon break at 13:30. Ten days later, the researcher invigilated the test at SJC, also after the second afternoon break at 13:30.

After data collection was completed, the tests were graded. For each sentence, it was possible to be assigned 0, 0.5 or 1 error(s). A sentence was assigned a 0.5 error if only one word was spelled incorrectly. If multiple words were spelled incorrectly or the sentence structure was incorrect, it was assigned 1 error. Using the obtained data, a statistical analysis, using Microsoft Office Excel and SPSS, has been carried out to assess if there was a significant difference between versions X and Y. The significance of such a difference was established by the use of a two-tailed t-test if the result were to be $p \leq 0.05$. The null hypothesis in this case meant that version X and version Y would have no different mean score.

Chapter 5: Results

5.1 Test

After grading the tests, it became apparent that the results obtained at SJC could not be used for this study. It can be seen from the data in Table 1 that the SJC scores were substantially lower than the minimal mean score of 2.0 for both version X and Y. In fact, only two of the twenty-one students at SJC managed to score 2.0 points or higher. The cause of this outcome will be discussed in the next chapter of this paper. Therefore, all further calculations in this study were done without the use of scores from the SJC students.

	Mean score, version X	**Mean score, version Y**	**Overall mean score**
VML, Class 1	4.88	5.82	5.37
VML, Class 2	6.42	6.29	6.36
SJC, Class 1	0.05	0.60	0.31

Table 1: Mean scores for each class in this study.

As can also be seen from the table (above), only in VML, class 2, version X yielded higher scores than version Y. Overall, this class scored higher than VML, class 1. Turning now to the VML data, the average scores of version X and version Y were compared. First of all, version X and version Y made by class 1 were compared. A t-test for these data showed that there was no significant difference (p = 0.23). Secondly, the same was done for class 2, which showed that the difference between version X and Y in that class was even less significant (p = 0.78). Because both classes showed an insignificant difference, the results of these classes were treated as a single set of data. Table 2 shows the mean scores for both versions and the range of scores scored by all VML students.

	Mean score	**Range**
Version X	5.83	1.5 - 8.0
Version Y	6.38	3.5 - 8.0

Table 2: Mean score and ranges of version X and version Y.

A two-tailed t-test was used to analyse the difference between the test scores of version X and of version Y. The null hypothesis (the mean scores of version X and version Y are equal) would be rejected if the result of the t-test were to be $p \leq 0.05$. However, no significant difference was found between version X and Y as the t-test showed that $p = 0.18$.

Figure 1 provides a clearer view of the likeliness of the scores of version X and version Y. It presents the frequency of each possible score in the test, i.e. how many times a score of 6.0 points was scored and how many times a score of 6.5 points was scored and so on. The curved line illustrates the distribution of the frequency of the scores, which is an estimation of the frequency of the possible scores if the test were to be reproduced.

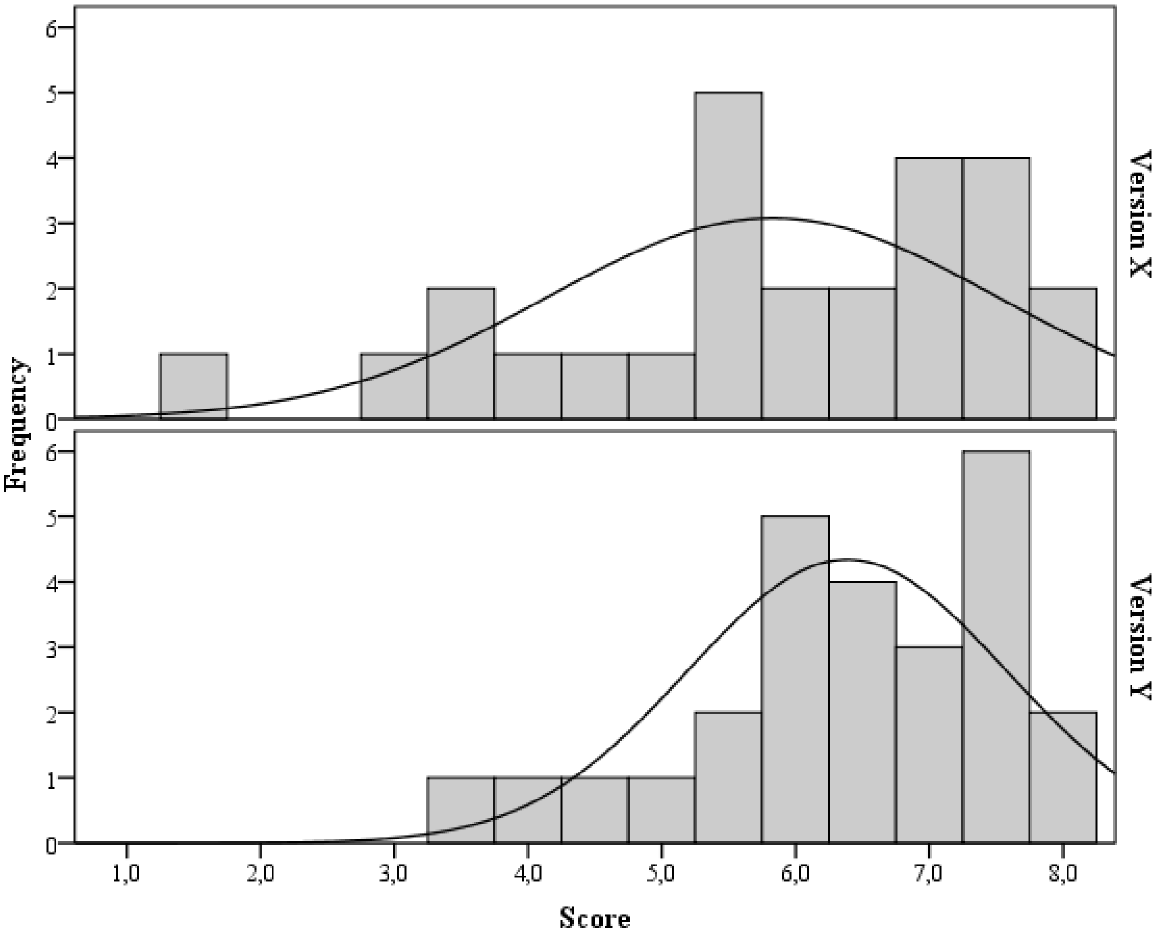


Figure 1: Score frequency with corresponding distribution.

All in all, Table 2 and Figure 1 show that the scores of version X were indeed lower than those of version Y but the difference was statistically insignificant.

5.2 Questionnaire

To assess the students' perception of the difficulty and enjoyment of the tests, a questionnaire was used (Appendix 3). A difficulty score of 1 meant the student found the test particularly easy, whereas a difficulty of 5 meant he or she found it extremely difficult. As regards to the enjoyment rate, a student with a score of 1 found the test really amusing but a student with a score of 5 found it totally uninteresting. Table 3 provides the mean scores of perceived difficulty and enjoyment for both test versions.

	Mean score, difficulty	**Mean score, enjoyment**
Version X	2.85	2.27
Version Y	2.42	2.46

Table 3: Mean scores of the questionnaire's results.

From this data, it can be seen that version X was perceived as more difficult but also more enjoyable. However, another couple of t-tests for these data sets pointed out that there were once again no significant differences between version X and version Y. The t-test comparing the difficulty of both test versions resulted to a *p*-value of 0.07. The enjoyment rates of both versions were found to be even more alike ($p = 0.36$).

Chapter 6: Discussion

This study aimed to assess the effect of sentence context in English grammar tests. By means of a case control experiment, an answer to the following research question was sought:

> *Does altered sentence context intended to involve students in a grammar test influence the test scores?*

The results show that the test scores of version Y, the control group's version, was made slightly better than version X, the version with altered sentence context intended to involve students. Contrary to the expectations noted in chapter 3 of this paper, however, these results did not show a significant difference between both versions. Therefore, altered sentence context intended to involve students in a grammar test does not seem to influence the test scores. On the other hand, this study has produced results which partly corroborate the findings of Clapham, Freedle and Kostin, Jafarpur, and Perkins in that the altered sentence context in version X seemed to lead to lower test scores. In contrast to earlier findings (Taylor; Xie and Andrews), however, this study has not found any evidence that these alterations would lead to significant differences.

The absence of a significant difference might be ascribed to the fact that one version of the test, version Y, used sentence context which could be found in the course book, while the other version did not. It could be suggested that both groups have learnt the sentences with the context presented in version Y. This could have enhanced the scores for version Y. However, it cannot be predicted with certainty that version X would yield significantly better results if this defect was taken into account. Future studies could tackle this issue by altering the sentence context of both groups, with one student involving version and one unembellished version. The results could then be in agreement with Celce-Murcia and Hilles' findings. These findings could also be tested in future research by accompanying the altered test method with an altered teaching method, as opposed to an unembellished teaching method and test. For instance, with respect to the relationship between language teaching and language testing, Arthur Hughes has suggested that: "[a] good test is an obedient servant since it follows and apes the teaching. […] Yet we cannot expect testing only to follow teaching. Rather, we should demand of it that it is supportive of good teaching and, where necessary, exerts a corrective influence on bad teaching (5). However, one must not forget that the present study involved just one specific test format and that studies into other test formats might bring forth totally different results (Woods).

The questionnaire, which polled the students about their perception of the test, did show that version X was perceived as more difficult and more enjoying, but the differences with version Y were once again insignificant. Some students added notes to their questionnaire to explain why they found it easy, difficult, enjoying or boring. One student commented on version X: "I think it makes little difference. It is even confusing at times because you are used to the sentences from the course book." Another student who made version X added: "I really enjoyed the fact that there were things like Candy Crush in it." However, some students found the test easier than what they were used to and noted: "It was a bit more amusing because the sentences were fun and already partly completed." These comments explain why the questionnaire generated these results.

Several comments of students who had made version Y showed that it certainly bore a resemblance to the students' customary tests. These comments read: "there is practically no difference between this test and the tests that we usually get," and "I thought it was almost the same!" However, there were more comments that showed that the students were more used to translating the whole sentence, rather than finishing a partly completed one. The comments: "it is nice that you have to complete partial sentences. This way you quickly remember the answers," and "It was nice that some words of the sentences were already given, so you would not have doubts about the order of words," would also explain why the test was deemed easier, as the context for test Y could trigger a student to remember the remainder of the sentence. For test X, this trigger was absent.

The two versions that were created for the purpose of determining the effect of sentence context might have lacked enough context that could be altered. A longer, more elaborate test might have produced results showing a significant difference, as the literature predicted.

In addition, the results of the SJC students could have contributed to the sample size, and therefore the reliability of this study's finding, but the students had, in fact, learned a different set of words and sentences than required. The comments on the questionnaires confirmed that the students had indeed studied other material. For example, one student said: "This had nothing to do with what we had to learn," while another student stated: "We have not learned these words and sentences yet." The students at SJC used a different edition of the *New Interface Coursebook 1 (t)hv Blue Label*. Unfortunately, this was unknown to the researcher before the experiment and while it was conducted.

Chapter 7: Conclusion

Earlier studies have noted a strong relationship between language testing and language learning. As mentioned in the theoretical framework, the teaching method and the test method may be interrelating. First of all, involvement of the students may be highly advantageous to language learning, so teachers should try to make this learning process as appealing as possible while retaining the students' attention to the subject (Celce-Murcia and Hilles; Fulcher and Davidson). Secondly, the relationship between language learning and language testing has been thoroughly investigated, and the term *washback* was coined to represent this relationship (Cheng, Watanabe and Curtis; Taylor). If the positive effects of student involvement in language learning can be extended through *washback*, this involvement might be beneficial for language testing as well. However, extensive quantitative and experimental studies into language testing revealed that context might influence the test scores negatively (Clapham; Jafarpur; Freedle and Kostin; Perkins; Xie and Andrews).

The present study was designed to determine the effect of context on test scores, or more specifically, the effect of sentence context on cloze deletion grammar tests. It was hypothesised that sentence context would have a significant effect on test scores. Moreover, this effect was expected to be negative if the sentence context involved the students more, and was therefore possibly more distracting.

An experiment was conducted among 73 Dutch secondary school students, which were evenly divided into two groups. Both groups made a version of an English cloze deletion grammar test. The control group made test version Y, which resembled the test format with which the students were most familiar. The other group made test version X, which had been altered with the purpose of involving the students in the test and invoking their interests. The test was followed by a questionnaire which goal was to assess the students' perception of the tests' difficulty and enjoyment rate.

The results of this experiment demonstrated that, although version Y was made slightly better, the difference between the scores of both versions was not statistically significant (p = 0.18). The questionnaire showed that the students tended to perceive version X as more enjoyable but version Y as less difficult. Nevertheless, these differences were also statistically insignificant (p = 0.36 and p = 0.07 respectively).

The findings of this study can be interpreted in a variety of ways. Version Y might have been made better because the sentence context did not distract the students. The difference can also be ascribed to the fact that the context of the sentences in version Y was more similar to the assigned learning material than version X.

To conclude, further research might assess the effect of sentence context more adequately by making the control group independent of the learning material. It may also be helpful to determine the influence on the test scores of student involvement if the teaching method itself is also altered.

Works Cited

Alderson, Charles J. and **Dianne Wall.** "Does Washback Exist?" *Applied Linguistics* 14.2 (1993): 115-29. Print.

Bachman, Lyle F. and **Adrian S. Palmer.** *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press, 1996. Print.

Bailey, Kathleen M. "Working for Washback: A Review of the Washback Concept in Language Testing." *Language Testing* 13.3 (1996): 257-79. Print.

Bishop, Dorothy V. M. *Test for the Reception of Grammar*. 2nd ed. Manchester: Age and Cognitive Performance Research Centre, University of Manchester, 1989. Print.

Bosschaart, Gerard, Annie Cornford, Kees van Daalen, Antoon van Eijk, Han Hordijk, Hedzer van der Kooi, Menno Lincklaen Arriëns, Arend Oosterlee, Elly Pothoven, Gertjan Spuy and **Diane van Steekelenburg.** *New Interface Coursebook 1 (t)hv Blue Label.* Utrecht/Zutphen: ThiemeMeulenhoff, 2003. Print.

Carroll, John B., Sidney M. Sapon, Daniel J. Reed, and **Charles W. Stansfield**. *Modern Language Aptitude Test: Manual.* Rockville: Second Language Testing Foundation, 2010. Print.

Celce-Murcia, Marianne, and **Sharon Hilles**. *Techniques and Resources in Teaching Grammar.* Oxford: Oxford University Press, 1987. Print.

Cheng, Liying, **Yoshinori Watanabe**, and **Andy Curtis**. *Washback in Language Testing.* Mahwah: Lawrence Erlbaum Associates Inc. Publishers, 2004. Print.

Clapham, Caroline. *Studies in Language Testing 4: The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension.* Cambridge: Cambridge University Press, 1996. Print.

Davies, Alan. *Principles of Language Testing (Applied Language Studies)*. Oxford: Blackwell, 1990. Print.

Freedle, Roy, and **Irene Kostin**. "Does the Text Matter in a Multiple-choice Test of Comprehension? The Case for the Construct Validity of TOEFL's Minitalks." *Language Testing* 16.2 (1999): 2-32. Print.

Fulcher, Glenn, and **Fred Davidson**. "Test Architecture, Test Retrofit." *Language Testing* 26.1 (2009): 123-44. Print.

Hughes, Arthur. *Testing for Language Teachers, Second Edition.* Cambridge: Cambridge University Press, 2003. Print.

Jafarpur, Abdoljavad. "Is the Test Constructor a Facet?" *Language Testing* 20.1 (2003): 57-87. Print.

McColl Millar, Robert. *Trask's Historical Linguistics.* London: Hodder Education, 2007. Print.

McDonough, Jo, and **Christopher Shaw**. *Materials and Methods in ELT: A Teacher's Guide.* Oxford: Blackwell Publishers, 1993. Print.

Perkins, Kyle. "The Effect of Passage Topical Structure Types on ESL Reading Comprehension Difficulty. *Language Testing* 9.2 (1992): 163-72. Print.

Scott, Linda. "Reviews: Teaching English Grammar: What to Teach and How to Teach It." *ELT Journal* 65.3 (2011): 346-48. Print.

Taylor, Lynda. "The Changing Landscape of English: Implications for Language Assessment." *ELT Journal* 60.1 (2006): 51-60. Print.

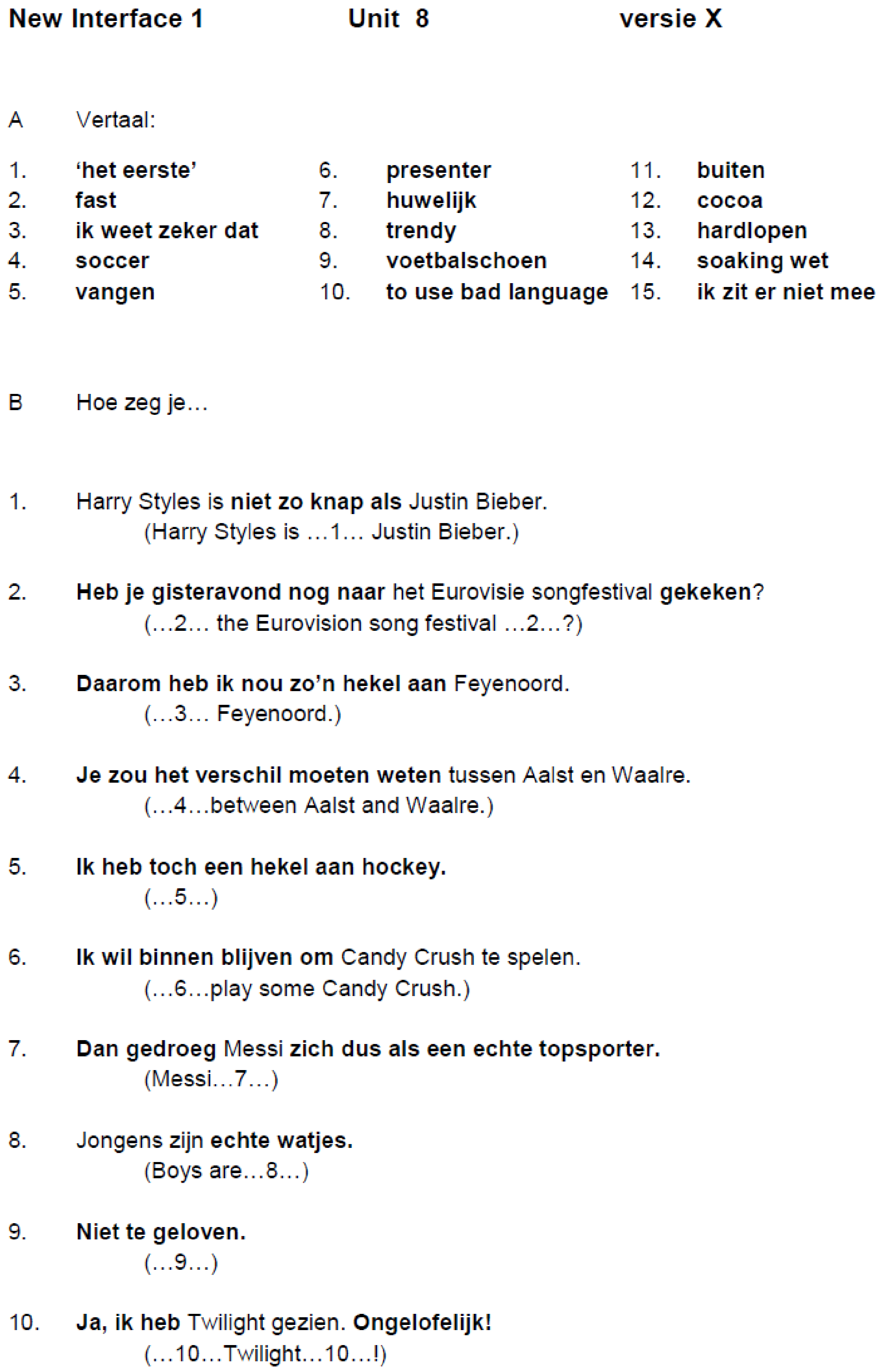
Vernon, Philip E. *The Measurements of Abilities.* London: University of London Press, 1956. Print.

Woods, Devon. *Processes in ESL Teaching: A Study of the Role of Planning and Interpretive Processes in the Practice of Teachers of English as a Second Language.* Diss. U of Utrecht, 1993. Ottawa: Centre for Applied Language Studies, 1993. Print.

Xie, Qin. "Is Test Taker Perception of Assessment Related to Construct Validity?" *International Journal of Testing* 11.4 (2011): 324-48. Print.

Xie, Qin, and **Stephen Andrews**. "Do Test Design and Uses Influence Test Preparation? Testing a Model of Washback with Structural Equation Modeling." *Language Testing* 30.1 (2012): 49-70. Print.

Appendix 1



New Interface 1 **Unit 8** **versie X**

A Vertaal:

1. **'het eerste'**
2. **fast**
3. **ik weet zeker dat**
4. **soccer**
5. **vangen**
6. **presenter**
7. **huwelijk**
8. **trendy**
9. **voetbalschoen**
10. **to use bad language**
11. **buiten**
12. **cocoa**
13. **hardlopen**
14. **soaking wet**
15. **ik zit er niet mee**

B Hoe zeg je…

1. Harry Styles is **niet zo knap als** Justin Bieber.
(Harry Styles is …1… Justin Bieber.)

2. **Heb je gisteravond nog naar** het Eurovisie songfestival **gekeken**?
(…2… the Eurovision song festival …2…?)

3. **Daarom heb ik nou zo'n hekel aan** Feyenoord.
(…3… Feyenoord.)

4. **Je zou het verschil moeten weten** tussen Aalst en Waalre.
(…4…between Aalst and Waalre.)

5. **Ik heb toch een hekel aan hockey.**
(…5…)

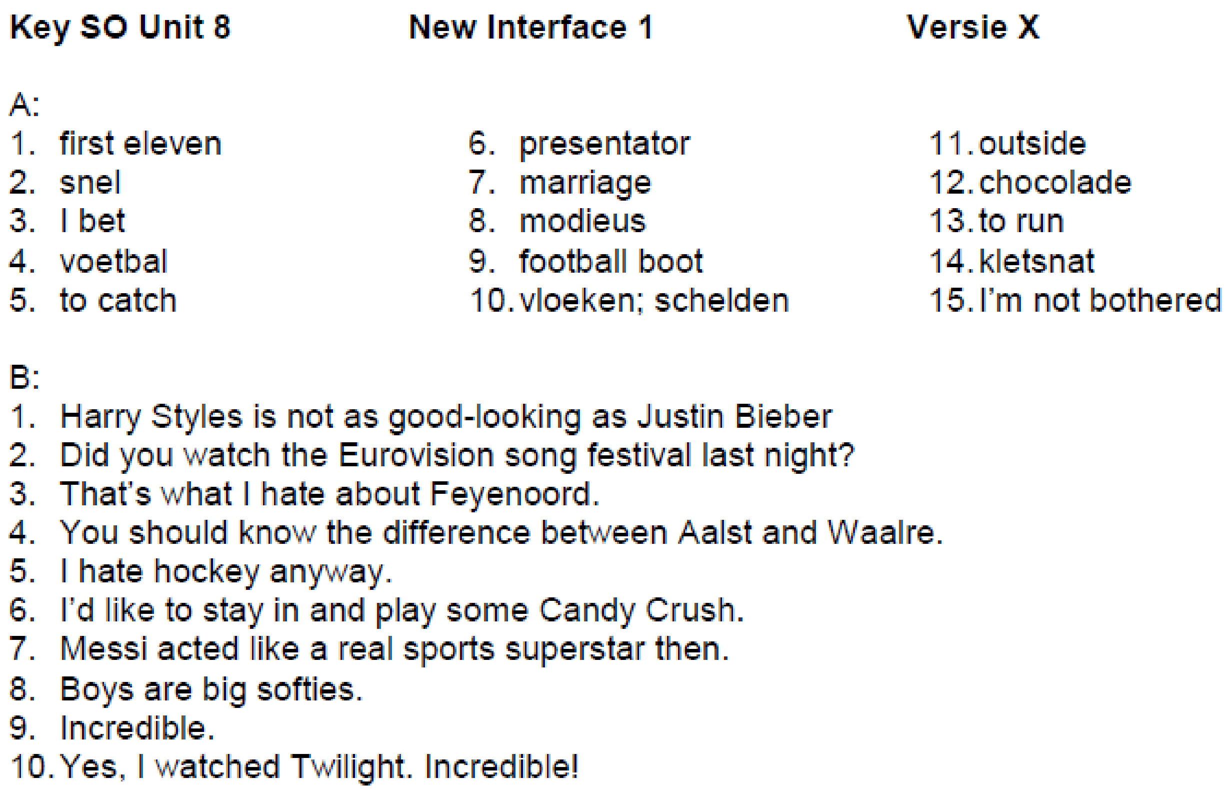
6. **Ik wil binnen blijven om** Candy Crush te spelen.
(…6…play some Candy Crush.)

7. **Dan gedroeg** Messi **zich dus als een echte topsporter.**
(Messi…7…)

8. Jongens zijn **echte watjes.**
(Boys are…8…)

9. **Niet te geloven.**
(…9…)

10. **Ja, ik heb** Twilight gezien. **Ongelofelijk!**
(…10…Twilight…10…!)



Key SO Unit 8 **New Interface 1** **Versie X**

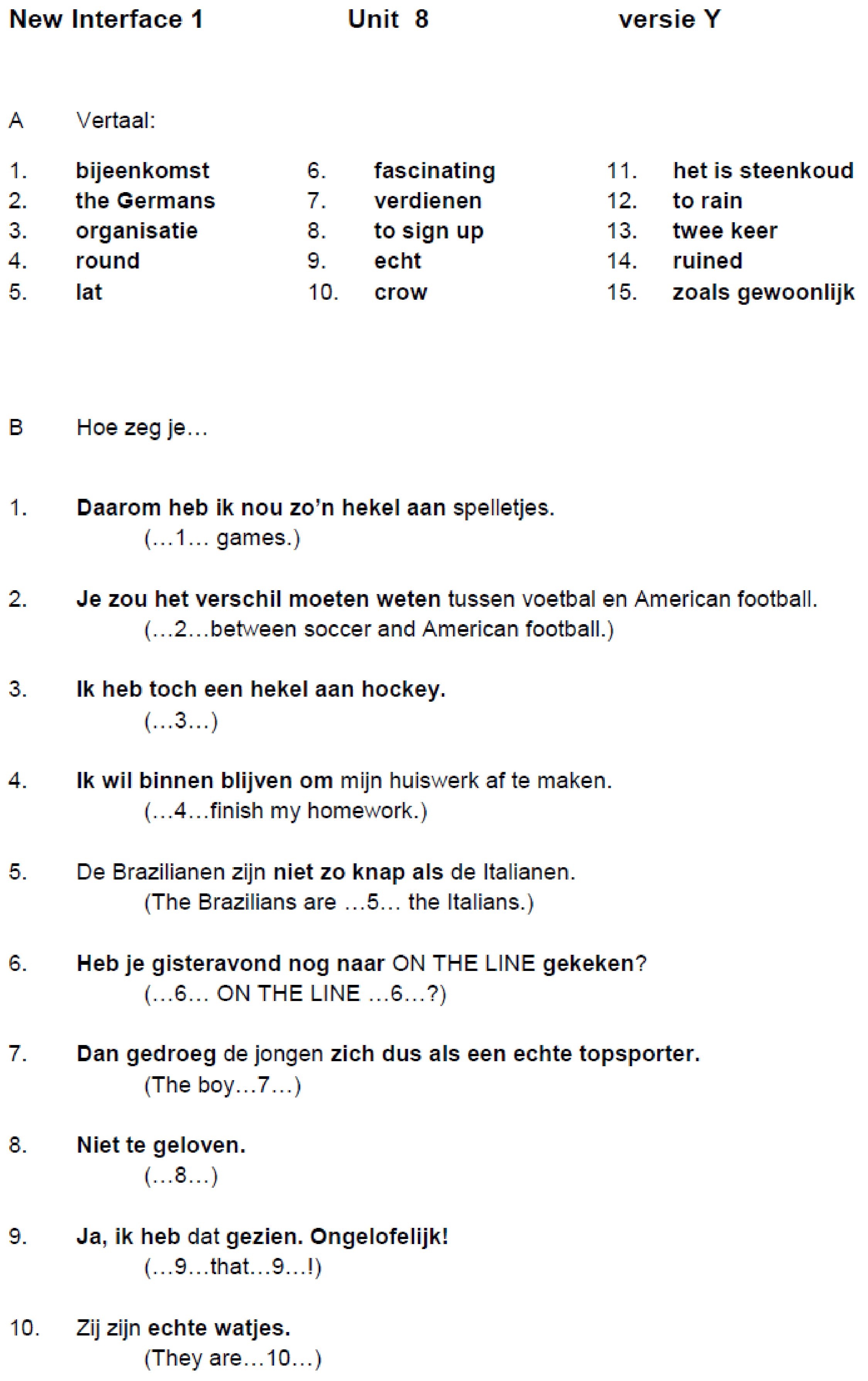
A:

1. first eleven
2. snel
3. I bet
4. voetbal
5. to catch
6. presentator
7. marriage
8. modieus
9. football boot
10. vloeken; schelden
11. outside
12. chocolade
13. to run
14. kletsnat
15. I'm not bothered

B:

1. Harry Styles is not as good-looking as Justin Bieber
2. Did you watch the Eurovision song festival last night?
3. That's what I hate about Feyenoord.
4. You should know the difference between Aalst and Waalre.
5. I hate hockey anyway.
6. I'd like to stay in and play some Candy Crush.
7. Messi acted like a real sports superstar then.
8. Boys are big softies.
9. Incredible.
10. Yes, I watched Twilight. Incredible!

Appendix 2



New Interface 1 **Unit 8** **versie Y**

A Vertaal:

1. **bijeenkomst**
2. **the Germans**
3. **organisatie**
4. **round**
5. **lat**
6. **fascinating**
7. **verdienen**
8. **to sign up**
9. **echt**
10. **crow**
11. **het is steenkoud**
12. **to rain**
13. **twee keer**
14. **ruined**
15. **zoals gewoonlijk**

B Hoe zeg je…

1. **Daarom heb ik nou zo'n hekel aan** spelletjes.
   (…1… games.)

2. **Je zou het verschil moeten weten** tussen voetbal en American football.
   (…2…between soccer and American football.)

3. **Ik heb toch een hekel aan hockey.**
   (…3…)

4. **Ik wil binnen blijven om** mijn huiswerk af te maken.
   (…4…finish my homework.)

5. De Brazilianen zijn **niet zo knap als** de Italianen.
   (The Brazilians are …5… the Italians.)

6. **Heb je gisteravond nog naar** ON THE LINE **gekeken**?
   (…6… ON THE LINE …6…?)

7. **Dan gedroeg** de jongen **zich dus als een echte topsporter.**
   (The boy…7…)

8. **Niet te geloven.**
   (…8…)

9. **Ja, ik heb** dat **gezien. Ongelofelijk!**
   (…9…that…9…!)

10. Zij zijn **echte watjes.**
    (They are…10…)

Key SO Unit 8 New Interface 1 Versie Y

A:

1. session
2. de Duitsers
3. organization; organisation
4. rond
5. crossbar
6. heel boeiend
7. to earn
8. contracteren
9. real; really
10. kraai
11. it’s freezing
12. regenen
13. twice
14. verpest
15. as usual

B:

1. That’s what I hate about games.
2. You should know the difference between soccer and American football.
3. I hate hockey anyway.
4. I’d like to stay in and finish my homework.
5. The Brazilians are not as good-looking as the Italians
6. Did you watch ON THE LINE last night?
7. The boy acted like a real sports superstar then.
8. Incredible.
9. Yes, I watched that. Incredible!
10. They are big softies.

Appendix 3

Universiteit Utrecht

Onderzoek Engelse Taal & Cultuur

Hieronder vind je een aantal vragen of stellingen over de schriftelijke overhoring die je zojuist hebt gemaakt. Vul bij iedere vraag steeds maar één antwoord in. Mocht er iets onduidelijk zijn, kun je altijd een vraag stellen aan de enquêteur of aan je docent(e).

- **1: Welke versie heb je gemaakt?**
  - Versie X
  - Versie Y

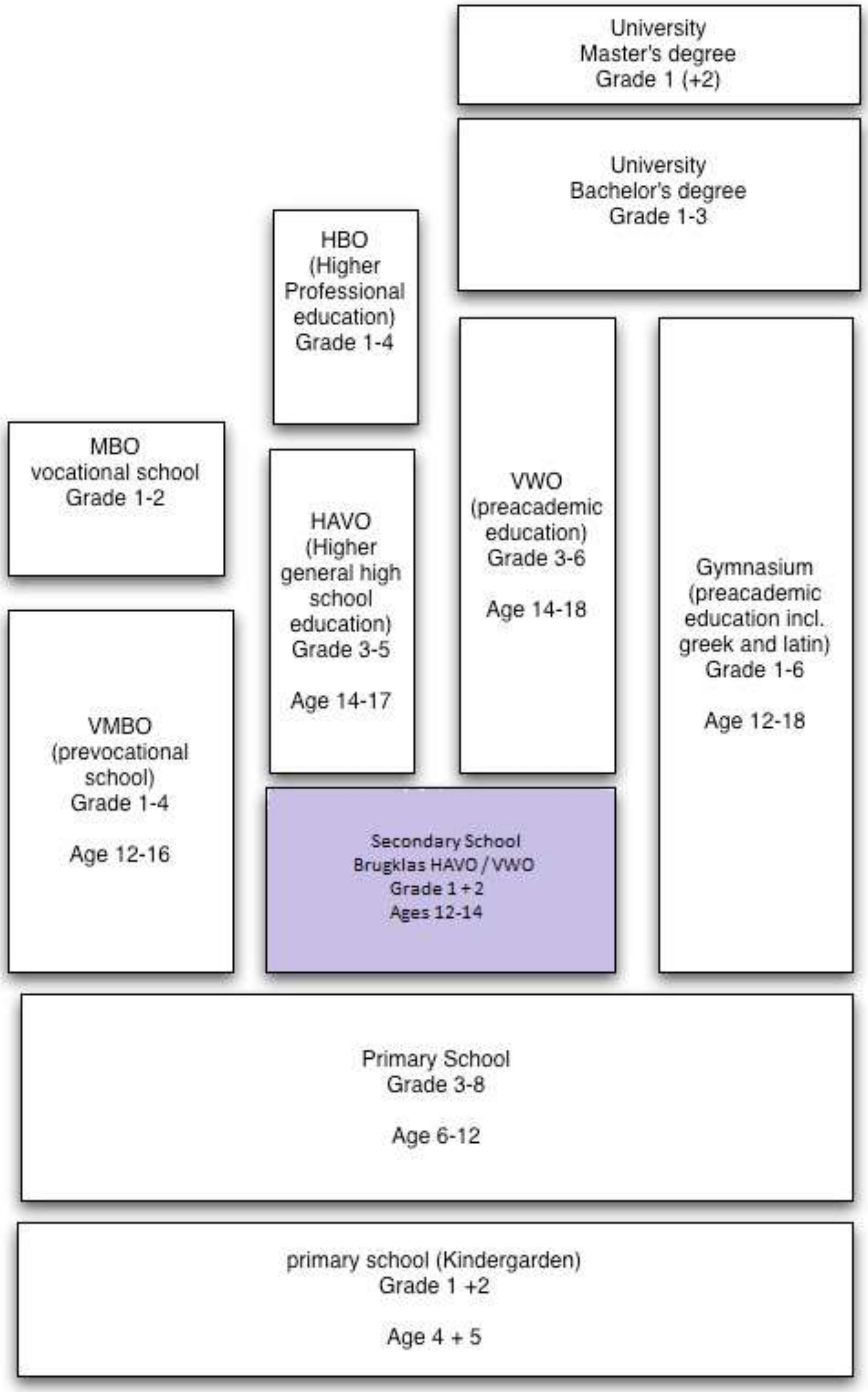
- **2: Ik vond de schriftelijke overhoring…**
  - … veel leuker dan normaal.
  - … iets leuker dan normaal.
  - … even leuk als normaal.
  - … iets minder leuk dan normaal.
  - … veel minder leuk dan normaal.

- **3: Ik vond de schriftelijke overhoring…**
  - … veel makkelijker dan normaal.
  - … iets makkelijker dan normaal.
  - … even makkelijk/moeilijk als normaal.
  - … iets moeilijker dan normaal.
  - … veel moeilijker dan normaal.

Eventuele opmerkingen kun je hieronder opschrijven:

Appendix 4

This image shows how the Dutch educational system is structured. The highlighted area shows which grade the students participating in this study were in.



Appendix 5

Lesson 36

Sentences

They are much fitter and faster than…	Ze zijn veel fitter en sneller dan…
…not as good-looking as the Italians	**…niet zo knap als Italianen.**
Their shoulders are broader and their hips are slimmer.	Hun schouders zijn breder en hun heupen zijn smaller.
They're not as tough as our players.	Ze zijn niet zo sterk als onze spelers.
You should come to important sessions.	Je zou naar belangrijke bijeenkomsten moeten komen.
You should see American footballers.	Je zou eens American footballers moeten zien.
You should know the difference…	**Je zou het verschil moeten weten…**
Girls' stuff.	Meidenwerk.
It was the first eleven football training.	Er was voetbaltraining voor het eerste elftal.
You just kick a ball about.	Jullie lopen maar een beetje tegen een bal te schoppen.
You shouldn't handle the ball.	Je mag de bal niet met je handen aanraken.
American football is all throwing…	Bij American football doen ze niet anders dan gooien…
They're big softies.	**Het zijn echte watjes.**
You're kidding.	Dat meen je niet.
Don't be sexist.	Doe niet zo seksistisch.
American girls could kill to…	Amerikaanse meisjes zouden ere en moord voor doen om…

Words

Gymnastics	Gymnastiek	Girls'stuff	Meidenwerk
First eleven	'het eerste'	Football training	Voetbaltraining
Session	Sessie; bijeenkomst	Important	Belangrijk
Kick (to…)	Trappen	Fit	Fit
Fast	Snel	The Dutch	De Nederlanders
(the) Brazilians	(de) Brazilianen	(the) Germans	(de) Duitsers
(the) Italians	(de) Italianen	Difference	Verschil
I bet	Ik weet zeker dat	Of course	Natuurlijk
Even	Zelfs	Go (to…) / went	Gaan / ging(en)
Organization	Organisatie	League	Competitie
Cross	(hier) kruising	Soccer	Voetbal
Rugby	Rugby	Oval	Ovaal
Round	Rond	Handle (to…)	Aanraken
Throw (to…)	Gooien; werpen	Catch (to…)	Vangen
Tackle (to…)	Onderuithalen	Score (to…)	Scoren
Quarterback	Quarterback	Touchdown	Met de bal de grond raken
Crossbarr	Lat	Posts	(doel)palen
Point	Punt	Amazing	Hier: fantastisch
Shoulder	Schouder	Broad	Breed
Trousers	Broeken	Tight	Strak
Helmet	Helm	For protection	Ter bescherming
Padded shirt	Shirt voorzien van schuimrubber kussentjes	A penalty shoot-out	Een beslissing d.m.v. een serie strafschoppen
Softie	Watje	Tough	Sterk; stoer
Crash into (to…)	Tegen elkaar botsen	Terrifying	Angstaanjagend
Cheerleaders	Cheerleaders	Kill (to…)	Doden; ergens een moord voor doen.

Lesson 37

Sentences

This is different.	Dit is anders.
Incredible.	**Niet te geloven.**
He's too young.	Hij is te jong.
He was awful, wasn't he?	Hij was vreselijk, vond je ook niet?
He acted like a real sports superstar then.	**Dan gedroeg hij zich dus als een echte topsporter.**
Did you watch…last night?	**Heb je gisteravond nog naar…gekeken?**
No, I missed that. When was it on?	Nee, dat heb ik gemist. Hoe laat kwam het?
Yes, I watched that. Incredible!	**Ja, ik heb het gezien. Ongeloofelijk.**
When was it on?	Wanneer was dat op TV?
It's just his parents.	Dat komt door zijn ouders.
Don't ask me.	(Dat) moet je mij niet vragen.
I can't remember his name.	Ik kan me zijn naam niet herinneren.
They're in it for the money.	Ze doen het voor het geld.
It was long past his bedtime.	Hij had al lang in bed moeten liggen.

Words

Sports programme	Sportprogramma	Trendy	Modieus
Last week	De vorige week	Race (to…)	Racen
Presenter	Presentator	Surprised	Verrast
Interview (to…)	Interviewen	Ordinary	Gewoon
Wife – wives	(getrouwde) vrouw – vrouwen	Sign up (to…)	Contracteren
Fascinating	Heel boeiend	Professional	Beroeps
Woman – women	Vrouw – vrouwen	Football boot	Voetbalschoen
Marriage	Huwelijk	Firm	Firma
Divorce	Echtscheiding	Head (to…)	Koppen
Earn (to…)	Verdienen	Real(ly)	Echt
Man – men	Man – mannen	Snooker	Snooker
Behave like (to…)	Zich gedragen als	Shout (to…)	Schreeuwen
Coach	Coach	Use bad language (to…)	Vloeken; schelden
Crow	Kraai	Act like a star (to…)	Zich als een ster gedragen
Photo session	Fotosessie	Aggressive	Agressief
Camera crew	Cameraploeg	Swear (to…)	Vloeken
Honestly	Eerlijk	A spoilt child	Een verwend kind

Lesson 38

Sentences

I can't believe this.	Niet te geloven.
I'm having a nightmare (a bad dream).	Dit kan niet waar zijn (Dit moet een nachtmerrie zijn.)
I'm sweating.	Ik zweet me rot.
That's what I hate about…	**Daarom heb ik nou zo'n hekel aan…**
Never mind.	Geeft niet.
I'd like to stay in and finish my homework.	**Ik wil binnen blijven om mijn huiswerk af te maken.**
Would you? I'd like to go home and watch the telly.	Ja? Ik wil naar huis om TV te kijken.
It's freezing.	Het is steenkoud.
We're going to do games outside on the playing field today.	We hebben vandaag veldgym (buitengym).
I'd like to go home.	Ik ga liever naar huis.
See? I was right.	Zie je wel! Ik had gelijk.
Never mind.	't Geeft niet.
I'm not bothered.	Ik zit er niet mee.
I hate hockey anyway.	**Ik heb toch een hekel aan hockey.**
That's why	Daarom…

Words

Nightmare	Nachtmerrie	Soaking wet	Kletsnat
Believe (to…)	Geloven	Ruined	Verpest
Outside	Buiten	Muddy	Modderig
Playing field	(hier) sportveld	Horrible	Afschuwelijk
It's freezing	Het is steenkoud	As usual	Zoals gewoonlijk
Finish (to…)	Afmaken	Slow	Langzaam
Go home (to…)	Naar huis gaan	Watch out (to…)	Uitkijken
Telly	TV	I'm not bothered	Ik zit er niet mee
Cocoa	Chocolade	Exercise	(lichaams)beweging
Come along (to…)	Meegaan; opschieten	Fresh air	Frisse lucht
Rain (to…)	Regenen	Skin	Huid
Hail	Hagel	Whole	Heel
Blizzard	Sneewstorm	Foot – feet	Voet – voeten
Shower	(hier) bui	Wet through	Doornat
Run (to…)	Hardlopen	Sweat (to…)	Zweten
Twice	Twee keer	Pick teams (to…)	Teams kiezen
Field	Veld		