SOURCES OF VARIANCE IN A YES/NO

VOCABURALY TEST

Wallapha Wongsirichan

A Thesis Submitted in Partial Fulfillment of the Requirements for

the Degree of Doctor of Philosophy in English Language Studies

Suranaree University of Technology

Academic Year 2015

สาเหตุของความแปรปรวนในแบบทดสอบคำศัพท์ แบบให้ตอบ ใช่/ไม่ใช่

นางสาววัลลภา วงศ์ศิริจันทร์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาศิลปศาสตรดุษฎีบัณฑิต
สาขาวิชาภาษาอังกฤษศึกษา
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2558

SOURCES OF VARIANCE IN A YES/NO VOCABURALY TEST

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

Thesis Examining Committee

(Assoc. Prof. Dr. Anchalee Wannaruk)

Chairperson

(Assoc. Prof. Dr. Jeremy Ward)

Member (Thesis Advisor)

(Asst. Prof. Dr. Bussba Tonthong)

Member

(Asst. Prof. Dr. Natjiree Jaturapitakkul)

Member

(Dr. Butsakorn Yodkamlue)

Member

(Prof. Dr. Sukit Limpijumnong)

Vice Rector for Academic Affairs and Innovation

(Dr. Peerasak Siriyothin)

Dean of Institute of Social Technology

วัลลภา วงศ์ศิริจันทร์ : สาเหตุของความแปรปรวนในแบบทดสอบคำศัพท์แบบให้ตอบ ใช่/ไม่ใช่ (SOURCES OF VARIANCE IN A YES/NO VOCABURALY TEST) อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.เจเรมี วอร์ด, 212 หน้า

งานวิจัยนี้มีวัตถุประสงค์ในการค้นหาว่าคำสั่ง (rubric) ที่แตกต่างกัน คำหลอก (nonword) ที่แตกต่างกัน และอัตราส่วนคำหลอกที่แตกต่างกัน ในแบบทดสอบคำศัพท์แบบให้ตอบใช่/ไม่ใช่ (แบบทดสอบ ใช่/ไม่ใช่) จะทำให้ผลสอบแตกต่างกันหรือไม่ ซึ่งแบบทดสอบใช่/ไม่ใช่ นี้ เป็นแบบทดสอบวัดจำนวนคำศัพท์เชิงรับ (receptive vocabulary size) โดยการจัดวางคำศัพท์ให้นักศึกษาตอบ ใช่ หรือ กาเครื่องหมายถูกหน้าคำศัพท์ที่ทราบความหมาย แบบทดสอบนี้จะใส่คำหลอกเข้ามาด้วย ซึ่งคำหลอกเหล่านี้มีไว้เพื่อวัดว่านักศึกษามีการเดาในการสอบหรือไม่ ข้อได้เปรียบของแบบทดสอบประเภทนี้คือ การสร้างแบบทดสอบและดำเนินการสอบนั้นง่ายกว่า เร็วกว่า และประหยัดกว่าแบบทดสอบแบบที่มีอยู่เดิม เช่น แบบทดสอบแบบเลือกตอบ (multiple choice tests) แต่อย่างไรก็ตาม เนื่องจากแบบทดสอบนี้เป็นแบบทดสอบที่ค่อนข้างใหม่ในวงการสอบ จึงมีรายงานว่ายังมีความไม่คงที่หลายประการในการใช้คำสั่ง ประเภทตัวลวง และอัตราส่วนตัวลวง ความแตกต่างเหล่านี้ นำมาสู่การวิจัยว่า คำสั่ง ประเภทคำหลอก และอัตราส่วนตัวลวงที่แตกต่างกันจะเป็นสาเหตุของความแปรปรวนในแบบทดสอบแบบให้ตอบ ใช่/ไม่ใช่ หรือไม่ ดังที่กล่าวมาแล้วข้างต้น

งานวิจัยนี้ได้ทำการเปรียบเทียบคำสั่ง 2 แบบ คือ R1 (ให้ตอบ *ใช่* หน้าคำที่นักศึกษารู้ความหมาย) และ R2 (ให้ตอบ *ใช่* หน้าคำที่นักศึกษาคิดว่ามีในภาษาอังกฤษ) เปรียบเทียบคำหลอก 2 แบบ คือ N1 (คำหลอกแบบเสียงเหมือนหรือคล้ายคำต้นแบบ เช่น *senter* จากคำต้นแบบ *center*) และ N2 (คำหลอกแบบเสียงต่างจากคำต้นแบบ เช่น *fenter* จากคำต้นแบบ *center*) และเปรียบเทียบอัตราส่วนคำหลอก 3 แบบ คือ P1 (คำจริง 50 : คำหลอก 50) P2 (คำจริง 67 : คำหลอก 33) และ P3 (คำจริง 90 : คำหลอก 10) คำสั่ง 2 แบบ คำหลอก 2 แบบ และอัตราส่วนคำหลอก 3 แบบดังกล่าวได้ถูกนำมารวมกันและทำให้เกิดแบบทดสอบที่ต่างกัน 12 ชุด เพื่อเป็นเครื่องมือในการวิจัยครั้งนี้

งานวิจัยนี้ได้ทำการศึกษากับนักศึกษาระดับปริญญาตรีจำนวน 600 คน จากมหาวิทยาลัยแห่งหนึ่งในจังหวัดนครราชสีมา โดยแบบทดสอบแต่ละชุดจะดำเนินการสอบกับนักศึกษาจำนวน 50 คน หลังจากนั้นนักศึกษาจะได้ทำแบบทดสอบแปลคำศัพท์ซึ่งเป็นคำเดียวกันกับคำศัพท์ในแบบทดสอบแบบให้ตอบใช่/ไม่ใช่ โดยนักศึกษาไม่ทราบล่วงหน้า หลังจากที่แบบทดสอบดังกล่าวได้รับการตรวจแล้ว มีการสัมภาษณ์แบบกึ่งมีโครงสร้างกับนักศึกษาจำนวน 72 คน เพื่อให้ทราบสาเหตุว่าทำไมผลสอบจึงออกมาในลักษณะหนึ่ง ๆ ความสัมพันธ์ระหว่างคะแนนของแบบทดสอบใช่/ไม่ใช่กับคะแนนแปลถูกคำนวณเพื่อดูคุณภาพในการทำนายความรู้จำนวนคำศัพท์ที่แท้จริง

(actual vocabulary size) ของนักศึกษา ส่วนคะแนนเฉลี่ยของแบบทดสอบแบบให้ตอบใช่/ไม่ใช่แต่ละชุดถูกนำมาคำนวณ เพื่อดูว่าแบบทดสอบแต่ละชุดมีความต่างอย่างมีนัยสำคัญหรือไม่

ผลการศึกษาพบว่าแบบทดสอบ R1 มีแนวโน้มที่จะมีคุณภาพในการทำนายความรู้จำนวนคำศัพท์ที่แท้จริงของนักศึกษาดีกว่าแบบทดสอบ R2 และแบบทดสอบ N1 มีแนวโน้มที่จะมีคุณภาพดีกว่า N2 ส่วนแบบทดสอบ P1 มีแนวโน้มที่จะมีคุณภาพดีที่สุดในบรรดาแบบทดสอบอัตราส่วนคำหลอก แบบทดสอบ P2 เป็นอันดับ 2 และแบบทดสอบ P3 เป็นอันดับ 3

สาขาวิชาภาษาต่างประเทศ ลายมือชื่อนักศึกษา________

ปีการศึกษา 2558 ลายมือชื่ออาจารย์ที่ปรึกษา________

WALLAPHA WONGSIRICHAN : SOURCES OF VARIANCE IN A YES/NO VOCABURALY TEST. THESIS ADVISOR : ASSOC. PROF. JEREMY WARD, Ph.D. 212 PP.

YES/NO VOCABULARY TESTS/TEST RUBRIC/ NONWORD TYPE/ NONWORD PROPORTION/PARTIAL KNOWLEDGE/TEST INSTRUCTION/ PSEUDOWORDS

This study aims to investigate whether different types of rubrics, nonword types, and nonword proportions in a yes/no vocabulary test (YN test) lead to different test results. A YN test is a kind of receptive vocabulary size test that is presented to learners in a form of a checklist where learners answer *yes* or tick the words they know the meaning of. The test includes nonwords, which are imaginary words added to the test to check whether a test taker do some guess work. The advantages of a YN test is that it is easier, quicker, and cheaper to write and administer compared with other traditional tests such as multiple choice. However, since the YN test is quite new to the test field, many inconsistencies of the test rubrics, nonword types, and nonword proportions have been reported. These discrepancies lead to the investigation whether different test rubrics, different nonword types, and different proportions will be sources of variance in a YN test as mentioned earlier.

The study compared 2 rubrics: R1 (Write Y (yes) if you ***know the meaning*** of the word), and R2 (Write Y (yes) if you ***think that the word exists*** in the language); 2 nonword types: N1 (near-homophone nonwords, which have the same or similar sound to their original words e.g. *senter* from the original word *center*), and N2 (non-

homophone nonwords, which have different sounds from their original words e.g. *fenter* from the original word *center*); and 3 nonword proportions: P1 (50 real words: 50 nonwords), P2 (67 real words: 33 nonwords), and P3 (90 real words: 10 nonwords). The 2 rubrics, 2 nonword types, and 3 nonword proportions were combined and resulted in 12 different YN test versions, which were used as the research tools in this study.

The study was conducted with 600 undergraduate students from a university in Nakhorn Ratchasima. Each test version was administered to 50 participants, followed by an unannounced translation test of the same target words. After the tests were marked, a semi-structured interviewed was conducted with 72 participants in order to explore why and how the results come up in a particular way. The correlation between the scores of each YN test version and the translation test were calculated to find out the quality of each test in predicting the actual vocabulary size of a test taker and the mean scores of each YN test version were compared to find out whether the YN tests yield significantly different test results.

The results suggest that R1 tests tend to have better quality in predicting the actual vocabulary size of the participants than R2, and N1 tests than N2. For the part of proportion tests, P1 tests are likely to have the best quality in predicting the actual vocabulary size of the participants, while P2 the second, and P3 the third.

School of Foreign Languages — Student's Signature________________

Academic Year 2015 — Advisor's Signature________________

ACKNOWLEDGEMENTS

I would like to express my gratitude to the following people for their support, guidance, and valuable comments in conducting this study.

First, I would like to express my deep gratitude to my thesis advisor Assoc. Prof. Dr. Jeremy Ward. It would not have been possible for me to complete this thesis without his great guidance, patience, generosity, and encouragement.

I would also like to express my thanks to Dr. Sutin Chanaboon and Asst. Prof. Thanawat Thitithananan for their great advice on statistics; Assoc. Prof. Dr. Anchalee Wannaruk, Asst. Prof. Dr. Bussba Tonthong, Dr. Butsakorn Yodkamlue, and Asst. Prof. Dr. Natjiree Jaturapitakkul for devoting their precious time reading and giving great comments to my thesis; Prof. Dr. Andrew Lian, Asst. Prof. Payom Konnaimuang, Mr. Peter Bint, Dr. Maneepen Apibalsri for all their help and advice.

I also feel deeply grateful to Khun Sarawut Ratburi, who constantly gives me valuable guidance for conducting this study and for all of his mental support.

My grateful thanks go to all the students who were important part of my pilot and main studies and all the staff of the universities where I conducted these studies.

I would like to thank all the staff at School of Foreign Languages, Institute of Social Technology for all their support.

Finally, I am really grateful to my family and my friends, Khun Norasing Srisiri, Khun Sa-Nguansri Klaeokla, Khun Chadarat Wilaiphol, Khun Pichet Promma, Khun Supitcha Meesukjaosamran, Khun Pornpilai Kittirattrakarn, Khun Chaiyut

Preecha, Khun Phuwasit Wongpian, Khun Sasiwimon Chokechai, Khun Surachai Panomai, Khun Soros Rattanachaisuwan, Khun Jirapa Poonsiri, Khun Sawang and Khun Nattaphat Tonglek, Khun Ampanida Pakarat, and Khun Ratchanee Singkajarn, and Khun Jitlada Chuenjandaeng, who constantly give me encouragement, support, and great advice.

Wallapha Wongsirichan

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF TABLES

Table **Page**

LIST OF TABLES (Continued)

Table **Page**

LIST OF TABLES (Continued)

Table **Page**

LIST OF TABLES (Continued)

Table **Page**

CHAPTER 1

INTRODUCTION

This chapter is an introduction to the present study. It provides background for this research work, research objectives, research questions, significance of the study, scope and limitations, and the operational definitions of the key terms.

1.1 Why study Yes/No vocabulary tests?

It is widely accepted that vocabulary is essential for language learning. A lot of research indicates that vocabulary knowledge correlates well with overall language proficiency (e.g. Saville-Troike, 1984; Laufer, 1997; Nation, 2001). Thus, it can be said that vocabulary knowledge is one crucial factor in learning a language successfully and, therefore, many scholars attempt to measure vocabulary knowledge of language learners in order to estimate their language proficiency. In research field, some researchers roughly measure participants' L2 proficiency from their vocabulary knowledge and in some schools, learners' vocabulary knowledge is used as a placement tool (Meara and Jones, 1988 and 1990; Harrington and Carrey, 2009).

Nevertheless, the term 'vocabulary knowledge' is not easy to be explained because there are no conclusive criteria about to what extent means knowing a word. Nation (2001) suggests that knowing a word involves several aspects as knowing its form, its meaning, and its usage. Some scholars suggest the level of knowing a word on a continuum ranging from recognizing the spelling or pronunciation of a word, to its meanings, to the ability to produce the word quickly and accurately in appropriate contexts (Wesche and Paribakht, 1996).

Knowing a word can also be explained in the following aspects: 1) knowing a word receptively and 2) knowing a word productively. Receptive vocabulary knowledge means knowing the word's form or meaning while productive vocabulary knowledge means knowing the word's meaning and being able to produce that word through writing or speaking (Nation, 2001). For example, if a learner can translate a word from L2 to L1, it means that he knows that word receptively but if a learner can translate a word from L1 to L2, it means that he can write or speak out that word correctly and means that he knows that word productively. Productive vocabulary knowledge also includes the ability to use the words in a context correctly in both semantic and grammatical aspects.

Measuring vocabulary knowledge is, thus, not an easy task because the criteria about to what extent means knowing a word is still unclear as mentioned above. Consequently, vocabulary knowledge tests available now are often created to measure only one aspect of vocabulary knowledge. Two kinds of vocabulary knowledge tests available at present are one that measures vocabulary breadth (or vocabulary size) and the other that measures vocabulary depth. The breadth tests measure how many words a learner knows and usually test at least one meaning of the word while the depth tests measure how well a learner knows other aspects of a word such as its synonyms, its multi-meanings, and its collocations (Zimmerman, 2004).

This present study investigates the Yes/No vocabulary test (YN test), which is a kind of vocabulary breadth test that deals with only receptive aspect of vocabulary knowledge. There might be a question why attention should be paid to the YN test despite the fact that it can test only breadth, not depth. The answer is that, according to many studies, breadth is a stronger indicator of overall language proficiency than depth (e.g., Meara 1996; Read, 2000; Laufer et al., 2004; Webb, 2005). And some research says that breadth is seen as an overlap with depth (Qian, 1999; Vermeer, 2001). That is, knowledge of breadth has some association with knowledge of depth. It was found in Vermeer (2001)'s study that the scores on the breadth test correlated well with those on the depth test. For this reason, breadth tests seem to be an adequate indicator of vocabulary knowledge, which means that the YN test can be one good tool for measuring vocabulary knowledge even though it measures only breadth dimension.

In Thailand, as well as other EFL or ESL countries, the need for measuring vocabulary knowledge of a learner is crucial both for teachers and learners themselves. That is to say, if the teachers know well about their students' vocabulary knowledge, the courses they teach may be adjusted to be more suitable for them. Also, if the learners know well about their vocabulary knowledge, they would have better understanding about their English ability and be able to prepare or improve themselves according to their vocabulary knowledge levels. Until now, although teachers in Thailand make use of a lot of vocabulary tests, problems still remain that it is time-consuming to create and administer the tests and to write a vocabulary tests such as a commonly used multiple-choice test (MC test) may require more time and money for the training of item writers in order to have a good quality test.

The YN test, which is one of the tools that can be used to measure vocabulary knowledge, is one interesting alternative for Thai schools and universities because it is easier to write, less expensive, and quicker both to write and administer when compared with many traditional vocabulary tests. When creating and administering a test are no longer complicated tasks, the use of vocabulary tests will be more often, which means that the more often the learners' vocabulary will be tested and this can bring about the greater improvement of English language teaching and learning in Thailand.

1.2 What is the YN Test?

The YN test is a kind of vocabulary test that has a very simple format. It presents the target words in a form of a checklist and then asks a test taker to indicate the words he knows the meanings of (Read, 2000; Beeckmans et al., 2001; Nation, 2001; Eyckmans et al., 2007). The YN test is designed to measure receptive vocabulary size of the test takers. The test may include nonwords, which are imaginary words added to the test to see whether the test takers overclaim their vocabulary knowledge. It is assumed that if a test taker overclaim his knowledge, he may do some guesswork by choosing the nonwords. Claiming that he knows the nonwords will result in a decrease of the test score. Following is the example of a YN test containing the nonwords *suggestment, montain, arrivetion, facila,* and *craim.*

Example of a Yes/No vocabulary test

<u>Directions</u>**:** Look at the following word listed below. Write Y (yes) in front of the words you know the meaning of, or write N (no) in front of the words you do not know the meaning of.

…..allow	…..suggestment	…..bandit	…..montain
…..observe	…..calculate	…..device	…..arrivetion
…..guess	…..facila	…..promote	…..admire
…..craim	…..argument	…..vision	…..clever

1.3 Advantages of the YN Test

The YN test is very attractive when compared with traditional vocabulary tests such as the multiple choice test (MC test) because it does not require a lot of time to write and administer (Meara and Buxton, 1987). It can also test much larger vocabulary than the MC test or other kinds of tests in a shorter period of time, which is the reason why the test can have more sampling rate than other test formats (Meara and Buxton, 1987; Read, 1988). It does not need a trained item writer to write because the test does not involve complicated process of creating distractors (Anderson and Freebody, 1983). In addition, it is also easy for children to answer (Anderson and Freebody, 1983).

Moreover, the YN test can be used as a placement test (Meara and Jones, 1988, 1990; and Harrington and Carrey, 2009) and can be used to assess learners' vocabulary size for other various purposes including the use in research and the use for diagnostic purposes (Meara, 1991; Read, 2007). These uses can also help in investigations of vocabulary growth and acquisition (Meara, 1991).

1.4 Problems of the YN test

There are, however, some problems on the YN test. For example, there is a question whether the YN test is a real test because it seems that the test is between a conventional language test and self-assessment (Beeckmans et al., 2001).

Another problem is when doing the test, some test takers tend to overclaim their knowledge of the target words (Meara, 1996).

There are also problems about scoring methods used for this test. That is, using different methods to calculate the YN test scores may lead to different test results (Huibregtse, Admiraal and Meara, 2002).

The test rubric is also another problem for this kind of test. That is, there is some inconsistency of the wording in the YN test rubrics. Some test rubrics say: "Write Y (yes) if you ***know the meaning*** of the word" (e.g. Meara and Buxton, 1987; Thoma, 2011), while some say: "Write Y (yes) if you think that the word ***exists*** in the language" (http://www.dialang.org).

Another big issue in the YN test is the problems of nonwords, which are imaginary words added to the YN test in order to measure whether any learners overestimate their knowledge as mentioned earlier. One of the problems is that there is no clear guideline of how to create nonwords. It seems that YN test writers tend to create nonwords according to the phonotactic and morphological rules of the language (Beeckmans et al, 2001). However, there is still a question to what extent the nonwords should differ from the original words (Beeckmans et al., 2001).

Another problem about nonwords is some types of nonwords present problems to some groups of learners. For example, nonwords that were created by adding wrong affixes to the real words such as *suggestment* (which is called pseudoderivatives) cause some problems to higher proficiency test takers. That is, these test takers tend to choose more pseudoderivatives than other groups of test takers (Beeckmans et al., 2001). This may be because higher L2 proficiency learners are likely to have better word formation knowledge (involving forming words by adding prefixes and suffixes) than those of lower proficiency and they may do more guesswork from this concept of knowledge.

Another issue on nonwords is the problem about their proportion in a YN test. The proportions of nonwords are not the same from different studies and the appropriate proportion is still inconclusive (Beeckmans et al., 2001; Eyckmans et al., 2007).

In addition, there is also an issue of proficiency of the test takers when doing the YN test. That is, different proficiency of the students can affect the test scores. The result of the test from the lower proficiency test takers seems to be inconsistent because those test takers tend to choose the nonwords unpredictably and then get very low scores because they are more likely to overclaim their vocabulary knowledge than those of higher proficiency (Meara, 1996).

Despite a lot of problems mentioned earlier, the YN test still remains an interesting topic for many researchers to explore possibly because of its great potentiality as a vocabulary test that is easier, quicker, and cheaper to write and administer than other traditional tests. The reasons why there are many problems or discrepancy in the YN test may be because the YN test is quite new to the test field. That is, it was introduced in 1983, which means it has been known for only 30 years now and still needs a lot of improvement. Further research work is still needed before it can reach its state-of-art status.

1.5 Sources of variance in the YN Test

The sources of variance in the YN test (i.e., reasons of the differences in the YN test results) will be presented in two parts: The variables that will be investigated in this study (independent variables) and those that will be controlled (control variables).

1.5.1 Investigated variables (independent variables)

As mentioned earlier in 1.4, there are a lot of issues or sources of variance that need to be taken into consideration when creating a YN test. In this present study, however, the researcher is investigating 3 variables, which are: 1) test rubric, 2) nonword type, and 3) nonword proportion, because a lot of inconsistency has been found.

1.5.1.1 YN test rubric

The inconsistency of the YN test rubrics is one of the major issues that need to be addressed. That is, there are 2 main types of YN test rubrics used at present:

1) Rubric 1: “Write Y (yes) if you ***know the meaning*** of the word”

2) Rubric 2: “Write Y (yes) if you ***think that the word exists*** in English”

Rubric 1 is seen in, for example, Meara and Buxton (1987), Meara (1996) and Thoma (2011) while Rubric 2 is used in the YN test that is a part of DIALANG, an on-line diagnostic language testing system of 14 European languages. These 2 rubrics seem to require different levels of vocabulary knowledge from test takers in that Rubric 1 wants the knowledge of the meaning of a word while rubric 2 requires only superficial knowledge of having seen a word, not its meaning. This leads to the doubt whether or not the 2 rubrics will yield the same YN test results.

This study, therefore, is designed to compare the results of the YN test using these 2 different rubrics so that we may gain some insight of how to write a more appropriate and consistent YN test rubrics.

1.5.1.2 Nonword type

There is also much discrepancy of nonword creation. That is, there are many types of nonwords available now, but the proper one that should be used in a YN test is still inconclusive. According to Thoma (2011), the nonword types available now are as follows:

1. Non-English like nonwords such as *bdcf*
2. English-like nonwords such as *membel*
3. Pseudohomophones or nonwords that are created to have the same sounds as their original words such as *cerum* from *serum*

4. Pseudo-compounds or nonwords that are created by combining 2 English words to be as a compound word but do not exist in the target language such as *dogtree* or *voicedoor*.
5. Pseudoderivatives or nonwords that are created by adding wrong affixes such as *suggestment* or *entertainism.*

Normally, the nonwords used in the YN test are English-like nonwords and pseudoderivatives. The other types are used widely in psycholinguistics field. There has been no report of using pseudohomophones in the YN vocabulary test (Thoma, 2011) because it might be possible that many researchers may feel that pseudohomophones are too close to real words. For example, Thoma (2011), in his yes/no business English vocabulary size test, mentioned clearly that the nonwords he created were not pseudohomophones or any nonwords that looked too close to real words. This leads to the researcher's interest in exploring whether or not pseudohomophones can be used in the YN test. In other words, the researcher is interested in investigating whether or not the similarity or difference in the sound of a nonword to its original word is a source of variance in a YN test. And this may give some ideas to the question raised by Beeckmans *et al* (2001) that to what extent a nonword should be different from a real word. Thus, 2 types of nonwords, which are:

1) near-homophone nonwords => those that have same/similar sounds to the original words (or Nonword Type 1) such as *willage* from *village.*
2) non-homophone nonwords => those that have different sounds to the original words (or Nonword Type 2) such as *cillage* from *village* will be compared to find out whether the nonwords that are similar (in the aspect of the sound) to the original words and those that are different will lead to different YN test results.

Hence, both types of nonwords will be created by changing only one letter from their original words but will be created to be similar or different in the sound to the original words.

The term *near-homophone nonwords* is an operational definition in this study to cover the nonwords that have one of the characteristics explained below:

1) those that are homophone to their original words (e.g. *sement* from *cement*)

2) those that deemed to be homophone for most Thai learners to their original words (e.g. *vindow* from *window*, *retter* from *letter*)

The sounds of the English alphabet that are considered for most Thai learners to be homophone will be explained later in 3.1.4.2.3 (Chapter 3).

The other term, which is *non-homophone nonwords,* means the nonwords that are not of the 2 characteristics mentioned above. In other words, they are the nonwords that sound different from their original words such as *pement* from *cement or metter from letter*.

1.5.1.3 Nonword proportions

The inconsistency of nonword proportions to the real words is also found in the YN test literature. That is, there are no conclusive guidelines of how many nonwords should be included in a YN test. Meara and Buxton (1987) used 60 real words and 40 nonwords (the proportion of 3:2) while Meara and Jones (1988) employed different proportion of 10 real words and 10 nonwords (the proportion of 1:1). In the DIALANG test battery, the YN tests consist of 50 real words and 25 nonwords (the proportion of 2:1). In Mochida and Harrington (2006), the YN test contains 90 real words and 60 nonwords (the proportion of 3:2) while Harrington and Carey (2009) used 72 real words and 28 nonwords (the proportion of about 3:1).

This discrepancy leads the researcher to the idea of finding out whether or not different proportions will be a source of variance in a YN test.

However, the proportions investigated in this present study will not be all the existing proportions reported in the YN test literature (i.e. 3:2, 3:1, 2:1, and 1:1 as mentioned above) because of some limitation of the research design used in this study. That is, the YN tests in this research will consist of equally 100 items in each test and the 4 mentioned proportions will result in little differences in the number of the real words and nonwords in each proportion. The equal items (i.e. 100 items) will result in equal time or effort of each student doing different versions of the YN tests, which means that more valid and reliable results could be obtained because time or effort will not be extraneous variables for this study. The little differences of the real words and nonwords in each proportion can be explained in this way. That is, from 100 items, the proportion of 3:1, 2:1, 3:2, and 1:1 will result in the number of real words : the number of nonwords as follows: 75:25, 67:33, 60:40, 50:50 respectively. The difference between 75:25 and 67:33 is only 8 real words and 8 nonwords; while the difference between 67:33 and 60:40 is only 7 real words and 7 nonwords; and the difference between 60:40 and 50:50 is only 10 real words and 10 nonwords, which seems to be unpractical and not cost-effective if all the 4 proportions are compared. Therefore, the researcher has designed 3 simple ranges, which are the proportions of 1:1 (the number of nonwords equals the number of real words), 2:1 (the number of nonwords is half of the number of real words), and 9:1 (there are only a few nonwords in the test). The proportions of 1:1, 2:1, and 9:1 will be 50:50, 67:33, and 90:10 respectively, which seems to be more practical to have the ranges that rank from a large number of nonwords (1:1), moderate number of nonwords (2:1), to a few number of nonwords (9:1) and each proportion has enough difference from one another. That is, the difference between 50:50 and 67:33 is 17 real words and 17 nonwords, while the difference between 67:33 and 90:10 is 23 real words and 23 nonwords.

Consequently, this present study will also focus on the nonword proportions by comparing 3 different proportions of real words and nonwords as mentioned earlier:

1) Proportion 1 => 50 real words : 50 nonwords (The number of nonwords is the same as that of the real words)
2) Proportion 2 => 67 real words : 33 nonwords (The number of nonwords is half of the real words)
3) Proportion 3 => 90 real words : 10 nonwords (There are only a few nonwords in the test.)

The 3 proportions will be compared with find out whether they will make significant differences to the YN test results. If yes, then the other proportions existing in YN test literature (i.e. 3:1 and 3:2) can be compared with obtain more precise information. However, if there is no significant difference among each proportion, the comparison among the other existing proportions will not be necessary.

1.5.2 Control variables

1.5.2.1 Test taker's English proficiency

The test takers' different proficiency may play roles in the YN test results because some studies found that higher proficiency test takers tend to choose more pseudoderivative nonwords than those of lower proficiency (Beeckmans et al., 2001). There is also a report that lower proficiency learners are likely to choose the nonwords unpredictably because they tend to do a lot of guesswork and then get very low scores (Meara, 1996).

This study is not investigating on the proficiency variable. Thus, in order to reduce the effect of the English proficiency as a source of variance in the YN test, each YN test version will be administered to the participants of higher, middle, and lower proficiency in equal proportions.

1.5.2.2 Test takers' backgrounds of study

The test takers' backgrounds of study, such as engineering, agricultural technology, or management technology, may be a source of variance in the YN test results. This is because those who have different backgrounds of study may have different attitudes or behaviors in doing the test. For instance, those who are from the school of engineering may be more mathematics-oriented students than those from the school of management, who can be more social-science-oriented. The mathematics-oriented students may have less enthusiasm to do any English tests including the YN tests in this study because their field of study may not emphasize much on learning English compared with the social-science-oriented, who tend to have better attitude toward English and maybe toward the YN tests in this study.

This present study does not investigate the test takers' backgrounds of study, so the researcher tries to reduce the effect of this variable by administering each version of the YN tests to students from various fields of study. This means that one version of the YN tests will be done by students from different majors, not from only one major, who may have the same attitude toward the test because of their same background of study.

1.6 Research Objectives

As mentioned earlier, one of the major problems of the YN test is the inconsistency in creating the rubric, the nonword types and their proportions. Therefore, the objective of this research is to investigate variance resulting from ways of creating different YN test rubrics, different kinds of nonwords, and different nonword proportions (as mentioned in 1.5.1) in order to seek for the information that can bring about the more consistency of creating rubrics and nonwords in a YN test.

1.7 Research Questions

According to the objective of the study (section 1.6), the research questions are formed as follows:

1. Do different test rubrics bring about differences in the test results?
2. Do different types of nonwords bring about differences in the test results?
3. Do different proportions of nonwords to real words bring about differences in the test results?

1.8 Significance of the study

If this present study can reveal whether or not different test rubrics, different nonword types and different nonword proportions affect YN test results, it will provide very useful information for creating more consistent rubric, nonword types, and nonword proportions of a YN test, which is a potential vocabulary test that is easier, cheaper, and quicker to write and administer than the traditional ones such as MC or matching tests. The findings of the present study will contribute greatly to the test writers and certainly to the teachers and learners of language learning alike.

1.9 Scope and Limitations of the Study

1.9.1 Test takers' L1 background

There are reports that different L1 background may lead to some problems on the YN test. For example, the problems of choosing cognate nonwords usually occur with the test takers of Latinate background than of other L1 backgrounds. (Cognates means words that have similar form and meaning in 2 languages such as *coccodrillo* in Italian and *crocodile* in English.) Learners of Arabic background also have some problems with the YN test in that they cannot judge, for example, whether *toilet* and *tilt* are different words because they are often blind to the vowels (Ryan and Meara, 1991).

In this present study, the participants are Thai speakers only.

1.9.2 Test takers' educational background

In this study, the test takers are the university students of a university in Thailand, meaning that they can represent only university students, not other educational levels.

1.9.3 Test takers' English proficiency

In this present study, the English proficiency of the test takers is divided to higher, middle, and lower proficiency according to the O-NET (Ordinary National Educational Test) scores of English. However, the word *higher* here, for example, means higher when compared with the other participants of this study (i.e. within the 600 test takers), not compared with all Thai students, which is considered a limitation of this study. This is because basically the students from the university where the researcher collected the data do not have very high O-NET scores of English (O-NET (Eng]). In this study, the researcher divided the 600 participants into 3 groups of 200 by their English proficiency. Therefore, the higher proficiency participants in this study means the first 200 students who have the highest scores of O-NET (Eng), which ranges from 31 to 85 marks out of 100, not the level of 80 to 100 as many people may think of. Then the second 200 participants who have lower scores compared with the first group (i.e. from 22 to 31 marks out of 100) are considered middle proficiency group, and then the other 200 participants, who get the lowest scores, (i.e. from 14 to 22 marks out of 100) are considered the lower proficiency group.

1.10 Operational definitions of the key terms

1.10.1 YN test

The term *YN test* in this study refers to the yes/no vocabulary test used for measuring receptive vocabulary size of a learner.

1.10.2 Nonword

In this study, the term *nonword* refers to an invented word that do not exist in the language and is added to the YN test in order to correct for guessing such as *vindow* (from *window*) and *girden* (from *garden*). Some scholars use the term *pseudoword, or imaginary word* interchangeably with the term *nonword*, but in this present study, only the term *nonword* is used.

1.10.3 Original word

In this study, the term *original word* is a real word that is used to create a nonword by changing one of its letters. For example, the nonword *lipten* was created from the original word *listen*.

1.10.4 Near-homophone nonwords

In the present study, *near-homophone nonwords* covers nonwords that are real homophones to their original words such as *senter* (from *center*); and those that, for most Thai learners, are homophones to their original words such as *vindow* (from *window*).

1.10.5 Non-homophone nonwords

The term *non-homophone nonwords* in this study refers to the nonwords that are not homophones (have different sound) to their original words such as *teekend* from *weekend*.

1.11 Summary

All the information mentioned earlier is an attempt to provide background and rationale for this present study, reasoning that the YN test is a very potential vocabulary test in that it is easier, quicker, and cheaper to write and administer than other conventional tests. It also includes the statement of the problems of a YN test, which focus on the YN test's discrepancy of test rubrics, nonword types, and nonword proportions. It also provides the purpose of this study, which is to find out whether the inconsistency of test rubrics, nonword types, and nonword proportions will lead to significant difference in the YN test results. In other words, this study aims to explore whether these 3 variables are sources of variance in the YN test. Then the significance of this study is also mentioned, stating that if we know clearly whether these 3 variables are the reasons for differences in the YN test results, it will contribute greatly to the test field in that we will have clearer information of how we can create a more consistent YN test, which will result in a more valid and reliable YN test in the future. In addition, the scope, limitation and definition of key terms are also provided. The detailed explanation of the theoretical background of this study will be presented in the next chapter.

CHAPTER 2

REVIEW OF LITERATURE

This chapter presents some theoretical background of this study. It begins with the importance of vocabulary and the reasons we measure vocabulary knowledge. Then the Yes/No vocabulary test is explained together with its history, its advantages, and its problems including the problems about the test rubrics, nonwords, and nonword proportions.

2.1 Importance of vocabulary

Vocabulary is now regarded by both researcher and teachers as central to language learning (Meara, 1980; Zimmerman, 1997; Read, 2007) and of critical importance to the typical language learner (Zimmerman, 1997). It plays crucial roles in learning a language because lack of vocabulary knowledge will lead to lack of meaningful communication as said by Wilkins (1972:111, cited in Milton, 2009), "Without grammar, very little can be conveyed. Without vocabulary, nothing can be conveyed." Milton (2009) also mentions that words are building blocks of a language and without words there is no language.

Research on reading indicates that vocabulary knowledge strongly predicts reading comprehension (e.g. Stahl, 1990; Blachowicz, Fisher, Ogle, and Watts-Taffe, 2006). It also has relationship with listening (e.g. Bonk, 2000; Rost, 2005), writing (e.g. Laufer & Nation, 1995; Laufer and Paribakht, 1998) and speaking (e.g. Levelt, 1993;

Adolphs and Schmitt, 2003). Moreover, many studies show that vocabulary knowledge has a strong relationship with overall language proficiency (e.g. Laufer, 1997; Nation, 2001). Hence, all this mentioned evidence may yield to the conclusion that vocabulary is essential to language learning and that attention should be highly paid to all aspects of vocabulary in order to be successful language learners and teachers alike.

2.2 What it means by knowing a word?

To explain the term 'vocabulary knowledge' is not an easy task because there are still inconclusive criteria about to what extent means knowing a word. Coady (1993) suggests that knowing a word involves knowing when and where to encounter a target word and its collocation, its register, its appropriate syntactic use, its derivations, its multi-meanings, and its associations. According to Nation (2001), knowing a word involves knowing its form, its meaning, and its usage, which means that knowing a word covers knowing its spelling, pronunciation, meaning, and usage in each context. Takac (2008)'s review of what it means by knowing a word concludes that knowing a word involves several aspects, which are phonological, orthographical, morphological, syntactic, and semantic knowledge, including the understanding of the position of a word in one's conceptual system and the ability to retrieve the word for active use (i.e. speaking and writing).

However, some scholars explain the level of word knowledge on a continuum ranging from recognizing the form of a word (its spelling or pronunciation), to its meanings, to the ability to quickly and accurately produce the word in proper contexts (Wesche and Paribakht, 1996). Similarly, Beck, Perfetti, and McKeown (1982) also suggest that the levels of word knowledge ranges from ‘unknown,’ ‘acquainted’ to ‘established’ knowledge.

Anderson and Freebody (1981) describes vocabulary knowledge in the following aspects: 1) breadth of vocabulary knowledge, which is how many words a learner knows their meanings (at least superficial meaning of the word); and 2) depth of vocabulary knowledge, which means how well a learner knows other aspects of a word such as its synonyms, its multi-meanings, and its collocations.

Knowing a word can also be explained in another important dimension; that is, knowing a word receptively and knowing a word productively. This dimension of vocabulary knowledge will be described in the following section (2.3).

2.3 Receptive and productive vocabulary knowledge

Receptive and productive vocabulary knowledge is one important perspective of vocabulary knowledge. Knowing a word receptively (sometimes referred to as passive vocabulary knowledge) means understanding what the word means as in reading or listening while knowing a word productively (sometimes called active vocabulary knowledge) means being able to produce that word through writing or speaking (Schmitt, 2000; Nation, 2001, 2005). An example of receptive knowledge is when a learner can give an L1 equivalent to a given L2 word, while of productive knowledge a learner can give an L2 equivalent to a given L1 word.

Some researcher see receptive and productive vocabulary knowledge as being on a continuum starting from receptive and then gradually developing to productive knowledge (Laufer and Paribakht, 1998; Henriksen, 1999) while some see it as separate parts (Meara, 1990; Nation, 2001). According to Meara (1990), receptive and productive knowledge are different kinds of knowledge that has some association with each other. Read (2000) suggested that receptive knowledge can be divided into two types of knowledge, which are *recognition* and *comprehension*, while productive knowledge can be divided to *recall* and *use*. The term *recall* means the ability to retrieve the L2 target word from memory while the term *use* means the ability to produce the word through speaking or writing. Laufer and Goldstein (2004) explain more elaborate framework on receptive and productive knowledge. They divide it into 4 different types: 1) *active recall* (retrieval of form by e.g. supplying an L2 word to a given L1), 2) *passive recall (*retrieval of meaning by e.g. supplying an L1 word to a given L2), 3) *active recognition* (retrieval of form by e.g. selecting an L2 word as an equivalent of a given L1), and 4) *passive recognition* (retrieval of meaning by e.g. selecting an L1 word to a given L2).

Receptive vocabulary of L2 learners is larger than productive vocabulary (Laufer, 1998; Laufer and Paribakht, 1998). This means that learners acquire words receptively first, and then productive knowledge is an advanced state of knowdlege (Laufer and Goldstein, 2004).

From the points of view mentioned above, it seems that receptive and productive knowledge is seen somehow differently by different researchers. That is, some view them as different kinds of knowledge while others perceive them as being on a continuum. In this present study, receptive and productive vocabulary knowledge is defined according to Schmitt (2000), and Nation (2001) and (2005), where receptive knowledge is compared with receptive skills (i.e., reading and listening) and productive knowledge is compared with productive skills (i.e., speaking and writing).

2.4 Breadth and depth of vocabulary knowledge

Vocabulary knowledge can be explained in another important dimension as breadth and depth. Breadth of vocabulary knowledge, which is generally used interchangeably with *vocabulary size*, is the term to explain how many words a learner know (at least one meaning and usually primary meaning of the word). In contrast, depth of word knowledge is described as a quality of knowing a word; that is, how well a learner know that word in other aspects such as its synonyms, antonyms, collocations, and usage. While vocabulary breadth involves form (both written and spoken) and meaning, vocabulary depth involves the word's multi-meanings, association, affixes, syntactic features, and other aspects (Nation, 2001).

Vocabulary breadth and depth are sometimes seen as an overlap. Qian (1999) suggest that knowledge of primary word meaning can often include knowledge of synonym, multi-meanings, and collocation.

On the other hand, according to Mezynski, 1983 and Beglar, 1999, learners could possibly know a word presented on a vocabulary breadth test but may not understand or use the word's inflections or derivations, or may not know how to use the word in particular context (Mezynski, 1983). That is, a learner who knows the word *participate* may be able to produce *participates, participated*, and *participating* because it is a regular inflectional patterns, but may not be able to produce the words *participant* or *participation.*

Vocabulary breadth is considered by some researchers (e.g., Meara 1996; Read, 2000; Laufer et al., 2004; Webb, 2005) as more important than depth. Meara (1996) points out that size is the only important aspect when learners know small numbers of words (i.e., 5,000 to 6,000 words). From Read (2000)'s point of view, vocabulary breadth may seems superficial, but it can give a more representative picture of the overall state of the learner's vocabulary than an in-depth knowledge of only limited numbers of words.

Vermeer (2001) suggests an idea that breadth and depth have some association and we cannot possibly say that vocabulary breadth is less valuable than vocabulary depth. In her study, she compared the results of a breadth and a depth test administered to 1,600 Dutch students and found that the results of breadth and depth test correlates well. Thus, she concluded that there is no real distinction between breadth and depth and explained that an in-depth knowledge of a word is a consequence of knowing more words; in other words, when learners know more words, they usually also have finer and deeper word knowledge as a consequence.

Additionally, vocabulary breadth has also been studied more than vocabulary depth (Nakanishi and Shimamoto, 2003), possibly because breadth can represent more overall picture of learner's vocabulary state as mentioned earlier by Read (2000).

Nevertheless, it is possible to examine vocabulary knowledge more comprehensively and systematically by examining both vocabulary breadth and depth of a learner, so that we can have the advantages of both breadth and depth perspectives (Nakanishi and Shimamoto, 2003).

2.5 Measuring vocabulary size (Breadth)

There are several purposes of measuring vocabulary size or breadth of vocabulary knowledge. For example, since vocabulary size has a high correlation with reading ability (e.g. Stahl, 1990; Beglar, 1999), vocabulary size tests are traditionally used in research on reading or literacy development (Read, 2007). It is useful to know how much vocabulary is needed for teaching learners in order for them to reach a threshold level that they can understand written text (Eyckman, 2004), which is assumed to be 95% of the words in a text (Laufer, 1992). Vocabulary size can tell how much an L2 learner lack their vocabulary that can be used to master reading, listening, speaking, or writing (Read, 2007).

2.5.1 Which word to measure?

To measure how many words a learner knows usually involves large samples of words. For L1 learners, normally the sample words come from a large dictionary of contemporary English in order to cover as many words as possible (Read, 2007). However, this sampling of words from dictionaries cannot tell how frequently the words are used; that is, which words are common and which words are rare (Read, 2007). This problem can be solved by using word frequency lists that are based on computer collections of words (corpora). Using words from frequency lists is more systematic in that it can actually tell which set of words learners actually knows (e.g. the first 1000 most frequent words, or the 2000 most frequent words). For L2 learners, who know narrower range of word than the native speakers, sampling of words should come from the frequently used words than the rarely used ones and word frequency lists provide opportunity for testing in this way.

The examples of word frequency lists are 1) the General Service List (GSL) (West, 1953), which is often seen as an old list that lacks modern terms; 2) the British National Corpus (BNC) (Leech, Rayson and Wilson, 2001), which is the most accessible list available both in books and on websites (Read, 2007); and 3) the Academic Word List (AWL) (Coxhead, 2000), the list of 570 word families that are usually occur in written texts in university context.

There is no conclusive idea about which list is the best to choose for creating a vocabulary size test. Test makers or language practitioners now tend to use the best available or the least unsatisfactory list (Read, 2007). In this present study, the BNC is used to create a vocabulary size test called Yes/No test (which will be described in detail in later section) because it is the most accessible source and it contains more modern terms than the GSL as mentioned above.

2.5.2 Vocabulary size (Breadth) tests

The examples of widely used vocabulary size tests now are: the Vocabulary Levels Test (VLT) (Nation, 1983, 2001; Schmitt, Schmitt and Clapham, 2001); the Eurocentres Vocabulary Size Test (EVST) (Meara and Jones, 1990); and the Productive Vocabulary Level Test (PVLT) (Laufer and Nation, 1995); the Vocabulary Size Placement Test (VSPT) of DIALANG (a web-based tests where learners of 14 European languages can assess their proficiency in the target language).

2.6 Measuring vocabulary depth

To be competent language users, only knowing an adequate numbers of words are not enough. They need to know other aspects of those words such as the word's inflections, synonyms, collocations, other meanings, and register. Measuring only breadth of vocabulary knowledge of a learner cannot provide a full description of how well that learner knows a word, so the vocabulary depth test is provided as a solution. Today, there have been rare studies on measuring vocabulary depth when compared with size (Eyckman 2004; Zimmerman, 2004; Read, 2007).

One widely used vocabulary depth test is the Vocabulary Knowledge Scale (VKS) (Paribakht and Wesche, 1993, 1997). The test uses five-point scale combining self report and some evidence by having the learner writing a synonym, L1 translation, or a sentence (Nation, 2001). The test asks how well a learner knows a word starting from "I don't remember having seen this word before", "I have seen this word but don't know the meaning", "I have seen this word before, and I think it means……….(giving synonym or translation), "I know this word. It means…………(giving synonym or translation)" , "I can use this word in a sentence: ……….(writing a sentence)".

Another test of vocabulary depth is Read's (1993, 1998) word associates format (Read, 2007). The purpose of this test is to measure deep word knowledge in a meaningful way. The test employs the concept of word association; that is, in one item there is one target word and six or eight other words. Half of those words are associated with the target words while the other half are not. The words are associated mainly in semantic or collocational aspects, and the format can measure some key elements of the core meaning of the target word, or more than one meaning of the word.

2.7 Yes/No vocabulary tests

2.7.1 What is Yes/No vocabulary Tests?

A Yes/No Vocabulary Test (YN test) is a kind of receptive vocabulary size test that is presented to learners in a very simple format. It is in a form of checklist where learners answer *yes* or tick the words they know the meaning of (see an example of a YN test in section 1.2). In order to prevent the learners from overclaiming their vocabulary knowledge, nonwords (imaginary words) are added to the YN test. That is, the YN test scores will be adjusted downward if the learners answer *yes* to the nonwords.

In calculating the YN test scores, 4 types of answers are considered as follows:

1) *hit* => the answer *yes* to a real word

2) *false alarm* => the answer *yes* to a nonword

3) *miss* => the answer *no* to a real word

4) *correct rejection* => the answer *no* to a nonword

However, the calculation of the scores varies for each scoring formula and there is still no conclusion which formula is the proper one for the YN test. At present, there are 4 main scoring formulae used in the field (Huibregtse et al., 2002), all of which have the same concept that false alarms will be used to adjust the score downward and are indicators for learners' guessing, while hits are rough indicators of their receptive vocabulary knowledge (See 2.7.4.5 for more detail).

2.7.2 History of Yes/No vocabulary tests

The YN vocabulary test was first used for L1 research (Sims, 1929; Tilley, 1936; Zimmerman et al., 1977; cited in Beeckmans et al., 2001). The first conventional format of YN test with nonwords was introduced by Anderson and Freebody (1983) to use with L1 learners while the first to use this current YN test format with L2 learners were Meara and Buxton (1987). Later on, Meara and his colleagues have developed several versions of YN tests. For instance, they developed the Eurocentres Vocabulary Size Test (EVST) (Meara and Jones, 1990), the Swansea EFL Vocabulary Size Test (Meara, 1992), the English Vocabulary Tests (Llex 10k) (Meara, 1996), and the Swansea Vocabulary Levels Test (X-Lex) (Meara and Milton, 2002). The YN tests is also used as a part of the test battery of the DIALANG project, which is an on-line diagnostic language testing system of 14 European languages funded by European Commission. With this program, the test-takers will know their vocabulary size in the target language (See www.dialang.org).

2.7.3 Advantages of Yes/No vocabulary tests

A YN vocabulary test is easier, quicker, and less expensive than other traditional vocabulary tests especially a multiple choice test (MC test) (Meara and Buxton, 1987). Meara and Buxton (1987) mentioned some disadvantages of the MC test and suggested the YN test as an alternative. They explained that for conventional MC tests, one item usually consists of 4 choices, which means that if a test writer wants to test 10 target words, he has to create at least 40 choices and they need to be very careful that those choices should not be too easy or too difficult to make each item reliable, which is a complicated process for test writers.

Meara and Buxton (1987) mentioned some other disadvantages of the MC test. That is, a test taker may know the word in the question but may not be able to answer this item correctly because he may not know the word in the multiple choices. For example, he may know the target word *exercise* but he may not know the meaning of the answer choice *work out*, so he cannot get the right answer. Another reason that he cannot answer this item correctly may be because he knows another meaning of the word *exercise*, which means *practice* but not the meaning in the given context that means *work out*. Furthermore, a test taker may get the correct answer if he knows what the other 3 choices mean without knowing the word *exercise*. If he makes a blind guess, he will have 25% chance of choosing the correct answer. If he knows one choice, he will have 33% of getting the right answer and if he knows 2 choices, he will have up to 50% chance to answer the item correctly.

To eliminate the mentioned problems that may occur when creating an MC test, a YN test could be a good alternative because it contains no multiple choices. Test takers only say whether they know each word or not in this kind of tests. Thus, writing an MC test takes more time and is more expensive when compared with writing a YN test.

Moreover, without multiple choices that may confuse the test takers as mentioned, the YN test is also easier for children to answer (Anderson and Freebody, 1983).

Besides, the YN test can test much wider range of vocabulary than other vocabulary tests, which means that the test allows a better sampling rate of vocabulary when compared with other kinds of vocabulary test formats (Meara and Buxton, 1987; Read, 1988). The samples of words can be drawn from frequency word lists or from some particular texts of specific areas in order to be a vocabulary test of special purposes (Meara, 1991).

Another important aspect of the YN test mentioned by Meara (1991) is that it can be used to create individual learner vocabulary profiles. These vocabulary profiles can be used to monitor a learner's vocabulary growth. Also, the vocabulary profiles can be used to diagnose which vocabulary set a learner need to improve. For example, in some cases, a learner may have better knowledge in academic or specialized vocabulary than the vocabulary of everyday use. The YN test can be a powerful tool to find out which types of vocabulary the learner need most.

The YN test is also often used in the research field where researchers normally study how much vocabulary is required for language use. For example, L2 learners need to know the first 2000 high frequency words for their general language use (Nation and Hwang, 1995), which are necessary for them before they move to other specific types of words that serve their particular needs. Learners need to have the knowledge of 5000-7000 word families to be proficient in speaking English (Hu and Nation, 2000) or to have 95% knowledge of the words in a text in order to understand that text (Hirsch and Nation, 1992). Thus, vocabulary size tests are useful for this kind of study and a YN test is a very practical tool because it is quick to write and administer and can test large vocabulary in a short period of time.

In addition, the YN test can also possibly be a good diagnostic test (Meara, 1991; Read, 2007). By using the YN test, teachers can see vocabulary size of learners and which types of vocabulary they lack. Also, knowing the threshold of vocabulary size that the learners need for doing a particular task such as reading comprehension can be a good guideline for teachers about how much and what kind of vocabulary they need to teach (Eyckmans, 2004).

The YN test also has potentiality to be a placement test (Meara and Jones, 1988 and 1990; Harrington and Carrey, 2009). Generally, a traditional placement test usually tests multiple skills of learners such as listening, grammar, reading, and listening, which means that it takes quite a long time to administer. In the places where time plays crucial roles, this conventional placement test may not be very practical. Meara and Jones (1988) developed a YN test as a new placement test for the Eurocentres Group during 1986-1987, which takes only 10 to 15 minutes for test takers to complete, instead of a time-consuming traditional placement test. The results suggested that this YN test works well as a placement tool.

Also, Harrington and Carrey (2009) suggested that a YN test can be an effective alternative for a traditional multiple-skill placement test. They studied the test results of new students of an Australian English language school who took an on-line version of the YN test of Meara and Buxton (1987) and a school's placement test battery. They compared the results of both tests and found that the results of the YN test correlated well with overall placement decisions, and concluded that the YN test can be a good alternative as a placement task.

Despite many good attributes of a YN test mentioned earlier, there are still some objections and problems that are raised to the YN test as follows:

2.7.4 Problems of the YN test

2.7.4.1 Problems about the test format

There is a question about the YN test format, which is whether the YN test is a real test or a self-assessment. Normally, a conventional test is intended to draw correct answer from the test takers while a self-assessment has no right or wrong answer. It depends on how test takers judge themselves (Oscarson, 1997). A YN test seems to be between a conventional language test and self-assessment and its ambiguity affects the interpretation of the test results (Beeckmans et al., 2001).

Because of this objection, there have been many attempts to prove that the YN test is a valid and reliable measure of vocabulary knowledge by comparing it with other kinds of traditional vocabulary tests such as multiple-choice tests (MC tests), translation tests, and etc. Sims (1929, cited in Eyckmans, 2004) compared the scores of 4 vocabulary tests that were used to measure vocabulary size in L1, which are 1) an MC test, 2) the test in which learners were required to give a definition, a synonym, or the word usage in a sentence, 3) a synonym matching exercise, and 4) a YN test. It was found that the reliability of the YN test is comparable to the other 3 tests although he concluded that a YN test still did not seem to offer acceptable construct validity. Anderson and Freebody (1981) compared a YN test with an MC test and found that the correlation between an MC test scores and a YN test scores was .84, which means that they correlated well with each other. Meara (1996) found in L2 research that a YN test correlated moderately well with other vocabulary tests and with tests of other language skills, especially integrative tests such as cloze, listening comprehension, and reading comprehension. Moreover, Read (2000) also suggests that the YN test has been a valid test and can potentially estimate vocabulary knowledge.

2.7.4.2 Problems about response bias

There is also a question whether the scores of the YN test reflect the real scores of the test takers because there seems to be a lot of space for response bias in a YN test. A response bias is a tendency of a respondent to answer *yes* or *no* without basing on his real knowledge. This answer *yes* or *no* is a result of the respondents' psychological, cognitive or socio-cultural profile rather than their vocabulary knowledge (Eyckmans, 2004). For example, in some YN test research, it was found that the result of a YN test from the lower-scored learners seems to be inconsistent because the learners tend to choose the nonwords unpredictably and, for this reason, they get very low scores because they overclaim that they know the nonwords (Meara, 1996).

The reasons of overclaiming their vocabulary knowledge may come from the following factors suggested by Nation (2007). Nation (2007) mentions that scores of any vocabulary test may not really reflect the real vocabulary knowledge of the test takers because of these following reasons:

1. The test takers may not take the test seriously.

2. They may have little test taking strategies.

3. The attitude of the test takers may lead to different scores, e.g. if the test takers need the high scores for some reasons (e.g. better grades or job

opportunities), they tend to do some guesswork in order to get higher scores (Shohamy, 2001).

Another reason for the response bias may be that of personal behaviors of the test takers in taking tests, which may be a personal trait of each individual. For example, some learners have more conservative behavior in taking tests. From the pre-pilot stage of this present study, the researcher found from the interview of 30 participants that when facing a word that the participants are uncertain about, some participants (23 out of 30) who have more conservative response behavior are likely to answer *no*, while those who are less conservative (7 out of 30) tend to answer *yes* without hesitation.

In short, it could be possible to say that the issue of how test takers of different attitudes or personal traits react differently to the test is the issue of response bias, which may be the main problem of the YN test because the nature of this test tends to be like a self-assessment test (Eyckmans, 2004). That is, the answer Yes of a test taker, for example, may be interpreted in many ways such as 1) The test taker really know the meaning of the word; 2) The test taker have seen the word but does not know its meaning; 3) The test taker are not sure about the word meaning; 4) The test taker knows parts of the target words; 5) The test taker is careless; or 6) The test taker guess blindly. There is no evidence that the test taker really knows the word compared with other kinds of vocabulary test such as MC tests or the True/False tests. For example, in the True/False test, the evidence whether he really answered correctly (e.g. *…T….Cats have four legs.*) is on the test question itself. Therefore, attempts to solve the response bias problem in the YN test may be one of the most important considerations in order that the YN test can reach its state of art.

The most potential ways to reduce the response bias in the YN test may be through a proper design of the test rubric and the nonwords, which may be the most important factors that can help in decreasing the response bias. This is because the test rubric is one of the most important elements that function as a direction of how a test should be done, so it is crucial that the problems of the YN test rubric should be addressed earlier. Also, the nonwords in a YN test, which act as the distractors of the test, are one very important mechanism in the YN test because they are used to adjust downward the scores of those who guess or overestimate their word knowledge. It may be possible to say that test takers who have a tendency to say *yes* to an unknown word or a word that they feel uncertain about should also have a tendency to say *yes* to a nonword, which means that the nonword will adjust the score of their overestimation automatically and, therefore, the response bias will not be a big problem for the YN test. However, there are still questions which kinds of nonwords will have enough quality to adjust the scores and how many nonwords will be adequate to do this job.

For this reasons, this present study will focus on the issues of the test rubric and the nonword issues, which will be explained in the later sections.

2.7.4.3 Problems about test rubrics

A test rubrics plays quite important roles in a test in that it gives information about the nature of the testing procedure, how test takers should respond to the test, setting the test takers’ expectation, and motivating them to do the test appropriately, carefully, and without cheating (Eyckmans, 2004). This means that a poor test rubric may lead to a misunderstanding of how to do a test; and therefore, results in scores that may not represent the real ability of the test takers. Until now, there are only few studies on the YN test rubrics and some of the rubric's inconsistency still remains.

The followings are test rubrics used in YN tests:

1). *Tick the words you know the meaning of, e.g. milk:* √ (Meara and Buxton 1987).

2). *Look through the French words listed below. Cross out words that you do not know well enough to say what they mean. Keep a record of how long it takes you to do the test*. (Meara and Jones 1988: 81).

3). *Read through the list of words carefully. For each word: if you know what it means, write Y (for Yes) in the box, if you don't know what it means, or if you aren't sure, write N (for No) in the box*(Meara, 1992).

4). *Read through the list of words carefully. For each word: if you know what it means, make a mark in the box beside the word. If you don't know what it means, or if you aren't sure, then leave the box empty* (Meara 1996: 43).

5). *Please read each word or phrase carefully. For each expression:*

- *if you know what the whole expression means, then cross the box with **Y** (for Yes);*
- *if you don't know what it means, or if you aren't sure, then cross the box with **N** (for No).*

Please respond to all expressions. Be honest and careful, the list contains real English expressions and expressions that do not exist in English (Thoma, 2011).

6). *In the test, you will be presented with a collection of 'words', some of which are real, and some of which are invented. For each word, you must press the "Yes" button if you think the word exists. If you think it is an invented word, press the "No" button* (http://www.dialang.org).

Rubric 1 to 5 tell what the word "know" means (which is knowing the meaning of a word), but do not mention whether there are nonwords in the test. Telling that there are nonwords in a YN test and not telling so may lead to different test scores because test takers may be more careful not to choose nonwords when they do the test. Abels (1994) studied about telling that there are nonwords and not telling so in a YN test and found that the participants had more careful behavior doing the test when they were informed that there were nonwords in the test.

Rubric 3, 4, and 5 include the forced decision character, which means that test takers need to write both Y and N to every word and, therefore, this characteristic provides no chance of omission of some words. On the other hand, Rubric 1, 2, and 4 allow omission, which may be interpreted as 'No' (The test taker don't know the meaning of the word) or the test taker omits a word because of his carelessness, even though they know that word.

Interestingly, Rubric 6 use different wording from "knowing the meaning of the word" to "knowing whether the word exist". The differences in using 2 different wordings (i.e. know the word's *meaning* and know that the word *exist*) lead the researcher to think whether these 2 test rubrics will affect the test takers in the same ways. The reason is because the 2 test rubrics seem to require different levels of word knowledge from the test takers. That is, the YN test rubric that use the phrase "knowing the meaning of the word" seems to require the knowledge on at least one meaning of a word, while the rubric that contains the phrase "knowing whether the word exist" seems to require different knowledge of the test takers. That is, it seems to require the test takers only to have seen the word or to have some ability of judging whether or not a word in the test should exist in English, without any necessity that the test takers know the meaning of that word.

2.7.4.4 Problems about nonwords

Nonwords are imaginary words added into a YN test in order to check whether the test taker overclaims their vocabulary knowledge. In other words, they act as distractor items in the YN test. There are also other terms used to called nonwords. For example, they are called *non-words* by Zimmerman et al. (1977) and Anderson & Freebody 1983; *imaginary words* by Meara and Buxton (1987); and *pseudowords* by Beeckmans et al. (2001).

Until now, there has been no clear guideline for the nonword creation (Eyckmans, 2004; Thoma, 2011). It seems that people in the YN vocabulary test field agree that nonwords should be constructed according to the phonotactic and morphological rules of the target language (Beeckmans et al, 2001). In other words, nonwords should not break orthographical and phonological rules of the target language.

In psycholinguistics field, nonwords are normally created to be used in a lexical decision task. A lexical decision task is a procedure that measures how quickly people classify stimuli as words or nonwords. Psycholinguists usually employ the lexical decision task as a tool for studying word processing. Thoma (2011) mentioned that nonwords created by psycholinguists are normally as follows:

1. Non-English like nonwords such as *kpkv*
2. English-like nonwords such as *cargin*
3. Pseudohomophones (or nonwords that are created to have the same sounds as their original words such as *senter* from *center*)
4. Pseudo-compounds (or nonwords that are created by combining 2 English words to be as a compound words but do not exist in the target language such as *ricedream* or *headgame*)

For YN vocabulary tests, nonwords that are normally used are English-like nonwords. Anderson and Freebody (1983), who were the first to introduce nonwords to be used in a YN test, created the nonwords according to the following two principles:

1. Changing one or two letters in a real word (e.g. *tand* from *land*;
2. *sancire* from *sincere*)
3. Forming wrong affixes to a base word (e.g. *observement, adjustion*) hich are called pseudoderivatives.

For principle 1, there is a question about to what extent a nonword should differ from its original real words. Abels (1994) suggested creating a nonword by changing more than one letter from an original word in order to prevent a test-taker from misreading it with a real word. However, this suggestion seems not to work well in some cases because changing 2 or 3 letters may result in getting a nonword that is different only one letter from another real word (Eyckmans, 2004). For example, the nonword *carden* that was created from the real word *pardon* by changing 2 of its letters can be interpreted as changing only one letter from another real word *garden*.

Also, creating nonwords following principle 2 (pseudoderivatives) presents some problems. That is, Anderson and Freebody (1983) reported that most of the false alarms the higher proficiency test-takers made were pseudoderivatives while the lower proficiency test-takers chose both types of nonwords (i.e. nonwords created from changing one or two letters and pseudoderivatives).

There has been no report of using pseudohomophones in the YN vocabulary test (Thoma, 2011) possibly because many researchers might feel that pseudohomophones are too similar to real words. For example, Thoma (2011), in his yes/no business English vocabulary size test, mentioned clearly that he did not use pseudohomophones or any nonwords that looked too similar to real words.

In some pseudohomophone studies in psycholinguistics field, there is evidence that people seems to take longer time to tell that a given nonword is a nonword or a real word when that given nonword looks very similar to its original word (e.g. the homophone- or near-homophone nonword such as *servey* from *survey*) compared with the given nonword that looks different from the original word (e.g. the English-like nonword such as *mebra* from *zebra* or even the non-English-like nonwords like *kpbdt*) (e.g. Borowsky and Masson, 1996; Gibbs and Van Orden, 1998). This suggests that the nonwords that were created to be similar to the original words may be more difficult to the test takers to reject than those created to be different to the original words.

Apart from the problems of how to create nonwords, there are also other problems of nonwords to different groups of test-takers. That is, there seem to be problems where the test-takers' L1 background is quite similar to the target language tested in the YN test. For example, Meara and Buxton (1987) found that some nonwords are more attractive to the speakers of a particular L1 background than the speakers of other languages. In other words, this was the problem of cognates (words that have similar form and meaning in 2 languages). They gave an example of the nonword *observement*, which is similar to a real word of Italian or French but not German. This means that it is easier for German speaking test takers to reject this nonword when compared with those who have latinate L1 background such as French or Italian test takers. The result is that this cognate effect leads to the very low scores of the test takers of latinate L1 background because they tended to choose many of this nonword type. However, there was no report that this phenomenon affected the reliability of the YN test nor any report about changing the real words or the nonwords for the French test takers.

Meara and Jones (1990) found the cognate effect again when administering a YN test to the French-speaking test takers. The results showed that the scores of the test takers of other L1 backgrounds correlated better with other language skills than the scores of the French test takers did. They suggested that this could be the results of the close similarity of some French and English vocabulary.

Nonetheless, the results of Meara and Jones (1990) were different from those of Meara, Lightbown and Halter (1994), who carried out the experiment with the French speaking learners in Montreal, Canada. In this study, they hypothesized that a YN test with considerable cognates in the learners' L1 could lead to their overestimation of their vocabulary size while the exclusion of the cognates would lead to the underestimation of their word knowledge. The results of this study suggested that a YN test which has the number of cognate nonwords that are close to the real number of cognates occurring in the language will not result in the less validity of the YN test and the results also showed that the scores of the YN test correlated well with other measures of linguistic skills.

Another problem of the nonwords is reported by Cobb (2000) that Arabic speaking test takers tend to choose a large number of nonwords. This may be because the Arabic speaking learners are likely to be blind to the vowels in their L1 and thus can misread the nonwords in the YN test. For example, they tend to confused the nonword *tilt* with *toilet*, or *mascarate* with *miscreate* (Ryan and Meara 1991).

However, Thoma (2011) pointed out different view of the nonword problems. That is, he suggested that different reaction to the nonwords of different groups of test takers can, in fact, reflect their real language use. For example, the finding that Arabic speaking learners tends to choose a large number of nonwords can be interpreted as real reflection of their difficulties in their L1 word recognition. Therefore, these learners, who are likely to be blind to vowels of the nonwords will also be blind to the vowels of the real words in the YN test. That is, they may confuse a real word with another word and so they claim that they know that real word, which in fact they do not know. This phenomenon can be compared with the *mock hits* phenomenon (Anderson and Freebody, 1983), where the test takers claim that they know a real word, which in fact they confused with another word. For example, they claim that they know the word *hell* which in fact they will translate it as *heel*.

Some researchers suggest that nonwords are not necessary because they do not contribute to the reliability of the YN test. Mochida and Harrington (2006) assessed YN test performance as a predictor of the later performance on the multiple-choice *Vocabulary Levels Test* (VLT) (Nation, 1990) in order to find out whether the test takers overclaim or underclaim their vocabulary knowledge. They found that raw *hits* was the best predictor of VLT scores. They also found that their advanced students seem to have more problems with nonwords at the 5000 frequency level than at the 10000 level.

Shillaw (1996), studying Japanese learners, similarly found little difference between YN tests results without nonwords and test results containing nonwords, and also suggested that YN tests containing only real words may be able to replace the present nonword versions.

Harrington and Carey (2009) suggested that “It is possible that nonwords are not necessary at all and that the appropriate rubric set will be sufficient to reduce guessing, especially for particular learner backgrounds ” (Harrington and Carey, 2009, p. 624).

However, this present study will continue to study nonwords because the results from the pre-pilot study showed that the number of nonwords correlate well with the number of wrong translation in the translation test administered after the YN test (23 out of 30 participants). This means that the nonwords may play important roles if we can find good quality nonwords and a proper number of them. Shillaw (1996), although suggesting that nonwords are unnecessary in the YN test, also mentioned that it is too early to say that YN test can work well without the presence of nonwords.

2.7.4.5 Problems about nonword proportion

Another problems found with nonwords is that the appropriate proportion of nonwords and real words in a YN test is still inconclusive. That is to say, there are no standard of how many real words and nonwords should be included in a YN test. Until now, there has been discrepancy in different studies. For example, Meara and Buxton (1987) used 60 real words and 40 nonwords while in the DIALANG test battery, the YN tests consist of 50 real words and 25 nonwords. In Mochida and

Harrington (2006), the YN test contain 90 real words and 60 nonwords while Harrington and Carey (2009) used 72 real words and 28 nonwords.

2.7.4.6 Problems about word selection

One problem of a YN test is about selecting words to be tested. That is, it is not clear which word categories should be selected, i.e. nouns, verbs, or adjectives. If we randomize the words, a YN test will consist of verbs, nouns, numerals, conjunctions, prepositions, and etc (Eyckman, 2004). Problems may also occur with some word categories such as conjunctions and prepositions because their meanings may depend more strongly on contextual clues (Eyckman, 2004).

2.7.4.7 Problems about scoring methods

At present, there are different scoring methods proposed for the YN test but there has been no conclusion of which scoring formula should be used until now.

Normally, the scores of the YN test are calculated based on the presence of the real words and the nonwords in the test. As a result, the responses of a test taker can be categorized into 4 types as follows:

- hit (H): answering *yes* to a real word
- false alarm (F): answering *yes* to a nonword
- miss (M): answering *no* to a real word
- correct rejection (CR): answering *no* to a nonword

However, there are different scoring formulae to calculate the YN test scores based on the nature of the YN test and the factors that are taken into account.

Huibregtse et al (2002) suggested that the following considerations should be included when calculating YN test scores.

1) There are 2 kinds of correct and 2 kinds of incorrect responses.

2) It is possible that the test takers can do sophisticated guesswork.

3) The test takers have individual response styles.

They mentioned the 2 important variables in a YN test as 1) guessing, and 2) individual response styles. When in doubt, the test takers may do some guesswork, for instance, it is possible that they may think they know a word but they are not sure about its meaning, so they make a sophisticated guess, meaning that they have some partial knowledge about that word and that they do not make a blind guess. According to Signal Detection Theory (SDT), sophisticated guess will be used when a person has some partial knowledge about the items in a test (Huibregtse et al, 2002). In psychology, Signal Detection Theory is used when psychologists want to measure the way a person make decisions under conditions of uncertainty, such as how a person would judge whether the distance would be near or far when they were under foggy conditions.

Besides, when in doubt, the second variable (individual response style) will also play roles. For example, some test takers may tend to answer *yes* while others may tend to answer *no* when they are not sure whether they know a target word (Nunnally and Bernstein, 1994). According to Huibregtse et al (2002), the individual response style plays important parts in a YN test because they affect the way test takers choose to answer *yes* or *no* on both real words and nonwords. For instance, some test takers who have more conservative response style will not answer *yes* very quickly to either real words or nonwords compared with those who are less conservative in their response behaviors. They concluded that response style is an individual trait.

In conclusion, the scoring methods used to calculate a YN test scores should take into account the 3 consideration mentioned above. Below are different YN test calculation methods available at presence.

Huibregtse et al (2002) suggested that the most straightforward way of calculating the YN test scores is to count the number of the correct responses, which are the number of hits and correct rejections. The formula is as follow:

1) H+CR

Nonetheless, Huibregtse et al (2002) do not seems to agree with the use of this formula. They gave the reasons that the 2 types of responses (hits and correct rejections) are considered to be unequal to scores because they are correct for different reasons. They gave a case to support their assumption that a test taker who answers *yes* to 58 out of 60 items (e.g. 39 hits and one correct rejection) gets the same score as the one who has 20 hits and 20 correct rejections. For them, this formula cannot discriminate between these very different types of response behaviors and, thus, may not be proper to be used for a YN test. Instead, Huibregtse et al (2002) suggested that the *hits minus false alarms* formula seems to make more sense because test takers can increase their scores with hits and decrease their scores with false alarms. The formula is as follow:

2) H-F

However, they mentioned that this formula may not be very accurate when taking individual response style into account. For example, a test taker who answers 8 hits (a hit rate of 0.20) and no false alarms, gets a score of 0.20. This response behavior seems to be a conservative response style because the test taker tends not to answer *yes*. It can be possible that the test taker actually knows more words than the scores he gets. This is not taken into account in this formula (H-F). An adjustment of the scores occurs only when the test taker tries to guess and has a false alarm rate larger than 0. In other words, this H-F formula was criticized for not meeting the condition of individual response style.

There are also some other more complex scoring methods, which are proposed for different reasons. For instance, in psychology field, Zimmerman et al. (1977) proposed using a scoring method based on Signal Detection Theory (SDT). They used this scoring method in their recognition test that measure vocabulary knowledge of a test taker and found that this scoring method is potential in that the method can explain for the possibility of response errors. This is because SDT is a theory that tries to quantify the ability to perceive and distinguish between a real signal and noise (i.e. any error or undesired disturbance of a useful information signal) (Pellicer-Sanchez and Schmitt, 2012).

Later on, Anderson and Freebody (1983) are the first to apply a scoring method based on SDT to the field of vocabulary testing. They employed a scoring method that is similar to what is used for multiple-choice or true/false tests to correct for guessing but relied on the hit and false alarm rate. This formula is called *correction for guessing* formula (cfg) (e.g. Huibregtse et al, 2002). It is also used in some of Meara's early work on YN tests (e.g. Meara and Buxton, 1987). This formula is as follow:

3) cfg = [p (H) - p (F)] / [1- p (F)]

(p = proportion)

However, the cfg seems to emphasize more on the hit rate than the false alarm rate and, as a consequence, Meara developed a new formula, which is also based on SDT, referred to as *delta m* (Δm) or Meara's Δm (e.g. Huibregtse et al, 2002). This formula will adjust scores according to how much each subject was guessing. This scoring method (Δm) equals the cfg formula minus the ratio F/H as follow.

4) $\Delta m = [(H - F) / (1 - F)] - (1 / H)$

Nevertheless, Huibregtse et al (2002) mentioned that even if Δm takes into account the guessing variable, it still cannot explain for the individual response style. They studied the differences of each formula (i.e., Formula 2 (*h-f*), 3 (*cfg*), and 4 (Δm)) and proposed their new formula as a better alternative. The new formula is called I_{sdt}, which is also based on SDT but takes both guessing and individual response style into consideration. The formula is as follow:

5) $I_{sdt} = 1 - [[4H(1 - F)] - [2(H - F)(1 + H - F)]] / [[4H(1 - F)] - [(H - F)(1 + H - F)]]$

In their study, they applied these different scoring methods to a YN test results of Meara (1992) and found that Δm formula always resulted in an underestimation of the standard scores, while the cfg led to an overestimation for large Hit proportions. The h-f formula mostly produced a near or sometimes equal values to those of the I_{sdt} method. Therefore, they proposed using their I_{sdt} formula.

There have been some attempts to compare different scoring methods. For example, Ward (2005) applied formula 2 to 5 (excluding formula 1, which is the number of hits and correct rejections) to 6 hit rates (H = .9, .8, .7, .6, .5, and .4) and 3 false alarm rates (.05, .10, and .15). Table 2.1 shows the scores that are the results of using these 4 formulae.

Table 2.1: Four formulae for the YN test

	h-f (formula 2)	**cfg (formula 3)**	**Δm (formula 4)**	**Isdt (formula 5)**
H= .9 F=.05-.1-1.5	85-80-75	89-89-88	84-78-72	85-80-75
H= .8 F=.05-.1-1.5	75-70-65	79-78-76	73-65-58	76-70-65
H= .7 F=.05-.1-1.5	65-60-55	68-67-65	61-52-43	68-62-56
H= .6 F=.05-.1-1.5	55-50-45	58-56-53	50-39-28	60-53-47
H= .5 F=.05-.1-1.5	45-40-35	47-44-41	37-24-11	52-45-38
H= .4 F=.05-.1-1.5	35-30-25	37-33-29	24-08-0	45-37-30

From this table, it can be noted that cfg yields slightly higher scores than H-F, 6%, 9%, 15% at false alarm rates of .05, .1, and .15 respectively. Secondly, Δm results in quite high decline of scores when false alarm rates increase. Thirdly, with high scores, I_{sdt} gives more declining scores for increasing false alarm rates than cfg, while it yields less declining scores with low scores.

Ward (2005) concluded that these 4 formulae yielded quite different results and the mathematical and statistical procedures involved in each formula are not easy to assess. He mentioned that, with his personal communication with Meara, Meara suggested using the simple H-F formula because “All the empirical work we’ve done suggests that p(h)-p(fa) (i.e. *h-f*) is as good as anything, and it’s easier for people to calculate and understand.” (Ward, 2005, p. 30). Meara’s suggestion is complied with Harrington and Carey (2009) because they adopted h-f to be used in thier study on the evaluations of the concurrent validity of an on-line YN test, which is designed to be used as a placement test at an English language school in Australia.

Mochida and Harrington (2006) also attempted to investigate the 4 formulae (and also the raw hits scores) by comparing these formulae in his study. In his study, he compared the scores of YN tests with those of Vocabulary Level Test (VLT), and applied the 4 formulae (plus the raw hits scores) to the YN test scores and found very little differences on the results. They mentioned that the raw hits scores were in fact the best mean predictor of the overall VLT scores, which were used as a standard for finding concurrent validity of the YN tests in his study. They also found that results from simple h-f can be compared with those of cfg and I_{sdt}. Therefore, they suggested that using h-f is an economical and serviceable alternatives, at least for the use of the YN test in a non-research level.

Nevertheless, Beeckmans et al. (2001) commented that each scoring formula seemed inadequate to account for a bias in a YN test, while they still supported that the YN test is a measurement that has enough accuracy. They gave this comment because they found the strange results of using these different formulae in their study of a YN test administered to their French-speaking students who are learning Dutch.

Although there are some attempts to compare each scoring method, there is still no conclusion of which method is the best to be used for a YN test (Eyckmans, 2004; Pellicer-Sanchez and Schmitt, 2012).

2.8 Summary

All this related theoretical information is an attempt to provide existing knowledge for this study, which will investigate whether different test rubrics, different nonword types, and different nonword proportions will be sources of variance for the YN test results. The results from this study will reveal the mentioned issues and will be valuable information for teachers or YN test writers when they want to create a more valid and reliable YN test. The methodology of this study will be presented in detail in the following chapter.

CHAPTER 3

METHODOLOGY

This chapter describes research methodology of the main study and the pilot study. The research methods of both studies include the research design, the participants, the procedures, the instruments, and the data analysis.

3.1 Main study

As mentioned earlier in Chapter 1 and 2, this study explored 3 independent variables: test rubric, nonword type, and nonword proportion. As a consequence, the results of this study were reported in 3 sections according to the 3 research questions.

Research Question 1 (Do different test rubrics bring about different test results?) investigates 2 different test rubrics:

- Test Rubric 1 (R1) => (Write Y (yes) if you ***know the meaning*** of the word)

- Test Rubric 2 (R2) => (Write Y (yes) if you think that the word ***exists*** in English).

For Research Question 2 (Do different types of nonwords bring about significant difference in the test results?), the researcher intended to explore 2 different nonword types:

- Nonword Type 1 (N1) => (near-homophone nonwords)
- Nonword Type 2 (N2) => (non-homophone nonwords)

As for Research Question 3 (Do different proportions of nonwords to real words bring about different test results?), the 3 different nonword proportions were investigated. The 3 proportions were as follows:

- Proportion 1 (P1) => 50 real words : 50 nonwords

  (The number of the nonwords was the same as that of the real words.)

- Proportion 2 (P2) => 67 real words : 33 nonwords

  (The number of nonwords was half of the number of the real words.)

- Proportion 3 (P3) => 90 real words : 10 nonwords

  (There were only few nonwords in the YN test.)

Therefore, 12 different YN test versions were created to serve this purpose. That is, the 12 YN tests were the results of the combination of the 3 factors mentioned above. Each YN test version varied in terms of its test rubric, its nonword type, or its nonword proportion (See 3.1.4.2 for more information).

3.1.1 Research design

The researcher employed mixed methods for this present study. That is, the study comprised 2 methods: 1) a quantitative method on the part of the YN test and translation test data, and 2) a qualitative method on the semi-structured interview data. However, the results of the study depended more on the quantitative part. The interview part, which was qualitative, would be a supplementary part that possibly give more ideas about some answers that were not obtained from the quantitative data such as the information whether the test rubrics were clear for them or any unexpected answers that were useful for this study.

The researcher employed a factorial experiment for this study. Normally, a factorial experiment is used for a study that investigates 2 or more independent variables (or factors). In this present study, the researcher wanted to know whether the 3 independent variables (or 3 factors), which were test rubric, nonword type, and nonword proportion, would be sources of variance in the YN test. The 3 factors in this experiment also varied in their levels. That is, the test rubric had 2 levels, which were Test Rubric 1 and Test Rubric 2. The nonword type also consisted of 2 levels, which were Nonword Type 1 and Nonword Type 2, while the nonword proportion was composed of 3 levels, which were Proportion 1, Proportion 2, and Proportion 3.

In a factorial design, all levels of each independent variable are combined with all levels of the other independent variables to produce all possible conditions. In this present study, there were 3 factors with 2 levels for the first factor (test rubric), 2 levels for the second factor (nonword type), and 3 levels for the third factor (nonword proportion), which were called 2*2*3 (two by two by three) factorial design structure and resulted in 12 different conditions. Here in this study, there were 12 YN test versions that represented these 12 different conditions. Each YN test version was the combination of the 3 factors with different variation of their levels. For example, Test 1 was the combination of Rubric 1, Nonword Type 1, and Proportion 1, while Test 12 was the combination of Rubric 2, Nonword Type 2, and Proportion 3. All the 12 YN test versions were shown in Table 3.1.

3.1.2 Participants of the Main Study

The participants of the main study were 600 undergraduate students at a university in Nakhon Ratchasima attending a mandatory basic English course of the university.

Each YN test version was distributed to 50 students. Each group of 50 students consisted of those of higher (17 students), middle (17 students), and lower English proficiency (16 students) according to their O-NET scores of English or O-NET (Eng). The reason why the researcher distributed each YN test version to 50 students who had similar proficiency to other 50 students was to limit their proficiency (receptive aspect) as a source of variance in this study.

The words *higher, middle, or lower proficiency* here meant higher, middle, or lower when compared with the other participants in this study (600 participants), not compared with other Thai students in general. For example, a student who had higher proficiency in this study meant a student who had the range of O-NET (Eng) scores from 31 to 85 marks out of 100, which was different from what many people might expect. That is, people might expect that higher proficiency students should be those who get O-NET (Eng) scores around 70 to 100. Therefore, what the researcher did was to divide the 600 participants into 3 groups of 200. The first 200 who scored the highest O-NET (Eng) (i.e. from 31 to 85 marks out of 100) were considered the higher proficiency group while the second 200 who had lower O-NET (Eng) scores than the first group (i.e. from 22 to 31 marks out of 100) were considered the middle proficiency group, and the last 200 who scored the lowest (i.e. from 14 to 22 marks out of 100) were considered the lower proficiency group.

Moreover, to limit the effect of fields of study as another source of variance in this study, the researcher distributed one YN test version to 50 participants who came from various majors such as engineering, information technology, medicine, or agricultural technology.

3.1.3 Procedures of the Main Study

1. The 12 YN test versions were distributed to the participants of higher-, middle-, and lower proficiency in equal proportion. This meant that 50 students of different proficiency did one version of the 12 YN tests (600 participants in total).
2. Then each participant did the translation test right after the YN test.
3. Then the semi-structured interview was conducted.
4. After that, the scores of the 2 tests and the interview data were analyzed.

3.1.4 Instruments of the main study

3.1.4.1 O-NET scores of English

The O-NET is a national test created by the National Institute of Educational Testing Services (NIETS), which is an organization responsible for writing national tests to assess the knowledge of the students of all Thai schools. The results of the national tests can tell the effectiveness of those schools in providing the education for their students and, thus, can be a good source of information for those schools' better educational policies. The O-NET is a kind of national test that NIETS created to assess 3 levels of Thai students: 1) primary level (the students who graduated Pratom 6), 2) secondary level (those who graduated Mattayom 3), and 3) high school level (those who graduated Mattayom 6). There are 8 O-NET tests according to 8 different fields of knowledge tested, which were: 1) Thai language, 2) social science, religion, and culture, 3) English language, 4) mathematics, 5) science, 6) health and physical science, 7) vocational education and technology, and 8) art.

The O-NET scores used in this present study were the O-NET scores of English (O-NET (Eng)), and of the high school level, which meant that all of the students who participated in this study had taken it when they graduated high school. The O-NET (Eng) was deemed to be the standard test of English for all Thai students of basic education level (i.e., school level, not university level). The O-NET (Eng) was written by Thai English teachers who had skills in creating English tests and it had been measured for its difficulty value and discrimination power of each item, and also the quality of all distractors. It had also been tested for its reliability (NIET's O-NET Handbook for Mattayom 6 students, 2010).

The O-NET (Eng) was in the form of multiple choice questions composed of 70-100 items that were claimed to measure speaking, writing, and reading ability, which meant that the listening skill was not included in this test. In addition, the items claimed to measure speaking ability in the test were in the form of filling the dialogues with appropriate expression, not the real speaking test that required test-takers to speak out. Besides, the items designed for writing ability were filling in the blank or error recognition questions intending to test grammar or vocabulary knowledge, not the real writing test that required students to write some sentences or essays. This meant that the test tended to measure receptive knowledge rather than productive. Therefore, the scores of the O-NET (Eng) possibly told a rough picture of the participants' proficiency in the receptive aspect, but might not be able to tell us exactly the productive ability of their speaking and writing skills or even the receptive skill in listening. Hence, the O-NET (Eng) scores could be a plausible tool for ranking their receptive proficiency, but might not be a precise tool for telling their real English competence of the 4 skills (reading, listening, speaking, and writing).

However, the YN tests in this study required only the knowledge of a receptive skill, so the scores of O-NET (Eng) were sufficiently used to divide the participants according to their receptive proficiency, which meant that it was not necessary to know their productive ability for this present study.

3.1.4.2 YN tests

As mentioned earlier, there were 12 versions of YN tests in this present study. The 12 version YN tests were the combination of the 3 factors investigated in this study, which were: 1) test rubric (Research Question 1), 2) nonword type (Research Question 2), and 3) nonword proportion (Research Question 3). The 12 different versions of the YN tests are explained in Table 3.1 below. The complete versions of the 12 YN tests are in the appendix.

Table 3.1: The YN tests of 12 different versions

(See the information under this table for the explanation of the codes in the right hand column.)

Test	Combination of
1	R_1 N_1 P_1
2	R_1 N_1 P_2
3	R_1 N_1 P_3
4	R_1 N_2 P_1
5	R_1 N_2 P_2
6	R_1 N_2 P_3
7	R_2 N_1 P_1
8	R_2 N_1 P_2
9	R_2 N_1 P_3
10	R_2 N_2 P_1
11	R_2 N_2 P_2
12	R_2 N_2 P_3

R_1 = *Test Rubric 1 (Write Y (yes) if you* ***know the meaning*** *of the word)*

R_2 = *Test Rubric 2 (Write Y (yes) if you think that the word* ***exists*** *in English)*

N_1 = *Nonword type 1 (near-homophone nonwords)*

N_2 = *Nonword type 2 (non-homophone nonwords)*

P_1 = *Proportion 1 (50 real words : 50 nonwords)*

P_2 = *Proportion 2 (67 real words : 33 nonwords)*

P_3 = *Proportion 3 (90 real words : 10 nonwords)*

3.1.4.2.1 Test rubrics

There were 2 types of test rubrics used in the 12 YN tests. That is, 6 YN tests employed Test Rubric 1 while the other 6 employed Test Rubric 2. The 2 test rubrics were both in English and Thai when administered to the participants. They were as follows:

1) Test Rubric 1:

Directions:

Write Y (Yes) in front of the word if you ***know its meaning*** *and write N (No) in front of the word if you* ***don't know its meaning****. There are some nonwords in this test.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษา****ทราบความหมาย*** *และเขียน N (No) หน้าคำที่นักศึกษา****ไม่ทราบความหมาย*** *ในข้อสอบนี้มีคำหลอกอยู่ด้วย*

2) Test Rubric 2:

Directions:

Write Y (Yes) in front of the word if you think that the word ***exists*** *in English language and write N (No) in front of the word if you think that the word* ***does not exist*** *in the language.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น****คำที่มีในภาษาอังกฤษ*** *และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็น****คำที่ไม่มีในภาษาอังกฤษ***

3.1.4.2.2 Real words

All the 12 YN tests were composed of the real words taken from BNC first and second thousand word frequency lists created by Nation (2004). The first 2000 high frequency words were considered appropriate for general language use (Nation and Hwang, 1995), which were necessary for L2 learners before they move to other specific types of words that served their particular needs. Half of the real words in each YN test were the head words (i.e., words without any inflections (*-s*, *-es*, *-ed*, *-ing*) or derivations (e.g., *-tion*, *-ment*, *-ity*)) drawn from the first thousand word frequency level and the other half drawn from the second thousand word frequency level. They were all content words (i.e. nouns, verbs, adjectives, or adverbs), not function words (i.e. articles, prepositions, or conjunctions). The number of the real words and nonwords varied according to the 3 proportions mentioned earlier. That is to say, the 12 YN tests were divided into 3 different groups according to the 3 proportions, which meant that one group contained 50 real words and 50 nonwords, another group contained 67 real words and 33 nonwords, while the other contained 90 real words and 10 nonwords. All the real words can be seen in the appendix.

3.1.4.2.3 Nonwords

There were 2 types of nonwords in this present study: 1) near-homophone nonwords and 2) non-homophone nonwords. The 2 types of nonwords were created to explore whether the phonological similarity of the nonwords to their original words (real words used to create nonwords) was a source of variance for the YN test results. The term *near-homophone nonwords* was an operational definition in this study to cover:

1) nonwords that were real homophone to their original words (e.g. *persent* from *percent*)

2) nonwords that deemed to be homophone for most Thai learners to their original words (e.g. *vindow* from *window*)

The near-homophone nonwords were created by changing one letter of the original word but the nonwords still had the same or similar sound to the original words. The word 'sound' here has the same meaning as the word 'phoneme', which is the smallest phonetic unit in a language that is able to convey a distinction in meaning, such as the *r* of *ring* and the *s* of *sing* in English. And the phrase 'similar sound' here means similar sound or exactly the same sound for many Thai learners. This phenomenon comes from the confusion of some English phonemes among Thai students. That is, some English phonemes that do not exist in Thai language lead to this confusion. The phonemes of Thai language that are the same as those of English language are: /b/ /d/ /f/ /s/ /h/ /m/ /n/ /l/ /w/ /y/; while the English phonemes that do not exist in Thai language are /g/ /v/ /o/ /ð/ /z/ /ʃ/ /3/ /tʃ/ /d3/ /r/(retroflex) (i.e. the /r/ sound that is produced with the tip of the tongue curled up) ; and those that exist in only Thai language are /c/ /ch/ /?/ /r/(trill) (i.e. the /r/ sound that is produced with the tip of the tongue touching the alveolar many times) (Wongkositkul, 1993). The phoneme /c/ (represented by the alphabet 'จ' in Thai) is close to /d3/ in English but is less forcefully when pronouncing it, while the phoneme /?/ sounds close to the word *oh* in English. The phoneme /ch/ (represented by the alphabet 'ช' in Thai) is close to /ʃ/ and /tʃ/ in English and Thai learners are often confused among these 3 sounds. This is possibly because the /ʃ/ and /tʃ/ sounds do not exist in Thai language, so most Thai learners may pronounce the sound /ch/, which is the nearest sound they can do, instead of the real pronunciation of /ʃ/ and /tʃ/ sounds. Therefore, the researcher employed this confusion between some sounds of English and Thai to create the near-homophone nonwords by changing one letter from the original word to be a letter that might cause confusion; for example, the confusion between /v/ and /w/ (e.g. *vorry* and *worry*); the confusion among /r/(retroflex), /r/ (trill), and /l/ (e.g. *retter* and *letter*); the confusion among /ch/, /ʃ/, and /tʃ/ (e.g. *mashine* and *machine*); and the confusion between /g/ and /k/ (e.g. *marget* and *market*).

As for the non-homophone nonwords, they were also created by changing one letter from the original words but they had different sounds from their original words such as *gertain* from *certain*.

Both 2 types of the nonwords were created by changing only one letter from the original words. Further investigation is needed for the change of 2 or more letters from the original words to form nonwords.

The 2 nonword types also came from the same original words. For example, the near-homophone nonword *senter* and the non-homophone nonword *fenter* were from the same original word *center*.

All the nonwords came from the two-syllable original words. The reason why the nonwords were two syllables instead of one was because it was found that creating nonwords from one-syllable yielded a problem. That is, it could not be easily judged which original words the one-syllable nonword was created from. For example, the one-syllable nonword *pind* might come from the original word *mind* or *pond* or *pine* by changing its first, second, or last letter respectively. This led to the researcher's decision to use the two-syllable original words instead because there was not much possibility that the two-syllable nonwords such as the nonword *surchase*, which came from the original word *purchase*, came from as much as 3 original words as the one-syllable nonword *pind* does as mentioned earlier. This was because the near-homophone and the non-homophone nonwords were created to investigate whether the degree of similarity of the nonwords to their original words was a source of variance in a YN test. Therefore, it should be better for the data analysis if we could exactly tell that a nonword was created from a certain original word, not too many possible original words like the one-syllable nonwords described earlier.

Also, the two-syllable original words are more common to be found in the 1000 and 2000 frequency bands than the three-or-more-syllable words.

The original words were from the first and second thousand word frequency levels. That is, the researcher randomly drew 50 original words having 2 syllables from only the first and the second thousand high frequency word level, not from any higher level.

The 12 YN tests had 3 different numbers of nonwords according to the 3 proportions of real words and nonwords mentioned earlier in 3.1.4.2.2. All the nonwords created for this study can be seen in the appendix.

3.1.4.3 Translation test

The participants did the translation test after the YN test. The researcher included this translation test in order to see whether the participants overclaimed their word knowledge or not. In this study, the translation test was used as the criterion to measure the YN tests' concurrent validity. The concurrent validity is the validity that can be found when comparing one test to another test deemed to be a criterion. In other words, the concurrent validity of the 12 YN tests was found by comparing the YN test results with the translation test results, which was set as the criterion here.

The YN and translation tests were on the same paper. The participants did the YN test in the first round and had to write the translation on the words they answered 'yes' on the same paper in the second round. The participants were informed at first that they have to translate the word they answer *yes*. In the first round (the YN test), the participants had to write with the pink pens provided by the researcher while in the second round (the translation test) with the green pens provided by the researcher. They were not able to use erasers or any wipe out. If they wanted to correct the answer, they were able to only cross out the old answers. The pink and green pens could tell whether or not the participants came back to correct the answer in the first round. In other words, the participants answer only *yes* or *no* in the YN test first and then got the new pen (after returning the pink pens of the first round) to do the translation of the words they answered *yes* for the second round. The translation scores were converted into percentages so that they could be compared with the scores of the YN test which were reported as percentages.

3.1.4.4 Semi-structured interview

The semi-structured interview was used to investigate more about why and how the results came in a particular way that the researcher did not clearly understand. It gave deeper details about the participants' behavior when doing the YN test and also some comments on the YN test they had taken. The researcher used the stratified random sampling method to select two higher-, two middle-, and two lower-scored students from each group (6 interviewees from each group), so there were 72 interviewees altogether from 12 groups. The interview was one-by-one and will be audio-recorded. The interview was conducted in Thai.

There were 9 interview questions, all of which were used for all the 72 participants. The 9 questions were as follows:

1) What were the problems when doing the test or any suggestion to help test takers when doing this kind of test?
2) Did you read the test rubric?
3) Was the test rubric clear for you?
4) How did you interpret the test rubric? Please explain.
5) What were the reasons you answered Y (yes) to some real words that you could not translate afterward?
6) What were the reasons you answered Y (yes) to some nonwords?
7) What were the reasons you answered N (no) to some nonwords?
8) Could you identify the nonwords from the real words and how?
9) How many nonwords did you think were in the YN test?

These different questions were designed to draw the answer for different research questions. That is, Question 1 was used to draw the answer for all the 3 research questions; Question 2-4 were designed to used for Research Question 1(test rubric study); Question 5-8 were designed for Research Question 2 (study on phonological similarity of the nonwords to their original words); and Question 9 was intended to be used for Research Question 3 (nonword proportion study).

3.1.5 Data analysis of the main study

All the 12 YN tests were validated by seeking for their concurrent validity. That is, the scores of the YN tests were compared with the scores of the translation tests using Pearson correlation as a statistical tool.

Also, the 12 YN tests were tested for their reliability by using Conbrach's Alpha, a statistical tool used to find out the internal consistency of a test.

As for the interview data, it was analyzed to gain an insight data as additional information to the quantitative data obtained from the test results.

The data analysis of each research question will be presented later in each section below.

3.2 Analysis of Research Question 1 (test rubric study)

3.2.1 Investigated variables

Research Question 1 is "Do different test rubrics bring about different test results?" This research question seeks for the answer whether the test rubrics that say "Write Y (yes) if you ***know the meaning*** of the word" (which will be mentioned as Test Rubric 1 or R1) and "Write Y (yes) if you think that the word ***exists*** in English" (which will be referred to as Test Rubric 2 or R2) lead to different test results. Since the two test rubrics seems to require different things from the test takers as mentioned earlier in 2.7.4.3, the researcher designed the different YN test versions that employed different test rubrics in order to find out whether it yielded different test results.

3.2.2 YN tests grouped for the analysis

For this analysis, the 12 YN test versions were categorized into 2 main types according to the 2 types of the test rubrics: 1) the R1 YN test type (i.e. Rubric 1 YN test type, which were Test 1-6), and 2) the R2 YN test type (i.e. Rubric 2 YN test type, which were Test 7-12). See Table 3.2 for the grouping of the 12 YN test versions according to the 2 test rubrics. The 6 YN test versions under the same test rubric l also varied in their nonword types and nonword proportions. The reason why the researcher analyzed the data of all 6 different test versions for one test rubric type instead of only one test version was because the researcher did not know whether or not different nonword types and different nonword proportions were the sources of variance for the YN test results, which was the reason why the factorial design was used in this study.

3.2.3 Number of papers

Three hundred papers were analyzed for Rubric 1 test type and the other 300 papers for Rubric 2 test type. That is, the data analysis of R1 test type was on the YN tests done by 300 students (6 YN test versions that employed Rubric 1(Test 1-6)), and the data analysis of R2 test type was on the YN tests done by the other 300 students (6 YN test versions that used Rubric 2 (Test 7-12)) (See Table 3.2).

Table 3.2: Grouping of the 12 YN test versions according to the 2 different test rubrics

Test Rubric 1 (R1)	**Test Rubric 2 (R2)**
Test 1 ($R_1 N_1 P_1$)	**Test 7** ($R_2 N_1 P_1$)
Test 2 ($R_1 N_1 P_2$)	**Test 8** ($R_2 N_1 P_2$)
Test 3 ($R_1 N_1 P_3$)	**Test 9** ($R_2 N_1 P_3$)
Test 4 ($R_1 N_2 P_1$)	**Test 10** ($R_2 N_2 P_1$)
Test 5 ($R_1 N_2 P_2$)	**Test 11** ($R_2 N_2 P_2$)
Test 6 ($R_1 N_2 P_3$)	**Test 12** ($R_2 N_2 P_3$)

R_1 = Test Rubric 1 (Write Y (yes) if you ***know the meaning*** *of the word)*

R_2 = Test Rubric 2 (Write Y (yes) if you think that the word ***exists*** *in English)*

N_1 = Nonword type 1 (near-homophone nonwords)

N_2 = Nonword type 2 (non-homophone nonwords)

P_1 = Proportion 1 (50 real words : 50 nonwords)

P_2 = Proportion 2 (67 real words : 33 nonwords)

P_3 = Proportion 3 (90 real words : 10 nonwords)

3.2.4 Semi-structured interview

As mentioned earlier in 3.1.4.4, the semi-structured interview was carried on in Thai with the 72 interviewees but the 3 questions that focused on the test rubric variable were as follows:

Interview Question 1: What were the problems when doing the test or any suggestion to help test takers when doing this kind of test?

Interview Question 2: Did you read the test rubric?

Interview Question 3: Was the test rubric clear for you?

Interview Question 4: How did you interpret the test rubric? Please explain.

As for Interview Question 1 (What were the problems when doing the test or any suggestion to help test takers when doing this kind of test?), the researcher intended to draw the information about the general problems the participants found when doing the test and also their suggestions, which was useful for enhancing the quality of the YN tests for further studies.

Interview question 2 aimed to investigate whether or not the participants read the test rubric. From this question, we had some ideas about the behaviors of the participants when doing the tests and it possibly explained whether the test results really came from the difference of the 2 test rubrics or other extraneous variables (e.g. the behavior of not reading the test rubric).

Interview question 3 and 4 were designed to explore whether the test rubric was clear for them and whether each test rubric was interpreted in the same direction or not. This information could help explain why the test results came up in a particular way.

3.2.5 Data Analysis of the YN tests

All the main study employed similar methods of data analysis. That is, the YN test scores of the test takers were the indicators of the differences caused by each investigated variables (i.e., test rubrics, nonword type, and proportion of the nonwords). The YN test scores were compared with the translation test scores, which was deemed to be the criterion in this research, in order to see their correlations. If the scores of one YN test that contained the investigated variable showed a higher correlation with the translation scores, it possibly meant that the investigated variable in that YN test tended to be more appropriate to be used than the compared variable. This was because the investigated variable made the test scores more similar to the translation scores.

For example, in this test rubric study, the variables investigated were the 2 test rubrics. Therefore, the scores of the 2 types of test rubric YN tests were compared to find out which set of scores (i.e., the scores of R1 YN tests (Test 1-6) and the scores of R2 YN tests (Test 7-12)) had higher correlation with the criterion scores, meaning that the particular test rubric was fitter to be used in a YN test than the other.

The YN test scores were calculated using the calculation method called *h-f* (see 2.7.4.7 for more detail). One point is given to the hit (i.e., the yes answer to the real word) and then the proportion of all hits was adjusted downward by subtracting with the proportion of false alarm (i.e. the yes answer to the nonword). One false alarm means -1 point. Then the final scores were converted to be in the form of percentage in order that the YN test scores and the translation test scores could be compared.

To determine which test rubric was more appropriate for a YN test, the researcher found the correlation of Rubric1 YN test scores with the translation scores as mentioned earlier. This correlation was calculated by using Pearson Correlation as a statistical tool. Then, the researcher sought for the correlation of the Rubric 2 YN test scores with the translation test scores in the same way. Then the 2 correlation values were compared and the higher value was able to tell which test rubric was better alternative to be used in a YN test.

3.2.6 Data analysis of the translation test

The scores of the translation test were calculated by the right answer to the real words in the YN test. The scores of the translation test were given by 3 English teachers in order to reduce the bias causing from the opinion of the raters when encountering the problem of ambiguity of the Thai translation the participants give. One point was given to the correct translation of a real word. There was only 1 point or 0 points rating (no 0.5 point) in order to make the point equal to the scoring of the YN test, which is also 1 or 0 point. Then these points were converted into percentage. Then, the scores of the YN test and the translation test were compared to find their correlation as mentioned earlier.

3.2.7 Data analysis of the interview

As mentioned in 3.1.4, the audio interview data were transcribed and reported into the form of percentage but focused on the 4 interview questions for this test rubric study mentioned earlier in 3.2.3.3.

3.3 Analysis of Research Question 2 (nonword-type study)

3.3.1 Investigated variables

Research Question 2 is “Do different types of nonwords bring about significant difference in the test results?” The variables to be explored in this question were the phonological similarity of a nonword to its original word (real words that are used to create nonwords). In other words, the researcher wanted to find out whether the nonwords that had the same/similar sound to their original words (near-homophone nonwords) led to different YN test results when compared with the nonwords that had different sounds to their original words (non-homophone nonwords).

3.3.2 YN tests grouped for the analysis

The 12 YN test version was divided into 2 groups according to the 2 different types of nonwords: 1) N1 test type (near-homophone YN test type), and 2) N2 test type (non-homophone YN test type). The N1 test type was Test 1, 2, 3, 7, 8, and 9 while the N2 test type was Test 4, 5, 6, 10, 11, and 12. See Table 3.3 for the classification of the 12 YN test versions according to the 2 types of nonwords. The 6 YN test versions that used the same type of nonwords varied in their test rubrics and nonword proportions.

3.3.3 Number of papers

The papers for this nonword-type study (Research Question 2) came from 300 participants for N1 test type (Test 1, 2, 3, 7, 8, and 9) and from the other 300 participants for the N2 test type (Test 4, 5, 6, 10, 11, and 12) (See Table 3.3).

Table 3.3: Grouping of the 12 YN test versions according to the 2 different nonword types

near-homophone nonwords (N1)	non-homophone nonwords (N2)
Test 1 (R1 N_1 P_1)	**Test 4** (R1 N_2 P_1)
Test 2 (R1 N_1 P_2)	**Test 5** (R1 N_2 P_2)
Test 3 (R1 N_1 P_3)	**Test 6** (R1 N_2 P_3)
Test 7 (R2 N_1 P_1)	**Test 10** (R2 N_2 P_1)
Test 8 (R2 N_1 P_2)	**Test 11** (R2 N_2 P_2)
Test 9 (R2 N_1 P_3)	**Test 12** (R2 N_2 P_3)

*R_1 = Test Rubric 1 (Write Y (yes) if you **know the meaning** of the word)*

*R_2 = Test Rubric 2 (Write Y (yes) if you think that the word **exists** in English)*

N_1 = Nonword type 1 (near-homophone nonwords)

N_2 = Nonword type 2 (non-homophone nonwords)

P_1 = Proportion 1 (50 real words : 50 nonwords)

P_2 = Proportion 2 (67 real words : 33 nonwords)

P_3 = Proportion 3 (90 real words : 10 nonwords)

3.3.4 Semi-structured interview for the analysis

The interview questions intended to draw the information involving the nonword type study were as follows:

Interview Question 5: What were the reasons you answered Y (yes) to some real words that you could not translate afterward?

Interview Question 6: What were the reasons you answered Y (yes) to some nonwords?

Interview Question 7: What were the reasons you answered N (no) to some nonwords?

Interview Question 8: Could you identify the nonwords from the real words and how?

For Interview Question 5 the researcher intended to draw the information about the reasons why the participants could not translate particular words that they answered *yes.* This question gave an insight of their guessing behavior on the real words.

Interview question 6 and 7 aimed to draw the information about why the participants decided to answer 'yes' or 'no' to some nonwords. This was because the answer 'yes' or 'no' to a nonword possibly meant: 1) They guessed blindly, or 2) When they were not sure about the nonword, some might guess by answering 'yes' while some, who had more conservative response style, might answer 'no', or 3) They confused the nonword with another real word so they answered 'yes'.

For Interview Question 8, the researcher intended to elicit whether it was easy or difficult for them to see whether an item was a real word or nonword and why. This question brought about some insight whether the nonwords created for this study had any advantages or drawbacks.

3.3.5 Data Analysis of the YN tests

As mentioned in 3.2.5, all the YN tests were analyzed in the same way both the scoring procedure and the use of the statistical tools (see 3.2.5 for more details. This nonword-type study also used the YN test scores to find out the differences that might result from each type of nonwords. That is, the researcher found the correlation of the N1 test scores with the translation scores (the criterion scores), and did the same with the N2 test scores. Then, the correlation values of the N1 and N2 tests were compared.

The higher correlation value might suggest which types of nonwords were more appropriate to add in a YN test.

3.3.6 Data analysis of the translation test

The analysis of the translation test for Research Question 2 was the same as those for Research Question 1 and 3 (see 3.2.6). The scores of the translation test were analyzed in the same way as mentioned in 3.2.6. However, apart from analyzing the *yes* answer to a real word, the additional step for Research Question 2 was also analyzing the *yes* answer to a nonword (i.e. false alarm). The data of the Thai translation of the false alarms were reported.

3.3.7 Data analysis of the interview

The interview data were analyzed in the same way as mentioned in 3.1.4 but the focus was on Interview Question 5, 6, 7, and 8 mentioned earlier.

3.4 Analysis of Research Question 3 (nonword proportion study)

3.4.1 Investigated variables

Research Question 3 is "Do different proportions of the nonwords to the real words bring about different test results?" This question investigated whether the different proportions of the nonwords to the real words were the sources of variance for the YN test results. The followings are the 3 proportions investigated in this study:

Proportion 1 => 50 real words: 50 nonwords

(the proportion of about 100:100)

Proportion 2 => 67 real words: 33 nonwords

(the proportion of about 100:50)

Proportion 3 => 90 real words: 10 nonwords

(the proportion of about 100: 10)

3.4.2 YN tests grouped for the analysis

For the analysis of this research question, the 12 YN test versions were categorized into 3groups according to the 3 nonword proportions. The 3 groups of YN tests were as follows:

1) P1 test type (i.e. Proportion 1 YN tests, which are Test 1, 4, 7, and10)

2) P2 test type (i.e. Proportion 2 YN tests, which are Test 2, 5, 8, and11)

3) P3 test type (i.e. Proportion 3 YN tests, which are Test 3, 6, 9, and 12)

See Table 3.4 for the grouping of the 12 YN test versions according to the 3 proportions. The 4 test versions that employed the same proportion varied in their test rubrics and nonword types.

3.4.3 Number of papers

The data of P1 test type were 200 papers from 200 students who did the 4 YN test versions that employed Proportion 1 (Test 1, 4, 7, and10); 200 papers from 200 students who did the 4 YN tests that employed Proportion 2 (Test 2, 5, 8, and11); and 200 papers from 200 students who did the other 4 YN tests that employed Proportion 3 (Test 3, 6, 9, and 12). (See Table 3.4)

Table 3.4: Grouping of the 12 YN test versions according to the 3 different nonword proportions

Proportion 1 YN tests (50 real words: 50 nonwords)	**Proportion 2 YN tests** (67 real words: 33 nonwords)	**Proportion 3 YN tests** (90 real words: 10 nonwords)
Test 1 (R1 N_1 P_1)	**Test 2** (R1 N_1 P_2)	**Test 3** (R1 N_1 P_3)
Test 4 (R1 N_2 P_1)	**Test 5** (R1 N_2 P_2)	**Test 6** (R1 N_2 P_3)
Test 7 (R2 N_1 P_1)	**Test 8** (R2 N_1 P_2)	**Test 9** (R2 N_1 P_3)
Test 10 (R2 N_2 P_1)	**Test 11** (R2 N_2 P_2)	**Test 12** (R2 N_2 P_3)

R_1 = *Test Rubric 1 (Write Y (yes) if you* ***know the meaning*** *of the word)*

R_2 = *Test Rubric 2 (Write Y (yes) if you think that the word* ***exists*** *in English)*

N_1 = *Nonword type 1 (near-homophone nonwords)*

N_2 = *Nonword type 2 (non-homophone nonwords)*

P_1 = *Proportion 1 (50 real words : 50 nonwords)*

P_2 = *Proportion 2 (67 real words : 33 nonwords)*

P_3 = *Proportion 3 (90 real words : 10 nonwords)*

3.4.4 Semi-structured interview for the analysis

As mentioned earlier in 3.1.4.4, the semi-structured interview was carried out with the 72 interviewees but the question that focused on the nonword proportion factor wasas follow:

Interview Question 9: How many nonwords did you think were in the YN test?

This interview question gave us more insightful detail about whether the participants have any idea about the proportion of the nonwords.

3.4.5 Data Analysis of the YN tests

As mentioned in 3.2.5, all the YN tests were analyzed in the same way both the scoring procedure and the use of the statistical tools (see 3.2.5 for more detail). However, the researcher focused on the nonword proportion for this analysis. Therefore, the researcher found the correlation of the P1 test scores, P2 test scores, and P3 test scores with the translation test scores as mentioned earlier in 3.2.5. Then the correlation values were compared and the higher value suggested which nonword proportion was more proper to be used in a YN test.

3.4.6 Data analysis of the translation test

As mentioned earlier, the data analysis of the translation test for Research Question 3 was the same as those of Research Question 1 and 2 (see 3.2.6 for more detail).

3.4.7 Data analysis of the interview

The interview data were analyzed in the same way as those of Research Question 1 and 2 (see 3.1.4) but the focus was on Interview Question 9 mentioned earlier.

3.5 Pilot study

3.5.1 Objectives

The objectives of the pilot study were to try out the instruments, to investigate the feasibility of the research procedures, to find out the appropriate time for each YN test version, for the translation test, and for the interview. It also aimed to find out whether there would be any unanticipated problems occurring during the study.

3.5.2 Participants

The participants were 72 students from the same university but were not the same participants of the main study.

3.5.3 Instruments

The instruments for the pilot study were the same as those in the main study.

3.5.4 Procedures

1. The nonwords created by the researcher were examined by a native speaker of English to see whether they looked orthographically and phonologically similar to the real words.
2. Then, the 12 YN test versions were examined (e.g. in terms of the clarity of the test rubrics or the proper length of the tests) by 3 English lecturers who were experienced in writing language tests.
3. Next, the 12 YN tests were adjusted according to the opinions of those 3 lecturers.
4. After that, each of the 12 YN test versions were distributed to 2 participants of higher, 2 of middle, and 2 of lower proficiency (according to their O-NET scores of English). This meant that 6 students of different proficiency did one version of the 12 YN tests (72 participants in total).
5. Then each participant did the translation test right after the YN test.
6. After that, the semi-structured interview was conducted with 12 participants. That is, from 6 participants who did each YN test version, only 1 was randomly selected to be the interviewees of this pilot study.

7. Then the scores of the 2 tests and the interview data were analyzed.
8. The timing of each test and interview session was concluded and the 12 YN tests were adjusted again if there were useful suggestions from the interviewees.

The concurrent validity (between the YN tests and the translation tests) and the reliability (internal consistency) of each YN test version were not reported in this pilot stage. They were reported in the main study instead because each YN test version was done by 50 participants in the main study (a sufficient number for statistical report).

3.5.5 Data analysis

Each test was analyzed the same way as those in the main study to obtain some rough picture about the results and any unanticipated problems. The rough idea about whether the 12 YN tests were correlated with the translation test, which was deemed to be the criterion to find the concurrent validity of the 12 YN tests, can also be obtained in this pilot stage.

Also, the researcher found the timing for each test and the interview of each participant by finding the mean time of each test from the 72 participants and the mean time of the interview session from the 12 interviewees.

In addition, the interview data gave some information whether the instruments need to be adjusted such as the clarity of the test rubrics and the proper size of the letters used in the tests.

3.5.6 Pilot study results

The researcher carried out all the steps mentioned above and found that all the steps were feasible for the main study. However, there were some minor changes on the instruments used in this study.

3.5.6.1 Some changes on nonwords

After consulting with a native speaker in order to find out whether the nonwords created by the researcher were similar to real words in the aspects of both sound and spelling, the nonword *senter* and *fenter* that came from the original word *center* were cut out. This resulted from the suggestion that problems of different spelling styles between British English and American English (e.g. *centre* and *center*) may occur if a student is familiar with only one spelling style and has no idea with the other.

3.5.6.2 Some changes on the format of the 12 YN tests

After consulting with the 3 lecturers who had experiences in writing English tests, there were some changes on the format of the YN tests. That is, according to their suggestions:

1) The space between each line of the test was wider in order to make the test easy to be read.

2) A square box was provided in front of each word so that it would be easy for test-takers to put Y (yes) or N (no) in the box and also easy for them to see whether they leave any box unanswered.

3) A line says, "Don't leave any box blank" was included under the test rubric to make sure that the test takers answer every item.

3.5.6.3 An addition of an interview question

After trying-out the interview questions with the participants, the researcher found out a new interesting question after question 4 (How did you interpret the test rubric?). Most interviewees in the pilot study gave quite the same answers to question 4 that they had to write Y in front of the word they know its meaning (for Rubric 1 test type) or the word they think exists in English (for Rubric 2 test type) or N if they do not know the meaning (for Rubric 1 test type) or think that the word do not exist in English (for Rubric 2 test type). These answers prompted the researcher to ask one more question "Could you tell me frankly which criteria you used to answer Y or N for each word?" and an interesting answer for Rubric 1 test type is "If I know the meaning of the word, I will answer Y and if I have seen it but don't know the meaning, I will also answer Y". Another interesting answer for Rubric 2 test type is "If I have seen the word before, I will answer Y, and if the spelling and the sound of the word I haven't seen before is ok for me when I read and pronounce it, I will also answer Y." These example answers seem to give clearer picture about the participants' guessing behavior, so the researcher decided to include the question "Could you tell me frankly which criteria you used to answer Y or N for each words?" in order to get more information to explain why the results of research question 1 (Do different test rubrics bring about significant difference in the test results?) comes out in a particular way.

3.5.6.4 Timing of the YN tests, the translation test, and the interview

At first, the researcher roughly set up 10 minutes for each YN test, 10 minutes for translation test, and 10 minutes for the interview of each participant. Then, the researcher asked the participants after administering each test whether the timing is appropriate for them**.** The researcher also recorded the time spent in each interview and found out the more appropriate time for the main study. That is, the time for each session will be:

1) 10 minutes for the YN tests

2) 15 minutes for the translation test

3) 15 minutes for the interview of each student

3.6 Summary

The information in this chapter was all about the methodology of this present study, which was designed to be mixed methods comprising a quantitative method on the part of 600 participants doing the YN tests and a qualitative method on the part of a semi-structured interview of 72 participants. The researcher used a factorial design to collect the data on the part of the YN tests, which was suitable for the study where more than one factor were investigated. In this study, three factors were explored, which were test rubric (2 test rubrics), nonword type (2 nonword types), and nonword proportion (3 nonword proportions). The combination of these three factors resulted in 12 different YN test versions, each of which varied in the test rubric, the nonword type, and the nonword proportion. The procedure of this study was also provided, starting from the participants doing the YN test, followed immediately by the unannounced translation test and then the semi-structured interview. The information about the pilot study, which intended to find out the feasibility of the main study and the timing of each session when collecting the data, was also explained in this chapter. The pilot study results showed that all the processes were possible, with some minor changes such as cutting out some nonwords from the YN tests and an interview question was added. The next chapter will provide detailed information about the results of the main study.

CHAPTER 4

RESULTS

This chapter presents the validity and reliability results for the 12 YN tests, the statistical results, and the semi-structured interview results of the present study.

4.1 Main study results

4.1.1 Reliability results of the 12 YN tests

The reliability results of the 12 YN test versions were calculated using Cronbach's Alpha as a statistical tool and are shown below in Table 4.1.

Table 4.1: Reliability results of the 12 YN tests

YN Test	Number of participants	Number of test items	Cronbach's Alpha
1 (R1N1P1)	50	100	.935
2 (R1N1P2)	50	100	.941
3 (R1N1P3)	50	100	.956
4 (R1N2P1)	50	100	.925
5 (R1N2P2)	50	100	.945
6 (R1N2P3)	50	100	.953
7 (R2N1P1)	50	100	.927
8 (R2N1P2)	50	100	.909
9 (R2N1P3)	50	100	.911
10 (R2N2P1)	50	100	.906
11 (R2N2P2)	50	100	.895
12 (R2N2P3)	50	100	.881

R_1 = Test Rubric 1 (Write Y (yes) if you ***know the meaning*** *of the word)*

*R_2 = Test Rubric 2 (Write Y (yes) if you think that the word **exists** in English)*

N_1 = Nonword type 1 (near-homophone nonwords)

N_2 = Nonword type 2 (non-homophone nonwords)

P_1 = Proportion 1 (50 real words : 50 nonwords)

P_2 = Proportion 2 (67 real words : 33 nonwords)

P_3 = Proportion 3 (90 real words : 10 nonwords)

In Cronbach's Alpha calculation, the acceptable reliability is at 0.700, and we can see from Table 4.1 that all versions of the YN tests yield higher reliability values than 0.700 ranking from .881 of Test 12 (R2N2P3)to .956 of Test 3 (R1N1P3), meaning that all of these YN tests are reliable.

4.1.2 Concurrent validity results of the 12 YN tests

In this present study, the concurrent validity results were obtained by finding the correlation between the YN test scores and the translation scores. The statistical tool Pearson Correlation was employed to calculate these correlations. The concurrent validity results of the 12 YN tests are presented in Table 4.2 as follows.

Table 4.2: Concurrent validity results of the 12 YN tests

Rank	YN Test	Number of participants	Pearson correlation	Sig. (2-tailed)
1	1 (R1N1P1)	50	.947**	**.000**
2	4 (R1N2P1)	50	.931**	**.000**
3	5 (R1N2P2)	50	.930**	**.000**
4	7 (R2N1P1)	50	.922**	**.000**
5	6 (R1N2P3)	50	.915**	**.000**
6	8 (R2N1P2)	50	.903**	**.000**
7	2 (R1N1P2)	50	.869**	**.000**

Table 4.2: Concurrent validity results of the 12 YN tests (Cont.)

Rank	YN Test	Number of participants	Pearson correlation	Sig. (2-tailed)
8	11 (R2N2P2)	50	.841**	**.000**
9	10 (R2N2P1)	50	.828**	**.000**
10	3 (R1N1P3)	50	.799**	**.000**
11	12 (R2N2P3)	50	.745**	**.000**
12	9 (R2N1P3)	50	.687**	**.000**

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

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

*R_1 = Test Rubric 1 (Write Y (yes) if you **know the meaning** of the word)*

*R_2 = Test Rubric 2 (Write Y (yes) if you think that the word **exists** in English)*

N_1 = Nonword type 1 (near-homophone nonwords)

N_2 = Nonword type 2 (non-homophone nonwords)

P_1 = Proportion 1 (50 real words : 50 nonwords)

P_2 = Proportion 2 (67 real words : 33 nonwords)

P_3 = Proportion 3 (90 real words : 10 nonwords)

The correlation between the YN test scores and the translation scores can tell the quality of each YN test version. If the YN test scores correlate well with the translation scores (i.e. the scores that are considered a criterion of the participants' vocabulary size in this present study), it suggests that this YN test version has high quality to predict actual vocabulary sizeof the participants.A correlation index or a correlationcoefficient is generally interpreted following a rule of thumb for interpreting the size of correlation coefficient by Hinkle et al (1998, p.120) as presented in Table 4.3 below.

Table 4.3: Rule of thumb for interpreting the size of correlation coefficient (Hinkle et al,1998, p.120)

Size of Correlation	Interpretation
0.90 to 1.00 (-0.90 to -1.00)	Very high correlation
0.70 to .90 (-0.70 to -0.90)	High correlation
0.50 to 0.70 (-0.50 to -0.70)	Moderate correlation
0.30 to 0.50 (-0.30 to -0.50)	Low correlation
0.00 to 0.30 (0.00 to -0.30)	Little if any correlation

From Table 4.2, it can be seen that the scores of all YN test versions have statistically significant correlation with those of the translation test at 0.01 level (*p* = .000), suggesting that all the YN tests created for this study could be good tools for predicting vocabulary size of the participants.

The scores of all YN test versions also have strong correlation with the translation scores, ranking from the highest correlation of .947of Test 1 to the lowest of .687 of Test 9.According to a rule of thumb for interpreting the size of correlation coefficient (Table 4.3), many of these YN test versions yield very high correlation between the YN scores and the translation scores ,i.e., Test 1 (.947), Test 4 (.931), Test 5 (.930), Test 7 (.922), Test 6 (.915), and Test 8 (.903). The othertests that yield high correlation between the YN scores and translation scores are Test 2 (.869), Test 11 (.841), Test 10 (.828), Test 3 (.799), and Test 12 (.745), while the only test that yields moderate correlation is Test 9 (.687). This suggests that, among the 12 YN test versions, Test 1, 4, 5, 7, 6, and 8 could have very high predictive power for the actual vocabulary size of test takers; Test 2, 11, 10, 3, and 12 could have highwhile Test 9 could have moderate.

4.1.3 Translation results of the 12 groups of participants

The translation test in this study is the tool that can give us information about the actual vocabulary size of the participants. The comparison of the mean translation scores of the 12 groups of participants can tell us whether or not each group has different vocabulary size level. Ideally, each group of participants doing each YN test version should not be different in their vocabulary sizeso that the comparison of the 12 YN test versions in this study will be reliable. In this present study, the grouping of 50 students by O-NET scores of English in order to have the same proportions of higher, middle, and lower English proficiency students in each group was successful. That is, the mean translation scores of each group are not significantly different (*p*> .05) as can be seen from Table 4.4 below. From this table, the mean translation scores rank from the highest at 49.68% of Test 9 to the lowest at 43.08% of Test 12.

Table 4.4: The mean translation scores of 12 groups of the participants

Rank	Group of the participants	Number of participants	Mean of translation scores (percent)	SD	Sig.
1	9 (R2N1P3)	50	49.68	23.19	.941
2	2 (R1N1P2)	50	49.31	22.83	
3	7 (R2N1P1)	50	48.84	24.78	
4	6 (R1N2P3)	50	48.51	23.08	
5	5 (R1N2P2)	50	48.08	21.89	
6	4 (R1N2P1)	50	47.84	24.06	
7	1 (R1N1P1)	50	47.04	25.95	
8	10 (R2N2P1)	50	45.52	24.34	
9	8 (R2N1P2)	50	45.49	21.27	
10	3 (R1N1P3)	50	45.24	21.14	
11	11 (R2N2P2)	50	44.14	22.13	
12	12 (R2N2P3)	50	43.08	22.04	
Total		600	46.90	22.98	

4.1.4 Research Question 1: Do different test rubrics bring about significant difference in the test results?

In this study, R1 tests are the tests that employ Rubric 1 *(Write Y (yes) if you **know the meaning** of the word)*, which are Test 1-6, while R2 tests are those that use Rubric 2 *(Write Y (yes) if you think that the word **exists** in English)*, which are Test 7-12.

In order to answer Research Question 1, the means of the YN test scores of R1 tests and R2 tests were compared. Also, the correlationvalues between the YN and translation scores of these two rubric test types were calculated.Furthermore, the results of Interview Question 1, 2, 3, 4, and 5 were reported as more insightful details for Research Question 1.

4.1.4.1 Mean difference and correlation results of R1 and R2 tests

The mean difference of R1 and R2 test types (calculated by t-test) can tell whether or not these 2 test types yield significantly different results, while the correlation coefficients between YN and translation scores were used to find out which test would better predict participants' actualvocabulary size.

The mean difference and correlation results of R1 tests and R2 tests are reported in Table 4.5 and 4.6 respectively.

Table 4.5: Mean difference of R1VS R2tests

YN tests	Number of the participants	Mean	SD	Sig. (2-tailed)
R1 tests	300	55.10	24.35	.831
R2 tests	300	54.67	24.51	

Table 4.6: Concurrent validity results of R1 tests VS R2 tests

YN Test	Number o the participants	Pearson correlation	Sig. (2-tailed)
R1 tests	300	.890**	**.000**
R2 tests	300	.812**	**.000**

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

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

The results show from Table 4.5 that there is no significant difference(p = .831) between the mean YN test scores of R1(55.10)and R2 tests (54.67), suggesting that R1 and R2 tests would not lead to differences in the test scores.

However, when moving on to Table 4.6, we can see that R1 tests yielda little higher correlation coefficient between the YN and translation scores (.890) than that of R2 tests (.812). This suggests that R1 tests may have better tendency to predict the participants' actual vocabulary size than R2.

4.1.4.2 Semi-structured interview results for Research Question 1

<u>Interview Question 1</u>:What were the problems when doing the test or any suggestion to help test takers when doing this kind of test?

This interview question aimed to ask about the general problems the participants found when doing the test and also their suggestions, which may be useful for the improvement of a YN test. The summary of finding of this question is presented in Table 4.7 below.

Table 4.7: Summary of finding from Interview Question 1:What were the problems when doing the test or any suggestion to help test takers when doing this kind of test?

Rank	Answers	Number of the interviewees (out of 72)	Percent
1	The test pattern is ok for me.	67	93.05
2	Writing Y or N is confusing for me; writing ⁄ or × is easier.	4	5.55
3	The letters are too small and there is limited space for writing the translation.	1	1.38

It can be seen from Table 4.7 that most of the interviewees (67 out of 72 or 93.05%) had no problems with the test pattern. However, a few of them (4 out of 72 or 5.55%) pointed out that writing Y or N was confusing while writing or × was easier for them.One interviewed students (1.38%) also reported that the letters in the tests were too small and the space for writing the translation was limited for him.

These interview results suggest that the YN tests used in this study caused not many problems for the test takers. However,in some particular contexts where test takers are more familiar with writing to represent the word *yes* and × to represent the word *no*, these 2 marks (and ×) might be a good alternative for those test takers instead of writing Y or N in a YN test.In addition,some typographical problems such as too small letter or limited space mentioned above are the points that a test writer should also be aware of.

<u>Interview Question 2</u>: Did you read the test rubric?

This interview question intended to investigate whether or not the participants read the test rubric. From this question, we may gain some ideas whether the test results really come from the differences of the test rubrics or from the behavior of not reading the test rubric. The finding of this question is displayed in Table 4.8 as follows.

Table 4.8: Summary of finding from Interview Question 2: Did you read the test rubric?

Answers	R1 test interviewees (out of 36)	Percent	R2 test interviewees (out of 36)	Percent
Yes	33	91.66	34	94.44
No (I consulted friends)	3	8.33	2	5.55

From Table 4.8, we can see that the number of the interviewees from R1 and R2 groups who answered that they read the test rubric are nearly the same (R1 groups = 33 out of 36 (91.66%); and R2 groups = 34 out of 36 (94.44%)). Nonetheless, there were a few interviewees who did not read the test rubric (R1 groups = 3 out of 36 (8.33%); and R2 groups = 2 out of 36 (5.55%)). They reported that, instead of reading the rubric, they asked how to do the test from their friends sitting near them.

From these results, it can be seen that most of the interviewees read the test rubric and only a few did not, which suggests that the behavior of not reading the test rubric may have little effect (if any) to the results of the differences between R1 and R2.

<u>Interview Question 3</u>: Was the test rubric clear for you?

This interview question was designed to explore whether the test rubric is clear for the participants and it may help explain whether the test results really come from the differences of the test rubrics or from the mistakes of the rubric writing.

Table 4.9: Summary of finding from Interview Question 3: Was the test rubric clear for you?

Answers	Reasons	R1 test interviewees (out of 36)	Percent	R2 test interviewees (out of 36)	Percent
Yes	-	35	97.22	35	97.22
No	1. The term “nonword” is not clear for me.	1	2.78	-	-
	2. The phrase “the word that do not exist in English” is not clear ; “the word that exist in an English dictionary” will be clearer for me.	-	-	1	2.78

It can be seen from Table 4.9that most of the interviewees (35 out of 36 [97.22%] from R1 groups and the same number from R2 groups) stated that the rubric was clear for them. Nonetheless, there were a few who said that the rubric was not clear for them (1 out of 36 [2.78%] from R1 groups and also the same number from R2 groups). The interviewee from R1 group reported that the term “nonword” was not clear for him while the one from R2 group stated that the phrase “the words that do not exist in English” was not clear and it would be better if the test said, “the words that exist in an English dictionary”.

From these results, we can see that the 2 test rubrics were clear for most interviewees but not clear for only a few of them. This suggests that the writing of the test rubric may have little effect (if any) to the results of R1 and R2 differences.

However, the point of the unclear wording like “nonword” in Rubric 1 as reported from an interviewee should be taken into consideration. That is, the definition or example of the term “nonword” should be added to make the rubric clearer. Also, for Rubric 2, the phrase “the words that do not exist in English” could be made clearer by adding parentheses containing the phrase “the words that exist in an English dictionary” as suggested by the interviewee from R2 group mentioned earlier.

***Interview Question 4*: How did you interpret the test rubric? Please explain.**

This interview question was designed to investigate whether each test rubric was interpreted in the same direction or not. This information can help explain whether the participants really understand each test rubric. The summary of finding is shown in Table 4.10 below.

Table 4.10: Summary of finding from Interview Question 4:How did you interpret the test rubric? Please explain.

Answers		R1 test interviewees (out of 36)	Percent	R2 test interviewees (out of 36)	Percent
1) I should answer Y (yes) if…..	1.1) I know the meaning of the word.	35	97.22	0	0
	1.2) I have seen/ heard/ or felt familiar with the word.	1	2.78	36	100
2) I should answer N (no) if…..	2.1) I do not know the meaning of the word.	35	97.22	0	0
	2.2) I have not seen/ heard/ or felt familiar with the word.	1	2.78	36	100

From Table 4.10, it can be seen that most of R1 test interviewees (35 out of 36 or 97.22%) interpreted Rubric 1 that they should answer Y (yes) if they know the meaning of the word, and answer N (no) to the word they do not know the meaning of. However, one of them (2.78%) interpreted that he should answer Y (yes) ifhe has seen/

heard/ or felt familiar with the word, and answer N (no) ifhe has not seen/ heard/ or felt familiar with the word.

This suggests that most of R1 test interviewees could interpret R1 correctly while only one wrongly interpret it and the wrong interpretation may have only little effect (if any) to the results of R1 tests.

As for R2 test type, allR2 test interviewees (36 out of 36 or 100%) interpreted Rubric 2 that they should answer Y (yes) to the word they have seen/ heard/ or felt familiar with and answer N (no) to the word they have not.This suggests that wrong interpretation of R2 might not occur and, therefore, might lead to no or little effect to the results of R2 tests.

<u>Interview Question 5</u>*: *Could you tell me frankly what criteria you used to answer Y (yes) or N (no) for each word?

This interview question aimed to explore what are the real criteria the participants used when doing the tests and it may reveal their guessing behavior that might occur during the tests. The summary of finding is illustrated below in Table 4.11.

Table 4.11: Summary of finding from Interview Question 5:Could you tell me frankly what criteria you used to answer Y (yes) or N (no) for each word?

Answer	Criteria	R1 test interviewees (out of 36)	%	R2 test interviewees (out of 36)	%
1) Answering Y (yes) when…..	1.1) knowing the meaning of the word	36	100	36	100
	1.2)having seen/ heard/ or felt familiar with the word	34	94.44	36	100
	1.3) making a guess	6	16.67	18	50.00
2) Answering N (no) when…..	2.1)not having seen/ heard/ or felt familiar with the word.	36	100	36	100
	2.2)being unsure whether the word existed.	9	25.00	1	2.78
	2.3)notknowing the meaning of the word	2	5.55	0	0

Please note that, from Table 4.11, one interviewee could have more than one answer. This is because an interviewee may have more than one criterion for answering Y or N to a word. That is, some students may have up to 3 criteria to answer Y to a word, e.g., they will answer Y 1) if they know the meaning of the word; 2) if they have seen/heard/or felt familiar with the word (i.e. know that the word exist but do not know its meaning or are not sure about its meaning); and 3)if they guess. Some may have two out of the mentioned criteria, and the others may have only one. Therefore, one answer will come from up to 36 participants as can be seen from 4.11 above.

From the table, we can see that, to answer Y to some words, all the interviewees of R1 and R2 group (72 out of 72 or 100%) usedCriterion 1.1 (knowing the meaning of the word).And all of the R2 test interviewees also used Criterion1.2 (having seen/heard/or felt familiar with the word), while not all R1 test interviewees (34 out of 36 or 94.44%) used it. In addition,6 of R1 test interviewees (16.67%) and up to 18 of R2 test interviewees (50%) reported that they also guessed when doing the tests (Criterion 1.3).

It can be seen that the interviewees of both R1 and R2 groups used all the 3 mentioned criteria to answer Y to some words. However, R1 test interviewees used Criterion 1.2 (having seen/heard/or felt familiar with the word) and Criterion 1.3 (guessing) less than did R2 test interviewees. This suggests that R2 tests may lead to more guessing behavior of the test takers than R1 tests.

For the criteria of answering N (no) to some words, all the interviewees of both R1 and R2 groups (72 out of 72 or 100%) used Criterion 2.1 (a word has not been seen/heard/or familiar with). Some of the R1 test interviewees (9 out of 36 or 25%) and one of the R2 test interviewees (2.78%) stated that they also used Criterion 2.2 (being unsure whether the word exists) to answer N (no).For Criterion 2.3 (not knowing the meaning of the word although knowing that the word exists), only 2 of R1 group interviewees (5.55%) reported the use of it while none of R2 group (0%) used it.

It can be seen that all the interviewees of both R1 and R2 groups have similar criteria of rejecting a wordthat they had not seen or heard (Criterion A). However, more interviewees of R1 group than those of R2 group rejected a wordusing Criterion B(beingunsure whether the word exists) and Criterion C (not knowing the meaning of the word although knowing that the word exists). It is possible to say that the interviewees who used these 2 criteria may have more conservative response style. Therefore, using more of these 2 criteria of R1 test interviewees also suggests that R1 may lead to less guessing behavior than R2.

4.1.5 Research Question 2: Do different types of nonwords bring about significant difference in the test results?

In this study, N1 tests are the tests that employ Nonword Type 1(near-homophone nonwords), which are Test 1, 2, 3, 7, 8, and 9, while N2 tests are those consisting of Nonword Type 2 (non-homophone nonwords), which are Test 4, 5, 6, 10, 11, and 12.

In order to investigate whether different types of nonwords results in different YN test scores, the mean differenceof the YN test scores of N1 tests and N2 tests were calculated and the correlation coefficients of these 2 nonword tests types' scores with the translation scores were compared. The results of Interview Question 6, 7, 8, and 9 were reported to provide more understanding of why the results of Research Question 2 come out in a particular way.

4.1.5.1 Mean difference and correlation results of N1 and N2 tests

The mean difference of the N1 and N2 tests (calculated by t-test) can give us an idea whether these 2 test types yield significantly different results. Also, thecorrelation coefficients between YN test scores and translation scores of the 2 test types are compared to find out which test would better predict participants' vocabulary size.

The mean difference and correlation results of N1 tests and N2 tests are illustrated in Table 4.12 and 4.13 respectively.

Table 4.12: Mean difference of N1 VS N2 tests

YN tests	Number of the participants	Mean YN test scores	SD	Sig. (2- tailed)
N1 tests	300	52.81	25.54	**.037***
N2 tests	300	56.96	23.09	

** The mean difference is significant at the 0.05 level*

Table 4.13: Concurrent validity results of N1 VS N2 tests

YN Test	Number of participants	Pearson correlation	Sig. (2-tailed)
N1 tests	300	.858**	**.000**
N2 tests	300	.858**	**.000**

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

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

From Table 4.12, we can see that the mean YN test scores of N1 (52.81)and N2 tests (56.96)are significantly different (*p* = .037), suggesting that the 2 types of nonwordsare likely to bring about different test results. However, from the data of Table 4.13, it can be seen that the correlation coefficients of the YN test scores with the translation scores of N1 and N2 tests are the same (i.e., .858 and .858), which suggests that N1 and N2 testscould be equivalent in predicting the participants' actual vocabulary size in the aspect of ranking.

However, in the aspect of the closeness of YN scores to the translation scores, N1 are closer to the translation scores than those of N2. (See Table 4.14 and 4.15 below)

Table 4.14: Mean translation scores of N1 VS N2 tests

YN tests	Number of the participants	Mean translation scores	SD	Sig. (2- tailed)
N1 tests	300	47.60	23.14	.455
N2 tests	300	46.20	22.85	

It can be seen from Table 4.14 that the mean translation score of N1 tests (47.60) is not significantly different from that of N2 (46.20)(*P* = .455), suggesting that these 2 groups of participants are similar in their actual vocabulary knowledge.

However, as addressed earlier, the mean of N1 YN test scores (52.81) is closer to that of N1 translation scores (47.60) (i.e. the mean difference of 5.21) when compared to the mean YN test score of N2 (56.96) with the mean N2 translation score (46.20) (i.e. the mean difference of 10.76). This suggests that N1 could lead to closer YN scores to the translation scores than N2. Table 4.15 presents the mean difference between YN test scores and translation scores of N1 and N2 tests.

Table 4.15: Mean difference between YN test scores and translation scores of N1 and N2 tests.

YN tests	N	Mean YN test score	Mean translation score	Mean difference
N1 tests	300	52.81	47.60	5.21
N2 tests	300	56.96	46.20	10.76

4.1.5.2 False alarm rate of N1 and N2 tests

As mentioned earlier, false alarm means an answer *yes* to a nonword and the calculation of YN scores for this study is to subtract the false alarm rate from the hit rate (i.e. an answer *yes* to a real word).The results showed that N1 and N2 tests yielded different false alarmrate as illustrated in Table 4.16 below.

Table 4.16: Mean difference of false alarm rate of N1 and N2 tests

YN tests	Number of the participants	Mean false alarm rate (%)	SD	Sig. (2- tailed)
N1 tests	300	23.22	17.38	**.036***
N2 tests	300	20.08	19.29	

** The mean difference is significant at the 0.05 level*

From Table 4.16, we can see that false alarm rate of N1 tests (23.22%) is significant higher than that of N2 tests (20.08%) (p = .036). This suggests that N1 were likely to be chosen more than N2 and this could explain why the participants who did N1 tests get lower YN scores (the mean of 52.81%) than those who did N2 tests (the mean of 56.96%) even though they have similar translation scores (the mean of 47.60% for N1 group and 46.20% for N2 group).

Table 4.17 below shows the false alarm rate of 50 N1 nonwords and 50 N2 nonwords ranking from the most popular to the least.

Table 4.17: False alarm rate of 50 N1 nonwords and 50 N2 nonwords

Rank	N1	N1 (%)	N2 (%)	N2	Rank
1	persent	66.0	70.7	metter	1
2	retern	63.5	49.0	chaster	2
3	serect	61.0	35.5	offine	3
4	conferm	56.0	32.5	fictor	4
5	westurn	44.0	32.0	discuit	5
6	cornor	43.0	32.0	disploy	6
7	deley	42.0	31.0	pariod	7
8	shapter	42.0	30.0	pergent	8
9	digree	39.7	28.0	porson	9
10	rabel	39.0	27.5	depair	10
11	cercle	38.5	27.5	shoilder	11
12	survay	37.3	27.0	cillage	12
13	teble	31.0	26.0	confarm	13
14	selious	30.0	25.5	morder	14
15	purshase	29.0	25.0	mathine	15
16	displey	27.0	25.0	wostern	16
17	sistem	26.5	24.5	morror	17
18	advanse	25.0	24.0	sinnal	18
19	bigin	24.5	24.0	socret	19
20	bisguit	23.3	23.5	surchase	20
21	marget	22.0	23.0	toble	21
22	meybe	22.0	19.0	canbel	22
23	retter	21.0	19.0	gertain	23
24	mirrer	21.0	19.0	sedect	24
25	trousors	18.3	18.8	carcle	25
26	factur	18.0	17.5	sedious	26
27	merder	18.0	17.0	porry	27
28	sertain	17.0	16.5	serdice	28
29	mashine	17.0	16.0	cirner	29
30	prefur	16.0	16.0	retorn	30

Table 4.17: False alarm rate of 50 N1 nonwords and 50 N2 nonwords (Cont.)

Rank	N1	N1 (%)	N2 (%)	N2	Rank
31	purson	16.0	14.5	mardet	31
32	sicnal	16.0	14.3	troisers	32
33	vorry	15.0	14.0	adhance	33
34	willage	15.0	13.7	dervash	34
35	shannel	14.7	12.5	prafer	35
36	wictim	14.5	12.0	chanbel	36
37	piriod	14.0	12.0	musin	37
38	traffig	14.0	12.0	traffin	38
39	sicret	13.5	11.3	lindow	39
40	lepair	13.0	11.0	bugin	40
41	cansel	13.0	11.0	pabel	41
42	durvish	12.0	10.5	bictim	42
43	bersurk	11.0	10.3	survoy	43
44	serwice	10.0	10.0	bersork	44
45	sistor	10.0	10.0	rinish	45
46	finich	9.5	10.0	sustem	46
47	vindow	9.0	8.0	deloy	47
48	musig	9.0	7.3	dogree	48
49	offise	8.5	3.5	moybe	49
50	shouldir	5.5	2.0	soster	50

From Table 4.17, the overall picture of how the participants chose N1 and N2 is displayed. We can see from the table that N1 are more popular than N2 in almost every rank except 1st rank (N2 (70.7%)> N1 (66.0%)); 20th (N2 (23.5%)> N1 (23.3%)); 21st (N2 (23.0%)> N1 (22.0%)); 25th(N2 (18.8%)> N1 (18.3%)); and 46th (N2 (10.0%)> N1 (9.5%)). This suggests that N1s could be more difficult to be rejected than N2s.

4.1.5.3 Semi-structured interview results for Research Question 2

<u>Interview Question 6</u>: What were the reasons you answered Y (yes) to some real words that you could not translate afterward?

This interview question was designed to draw the information about the reasons the participants could not translate some real words that they answered Y (yes). This question may give an insight of their overestimation on the real words. Table 4.18 below illustrates the summary of finding of this interview question.

Table 4.18: Summary of finding from Interview Question 6:What were the reasons you answered Y (yes) to some real words that you could not translate or translated it wrongly afterward?

Reasons	N1 test interviewees (out of 36)	Percent	N2 test interviewees (out of 36)	Percent
1. I had seen/heard/or felt familiar with the word.	35	97.22	35	97.22
2. I made a guess.	10	27.77	14	38.88

Please note that, like Interview Question 5, one interviewee could have more than one reason for this interview question. Thus, each reason could come from up to 36 interviewees per group.

From Table 4.18, it can be seen that the same number of the interviewees from both N1 and N2 groups (35 out of 36 or 97.22%) reported that the reason they answered Y to the real words they could not translate afterward was that they had seen/heard/or felt familiar with the words. In addition, 10 out of 36 (27.77%) of N1 group and 14 out of 36 (38.88%) of N2 group stated that they also made a guess to some real words.

This suggests that their overestimation on the real words couldbe based upon both their word familiarity and their blind guessing.

<u>Interview Question 7</u>:What were the reasons you answered Y (yes) to some nonwords?

This interview question aimed to draw the information about why the participants decided to answer ‘yes’ to some nonwords.Initially, it was intended to ask the 72 interviewees, but because the information on the translation test papers could be a good source to answer this question so the researcher decided to use the information on the paper instead. That is, on the translation test paper, the participants were told that they had to translate all the words (and nonwords) they answered Y. If they could not translate them, they had to write, on each word, the reason why they could not translate them. Therefore, the results in Table 4.19 below come from 600 papers instead of the 72 interviewees.

Table 4.19: Summary of finding from Interview Question 7:What were the reasons you answered Y (yes) to some nonwords?

Answers	Details	N1 test participants (out of 300)	N2 test participants (out of 300)
1. Having some partial knowledge on the nonword	1.1 Confusing the nonword with another word by giving a wrong translation	43.24%	13.27%
	1.2 Having seen/heard/or felt familiar with the nonword.	21.01%	11.16%
	Total	**64.25%**	**24.43%**
2. Guessing	-	34.84%	75.28%
3. Misreadingthe nonword	-	0.90%	0.29%

Please also note that, from Table 4.19, one participant could have more than one reason for this question. Therefore, each reason could come from up to 300 participants.

From the table, we can see that there were 3 reasons for answering Y to the nonwords. Reason 1 is the participants had some partial knowledge on the nonword. That is, 1.1) they confused the nonword with another word by giving a wrong translation; and 1.2) they confused the nonword with another word claiming that they had seen/ heard/ or felt familiar with the nonword. Reason 2 is they made a guess while Reason 3 is they misread the nonword (i.e. the participants answered Y to a particular nonword in the YN test, but when they did the translation test, they found out that they chose the nonword because they misread it and they wrote this reason on the test paper (beside that particular nonword).

All the mentioned reasons were reported by the participants themselves on the test papers except Reason 1.1, which was summarized from the evidence that they gave translation to the nonwords.

It can be seen that the first reason (64.25%) for choosing N1 is Reason 1 (having some partial knowledge on the nonword), the second (34.84%) is Reason 2 (guessing), and the third (0.90%) is Reason 3 (misreading the nonwords).

However, the first reason (75.28%) for choosing N2 is Reason 2 (guessing), the second (24.15%) is Reason 1 (confusing the nonwords with another word),and the third (0.29%) is Reason 3 (misreading the nonwords).

As can be seen, the first reason for choosing N1 is the participants had some partial knowledge on the nonwords, while the first reason for choosing N2 is they made a guess.This suggests that N1 may lead to more confusion with another real word than N2, while those who chose N2 tend to be those who guessed more than those who had partial knowledge on real words and nonwords (See Table 4.20 and 4.21 for the examples of their confusion caused by having only partial knowledge on real words and nonwords).

In addition, there is a report of both N1 and N2 misreading but at a low percentage (N1=0.90%, N2 = 0.29%), suggesting that the misreading of the nonwords may have little effect (if any) to the results of this nonword study.

Table 4.20 and 4.21 below presents the examples of translations given to the 2 most popular and the 2 least popular N1 and N2.

Table 4.20: Examples of N1 translations

Rank	Nonwords (N1)	Translation	Percent
1	persent	นำเสนอ (present(verb))	53.40%
		เปอร์เซ็น (percent)	22.33%
		ปัจจุบัน (at present)	7.77%
		ของขวัญ (gift, present (noun))	3.88%
		มาเข้าเรียน(to be present)	1.94%
		บุคคล (person)	1.94%
		ผู้ปกครอง (parents)	1.94%
		บำบัด (therapy)	0.97%
		พบ,เจอ (meet, encounter)	0.97%
		ขั้นสูง (high level)	0.97%
		ส่ง (submit)	0.97%
		สัดส่วน (proportion)	0.97%
		แสดง (show)	0.97%
		อธิบาย (explain)	0.97%

Table 4.20: Examples of N1 translations (Cont.)

Rank	Nonwords (N1)	Translation	Percent
2	retern	กลับมา (return)	94.78%
		ด้านบน (top)	0.87%
		ตอบ (answer)	0.87%
		ทำซ้ำ (repeat)	0.87%
		รีเทอร์ (r-e-t-u-r)	0.87%
		เริ่มใหม่ (begin)	0.87%
		อีกครั้ง (again)	0.87%
49	musig	ดนตรี (music)	77.77%
		ขบขัน (funny)	11.11%
		จามไอ (cough)	11.11%
50	shouldir	ความหมาย (meaning)	50.00%
		รองเท้าส้นสูง (high-heeled shoes)	50.00%

Table 4.21: Examples of N2 translations

Rank	Nonwords (N2)	Translation	Percent
1	metter	เมตร (meter : as a unit of length)	43.25%
		เกิดอะไรขึ้น (What's the matter?)	11.70%
		ดี (good)	4.50%
		พบปะ (meet)	3.60%
		เรื่องราว, เหตุการณ์ (issue/matter)	3.60%
		มิเตอร์ (meter : as a measuring device)	2.70%
		ปัญหา (problem)	2.70%
		วัตถุ (matter/object)	2.70%
		กลาง (middle)	2.70%
		มาก (many, a lot)	1.80%
		อาการ (symptom)	1.80%

Table 4.21: Examples of N2 translations (Cont.)

Rank	Nonwords (N2)	Translation	Percent
		กว่า (more/better)	1.80%
		อากาศ (weather)	1.80%
		แข่งขัน (compete)	0.90%
		คนขายเนื้อ (butcher/meat seller)	0.90%
		เครื่องกล (machine, motor)	0.90%
		จดหมาย (letter)	0.90%
		เซนติเมตร (centimeter)	0.90%
		ดีที่สุด (best)	0.90%
		ติด (stick)	0.90%
		เนื้อ (meat)	0.90%
		บังคับ (force)	0.90%
		ผิดปกติ (abnormal)	0.90%
		พอดี (fit)	0.90%
		มี (have)	0.90%
		แม้ว่า (even though)	0.90%
		เล็กน้อย (not much, little. few)	0.90%
		สาระ (content)	0.90%
2	chaster	เชสเตอร์กิล (Chester Grill)	28.58%
		คุยสื่อสาร (chat, communicate)	14.29%
		ถ่ายภาพติด (shutter)	14.29%
		สะสม (collect)	14.29%
		หน้านี้ (this page)	14.29%
		หมากรุก (chess)	14.29%
49	moybe	may be (may be)	100%
50	soster	-	-

From Table 4.20 and 4.21, it can be seen that the 2 most popular N1 and N2 tend to be those that caused a lot of confusion to the participants, but the 2 least popular are likely to be those that caused a little.

For example, there were about 14 different translations given to the most popular N1*persent*such as *present* as a noun, a verb, and an adjective (66.99%), *percent* (22.33%), *person* (1.94%), *parents* (1.94%), and many others (6.80%). While the least popular N1 *shouldir* was given only 2 different translations as *meaning* (50%) and *high-heeled shoes*(50%).

For the most popular N2 *metter*, there were about 29 different translations given to this nonword such as *meter* as a unit of length (43.25%), *matter* as in the sentence "What's the matter?" (11.70%), *meet* (3.60%), *meter* as a measuring device (2.70%), *middle* (2.70%), *better* (1.80%), *weather* (1.80%), *letter* (0.90%), *meat* (0.90%), and many others (30.65%). While the least popular N2 *soster* was given no translation at all (i.e. the participants answered Y to this nonword but gave no translations).

These examples suggest that the participants tend to confuse the nonwords with other words that have similar spelling or sound to the nonwords. (e.g. *metter* was confused with *meter*, *matter*, *meet*, *middle*, *better*, and etc.) However, the other translations that seems to have no link to any word such as *meaning* and *high-heeled shoes* given to the nonword *shouldir* may be the results of blind guessing or another kind of confusion that needs further investigation.

***Interview Question 8**: What were the reason you answered N (no) to some nonwords?*

This interview question intended to draw the information about why the participants decided to answer N (no) to some nonwords. The information may help explain the thinking process or the guessing behavior of the participants in this present study.

Table 4.22: Summary of finding from Interview Question 8:What were the reasons you answered N (no) to some nonwords?

Reasons	N1 test interviewees (out of 36)	Percent	N2 test interviewees (out of 36)	Percent
1. I knew that it was a nonword.	36	100	36	100
2. I was not sure whether it was a real word or a nonword.	6	16.67	4	11.11
3. I did not know that it was a nonword. (I thought it was a real word that I had never seen or heard.)	16	44.44	18	50.00

Also for this interview question, one interviewee could have more than one reason. Hence, each reason could come from up to 36 interviewees.

From Table 4.22, we can see that all of the interviewees both from N1 and N2 groups (36 out of 36 or 100%) stated that they answered N to a nonword when they exactly knew that it was a nonword. While some of them (6 out of 36 (16.67%) for N1 group; and 4 out of 36 (11.11%) for N2 group) also reported that they answered N when they were not sure whether it was a realword or a nonword.However, some of the interviewees from N1 group (16 out of 36 or 44.44%) and N2 group (18 out of 36 or 50%) addressed that they answered N to a nonword because they thought that it was a real word that they had never seen or heard (i.e. they did not know that it was a nonword).

It can be seen that the 3 reasons mentioned above came from not very different number between N1 and N2 group interviewees, suggesting that both N1 and N2 may be rejected based upon similar basis mentioned earlier.

***Interview Question 9*: Could you identify the nonwords from the real words and how?**

This interview question aimed to explore whether it was easy or difficult for the participants to differentiate between a real word and a nonword and why. Table 4.23 below presents the summary of finding of this interview question.

Table 4.23: Summary of finding from Interview Question 9:Could you identify the nonwords from the real words and how?

Answers	Details	N1 test interviewees (out of 36)	Percent	N2 test interviewees (out of 36)	Percent
Yes	The nonwords were created by changing some letters from their original words.	31	86.11	23	63.88
No	-	5	13.88	13	36.11

From Table 4.23, we can see that most interviewees both from N1 (31 out of 36 or 86.11%) and N2 groups (23 out of 36 or 63.88%) could identify the nonword from the real word explaining that the nonwords were created by changing some letters from their original words. However, some interviewees (5 out of 36 (13.88%) from N1 group; and 13 out of 36 (36.11%) from N2 group) reported that they had no idea how the nonwords were created. This suggests that most participants tend to have awareness on how to identify a nonword from a real word when doing the tests. In addition, the interviewees from N1 group has higher percentage (86.11%) on this awareness than those from N2 group (63.88%), suggesting that it may be easier to find the pattern of N1 than that of N2 possibly because N1 (near-homophone nonwords) have more similar sound to their original words than N2 (non-homophone nonwords).

4.1.6 Research Question 3: Do different proportions of nonwords to real words bring about significant difference in the test results?

For this proportionstudy, the YN tests are divided into 3 types according to the different proportions of real words and nonwords. P1 tests (Proportion 1 tests) use 50 real words and 50 nonwords, which are Test 1, 4, 7, and 10. P2tests (Proportion 2 tests) use 67 real words and 33 nonwords, which are Test 2, 5, 8, and 11; while P3 tests (Proportion 3 tests) use 90 real words and 10 nonwords, which are Test 3, 6, 9, and 12.

In order to answer Research Question 3,the mean difference of the YN test scoresof P1, P2, and P3 tests were compared and the correlation between the YN and translation scores of these 3 proportion tests were calculated.Also, the results of Interview Question 10 were reported to give deeper details for this research question.

4.1.6.1 Mean difference and correlation results of P1, P2, and P3 tests

The mean difference of P1, P2, and P3 tests (calculated by one-way ANOVA) can tell whether these 3 test types yield significantly different results, while the correlation coefficient between YN and translation scores can give us an idea which test would better predict participants' actual vocabulary size.

The mean difference and correlation results of P1, P2, and P3 tests are displayed in Table 4.24 and 4.25 respectively.

Table 4.24: Mean difference of P 1, P2, and P3 tests

YN tests	Number of the participants	Mean	SD	Sig.
P1 tests (50:50)	200	53.40	26.18	.301
P2 tests (67:33)	200	57.01	24.11	
P3 tests (90:10)	200	54.24	22.82	

Table 4.25: Concurrent validity results of P1, P2, and P3 tests

YN Test	No. of the participants	Pearson correlation	Sig. (2-tailed)
P1 tests	200	.903**	**.000**
P2 tests	200	.871**	**.000**
P3 tests	200	.773**	**.000**

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

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

From Table 4.24, it can be seen that there is no significant difference between the mean YN test scores of P1, P2, and P3 tests (p = .301), suggesting that P1, P2, and P3 tests would not lead to differences in the test scores.

Nevertheless, as can be seen from Table 4.25, P1 tests yield the highest correlation coefficient between the YN and translation scores (.903**), while P2 tests yield the second (.871**), and P3 the third (.773**). This suggests that, among the 3 proportion tests, P1 tests may have the best tendency to predict the participants' actual vocabulary size, while P2 the second best, and P3 the third.

4.1.6.2 Semi-structured interview results for Research Question 3

<u>Interview Question 10</u>: How many nonwords did you think were in the YN test?

This interview question may give us more insight about whether the participants have any idea about the proportion of the nonwords. This is because if they knew the proportion, it may be possible that they made more guesses. For example, if a participant thought that the proportion might be 50 real words : 50 nonwords, they might count their Y (yes) answers and their N (no) answers. And if the answer Y was 45 and the answer N 55 (each YN test consists of 100 items), they might try to answer 5 more Y by guessing in order to have 50 Y answers. This phenomenon may help explain why the results of this study come up in a particular way.

Table 4.26: Summary of finding from Interview Question 10:How many nonwords did you think were in the YN test?

Answers	P1 test interviewees (out of 24)	%	P2 test interviewees (out of 24)	%	P3 test interviewees (out of 24)	%
1. I did not think about it at the time I did the test.	20	83.33	18	75.00	21	87.50
2. I think there were more nonwords than real words.	1	4.17	1	4.17	-	-
3. I think there were more real words than nonwords.	-	-	2	8.33	2	8.33
4. I think the nonwords accounted for 20%.	-	-	-	-	1	4.17
5. I think the nonwords accounted for 30%.	1	4.17	-	-	-	-
6. I think the nonwords accounted for 40%.	1	4.17	-	-	-	-
7. I think the nonwords accounted for 50%.	1	4.17	3	12.50	-	-

From Table 4.26, we can see that most interviewees (20 out of 24 or 83.33% from P1 group; 18 out of 24 or 75.00% from P2 group; and 21 out of 24 or 87.50% from P3 group) stated that they did not think about the proportion of the real words and the nonwords at the time they did the tests.

As can be seen, only one interviewee (out of 72) gave the exactly right answer that the test he did (a P1 test) contained 50% nonwords while some had only rough ideas about the test proportion and the others even had a wrong picture of the test proportion. For instance, 2 out of 24 (8.33%) from P2 group and also 2 out of 24 (8.33%) from P3 group thought that there were more real words than nonwords, suggesting that they had only rough idea about the proportion. While 1 out of 24 (4.17%) from P1 group and also 1 out of 24 (4.17%) from P2 group thought that there were more nonwords than real words, suggesting that they had a wrong picture of the proportion in the YN tests they did.

To summarize, it can be seen that most of the interviewees had no ideas about the proportion when they did the tests, while only 1 out of 72 had an exactly right idea, and the others had only rough or even wrong ideas. This suggests that the guessing behavior that might come from knowing the test proportion may have little effect (if any) to the results of this study.

4.2 Summary

All the information in this chapter is an attempt to provide the results of this present study, starting from the reliability results of the 12 YN tests, followed by the concurrent validity results, the translation results,the results of the 3 research questions and ofthe semi-structured interview.

In the part of the reliability test (using Cronbach's Alpha), the results showed that the 12 YN testswere reliable ranking from .881 of Test 12 (R2N2P3) to .956 of Test 3 (R1N1P3).(The acceptable reliability is at .700).

For the concurrent validity results, which come from the correlation between YN and translation scores, it was found that the scores of all YN test versions have statistically significant correlation with those of the translation test at 0.01 level (p = .000), suggesting that all the YN tests created for this study could be good tools for predicting vocabulary size of the participants ranking from the highest correlation coefficient of .947** of Test 1 (R1N1P1) to the lowest of .687** of Test 9 (R2N1P3).

It was also found that the mean translation scores of each YN test version participants (50 participants per test) were not significantly different (means = 46.90% of 600 participants), which suggests that each group tends not to be different in their vocabulary size, and means that the comparison of each YN test version would be reliable because of the participants' equivalent vocabulary size.

As for the results of Research Question 1 (Do different test rubrics bring about differences in the test results?), it was found that R1 tends to be better than R2 in predicting the actual vocabulary size of the participants because R1 tests yield slightly higher correlation coefficient than R2 (.890** and .812** respectively).

In the part of Research Question 2 (Do different types of nonwords bring about differences in the test results?), the results suggest that N1 tests tend to be better than N2 tests because N1 tests could lead to closer scores to the translation test than N2 (N1=5.21% and N2=10.76% different from the translation test) although the correlation with the translation scores of N1 and N2 tests are the same at .858**.

For Research Question 3 (Do different nonword proportions bring about differences in the test results?), the results suggest that P1 test (*r* = .903**) tend to be the best in predicting the actual VS of a test taker, with P2 the second (*r* = .871**), and P3 the third (*r* = .773**).

In the part of Interview Question 1 (What were the problems when doing the test or any suggestion to help test takers doing this kind of test?), the results suggest that the YN tests used in this study tend not to cause problems to the participants (67 out of 72 or 93.05%). However, some interviewees (4 out of 72 or 5.55%) reported that answering *yes* or *no* was confusing for them while answering ✓ or × was easier, and one interviewees (1.38%) stated that the letters were too small and the space for writing the translation was limited for him.

For Interview Question 2 (Did you read the test rubric?), the results revealed thatmost of the interviewees (R1 tests=91.66%, R2 tests=94.44%) read the test rubric and only a few did not (R1 tests=8.33%, R2 tests=5.55%), which suggests that the behavior of not reading the test rubric may have little effect (if any) to the results of the differences between R1 and R2.

For Interview Question 3 (Was the test rubric clear for you?), the results suggest that the rubric used in this study tend to be clear for most of the interviewees (35 out of 36 (97.22%) from R1 groups and the same number from R2 groups). Only a few reported that the rubric was not clear for them (1 out of 36 (2.78%) from R1 groups and also the same number from R2 groups).

For Interview Question 4 (How did you interpret the test rubric? Please explain.), the results suggest that the 2 rubrics were relatively well understood by most participants because only 1 interviewee out of 72 (1.38%) have a wrong interpretation for R1. That is, he interpreted R1 "Write Y (yes) in front of the word you know the meaning of" as "write Y (yes) in front of the word you have seen or heard".

As for Interview Question 5 (Could you tell me frankly what criteria you used to answer *yes* or *no* for each word?), it was found that to answer Y to some words, all the interviewees of R1 and R2 groups (72 out of 72 or 100%) used Criterion 1 (knowing the meaning of the word). And all of the R2 test interviewees also used Criterion 2 (a word has been seen/heard/or familiar with), while not all R1 test interviewees (34 out of 36 or 94.44%) used it. In addition, 6 of R1 test interviewees (16.67%) and up to 18 of R2 test interviewees (50%) reported that they also guessed when doing the tests (Criterion 3). This suggests that R2 tends to lead to more guessing than R1.

For Interview Question 6 (What were the reasons you answered *yes* to some real words that you could not translate afterwards?), the results suggest that the participants' overestimation on the real words could come from more of partial knowledge overestimation (i.e. they had seen, heard, or felt familiar with the words) than random guessing (97.22% and 33.33% respectively).

In the part of Interview Question 7 (What were the reasons you answered *yes* to some nonwords?), the results suggest that N1 and N2 were chosen for 3 reasons of partial knowledge overestimation, random guessing, and misreading, but at different levels. That is, the first reason for choosing N1 is the participants had some partial knowledge on the nonword, while the first reason for choosing N2 is they made a guess. This suggests that N1 may lead to more confusion with another word than N2, while those who chose N2 tend to be those who guessed more than those who had partial knowledge on the words. Also, both N1 and N2 were chosen for the reason of misreading but at a low percentage (N1 = 0.90%, N2 = 0.29%), suggesting that the misreading of the nonwords may have little effect (if any) to the results of this nonword study.

For Interview Question 8 (What were the reason you answered *no* to some nonwords?), the results showed that both N1 and N2 were similarly rejected for 3 reasons, which are 1) the participants reject N1 or N2 when they knew that it was a nonword (72 out of 72 or 100%); 2) when they were not sure that it was a real word or a nonword (34 out of 72 or 47.22%); and 3) when they did not know that it was a nonword (i.e. they thought it was a real word that they had never seen or heard) (10 out of 72 or 13.88%).

As for Interview Question 9 (Could you identify the nonwords from the real words and how?), the results suggest that most interviewees tend to have awareness on how to identify a nonword from a real word when doing the tests (54 out of 72 or 75%). They reported that they knew the nonwords were created by changing some letters from their original words. Furthermore, the interviewees from N1 group has higher percentage (86.11%) on this awareness than those from N2 group (63.88%), suggesting that it may be easier to find the pattern of N1 than that of N2 possibly because N1 (near-homophone nonwords) have more similar sound to their original words than N2 (non-homophone nonwords).

Finally, on the part of Interview Question 10 (How many nonwords did you think were in the YN test?), the results revealed that most of the interviewees had no ideas about the proportion when they did the tests (59 out of 72 or 81.94%), while only 1 out of 72 (1.38%) had an exactly right idea, and the others had only rough (9 out of 72 or 12.50%) or even wrong ideas (4 out of 72 or or 5.56%). This suggests that the guessing that might come from knowing the proportion of real words and nonwords may have little effect (if any) to the results of this study.

The discussion of the results of this present study will be presented in the following chapter.

CHAPTER 5

DISCUSSION

This chapter repeats some figures from chapter 4 and discusses them according to the research questions.

5.1 The uses of a YN test

It seems that now a YN test might be more proper as a placement tool than as a means to exactly predict the actual vocabulary size of a test taker because the YN test results correlate with the actual vocabulary size but seems not to yield the exact vocabulary size.

In this present study, a translation test is the test of concurrent validity; that is, it is the means to find out the actual vocabulary size of the participants (See section 3.1.4.3 for more details). The correlation between the scores of a YN test and a translation test can tell the concurrent validity of that YN test version. The correlation between the 2 tests means that if a test taker gets a high score in a YN test, we can predict that he will also get a high score in a translation test (suggesting that his actual vocabulary size is large).

However, as mentioned above, although the YN scores correlate well with the translation scores, the YN scores could not be used to predict the actual translation scores (i.e. the scores showing the actual vocabulary size of a test taker). This is perhaps due to the inconsistencies found in different YN test versions now, such as different test rubrics, nonword types, nonword proportions, calculation methods, and etc. In other words, the difference between the YN scores and the actual vocabulary size scores may vary according to each YN test version. For example, if a test taker got an 80% YN test score, we could predict that he would get a high score on the translation test, but the exact translation score is unpredictable- it could be 90%, 85%, or 72%, etc, perhaps according to different YN test versions mentioned earlier.

Therefore, it seems more appropriate to use a YN score for ranking than for telling the actual vocabulary size score at present. However, further studies may find out the way to develop a YN test that can predict the more exact, or closer, actual vocabulary size of a test taker.

To sum up, this present study pays more attention to the ability to rank a test taker than the ability to tell the actual vocabulary size of a YN test. Thus, the correlation coefficient (*r*) between a YN test and a translation test will be the most important indicator for the quality of a YN test version in this study. In other words, a YN score will not be as important in itself as the correlation coefficient between the YN and translation scores in this study.

5.2 Research Question 1: Do different test rubrics bring about differences in the test results?

In this present study, Rubric 1 (R1) tells test takers to answer Y (yes) to the word they *know the meaning* of while Rubric 2 (R2) informs them to answer Y to the word they think *exist* in English. These 2 rubrics seem to require different levels of vocabulary knowledge from test takers. That is, R1 may need more knowledge of the word meaning while R2 may need less (i.e. only having seen the word or even guessing). This led to the doubt whether or not the 2 rubrics yield the same YN test results. Therefore, Research Question 1 was framed to find out whether these 2 rubrics bring about different results.

Please note that, in this present study, the difference between R1 and R2 YN scores (calculated by t-test) can tell whether or not these 2 test types yield significantly different YN test results. However, the correlation between YN and translation scores, which are considered more informative than the YN scores by themselves, were used to find out the quality of each YN test version, i.e., which test would best predict participants' actual vocabulary size.

5.2.1 Why do R1 tests tend to be better in predicting actual vocabulary size of a test taker than R2 tests?

The results suggest that R1 tests are likely to have better quality in predicting actual vocabulary size than R2 tests because the correlation coefficients between the YN scores and the translation scores of R1 tests are higher than that of R2 tests. Table 5.1 below presents the correlation between YN and translation scores of R1 (300 participants) and R2 tests (300 participants) and Table 5.2 reports the correlation between YN and translation scores of the 6 different R1 test versions compared with the 6 different R2 test versions.

Table 5.1: Correlation between YN and translation scores of R1 VS R2 tests

YN Test	**Number of the participants**	**Pearson correlation**	**Sig. (2-tailed)**
R1 tests	300	.890**	**.000**
R2 tests	300	.812**	**.000**

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

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

Table 5.2: Correlation between YN and translation scores of six R1 tests VS six R2 tests

R1 tests			R2 tests		
Test	**N**	**Pearson correlation**	**Pearson correlation**	**N**	**Test**
1 (R1N1P1)	50	**.947****	**.922****	50	7 (R2N1P1)
2 (R1N1P2)	50	**.896****	**.903****	50	8 (R2N1P2)
3 (R1N1P3)	50	**.799****	**.687****	50	9 (R2N1P3)
4 (R1N2P1)	50	**.931****	**.828****	50	10 (R2N2P1)
5 (R1N2P2)	50	**.930****	**.841****	50	11 (R2N2P2)
6 (R1N2P3)	50	**.915****	**.745****	50	12 (R2N2P3)

*R_1 = Test Rubric 1 (Write Y (yes) if you **know the meaning** of the word)*

*R_2 = Test Rubric 2 (Write Y (yes) if you think that the word **exists** in English)*

N_1 = Nonword type 1 (near-homophone nonwords)

N_2 = Nonword type 2 (non-homophone nonwords)

P_1 = Proportion 1 (50 real words : 50 nonwords)

P_2 = Proportion 2 (67 real words : 33 nonwords)

P_3 = Proportion 3 (90 real words : 10 nonwords)

From Table 5.1, when calculated by combining the 6 different R1 tests altogether (300 participants) and also the 6 different R2 tests (300 participants), R1 tests yield a slightly higher correlation coefficient between the YN and translation scores (.890) than that of R2 tests (.812). However, from Table 5.2, when comparing each YN test pair that differs only the rubric, but all other things are equal (i.e. they share the same nonword type and nonword proportion), it can be seen that most R1 test versions yield clearly higher correlation coefficients between the YN and translation scores than those of R2, i.e., Test 1 (.947) higher than Test 7 (.922); Test 3 (.799) higher than Test 9 (.687); Test 4 (.931) higher than Test 10 (.828); Test 5 (.930) higher than Test 11 (.841); and Test 6 (.915) higher than Test 12 (.745).

The reason why R1 tests may better predict actual vocabulary size of test takers than R2 is possibly because R1 tells test takers to write Y (yes) to the word they know the meaning of, which should directly correlate with their translation scores. In other words, claiming to know the meaning of a word seems to strongly suggest the ability to translate a word. Compare this with the YN scores under R2, which says "write Y (yes) in front of the word you think exists in English", the scores may come from 1) words whose meaning is actually known; 2) words they just have seen or heard, but whose meaning is actually unknown; or 3) words that they simply guess about. In other words, the scores of R2 tests may not come from knowing the meaning of the word alone, but may also come from familiarity with the word or even from guessing, which could lead to less correlation between the YN scores and the translation scores compared with R1.

This assumption is also supported by the information obtained from Interview Question 5 (See Table 5.3 below) that R2 may lead to more guessing behavior than R1 and that R1 may influence some participants, though not all, to be more conservative (tend not to guess) when doing the test than R2.

Table 5.3: Summary of finding from Interview Question 5: Could you tell me frankly what criteria you used to answer Y (yes) or N (no) to each word?

Answer	Criteria	R1 test interviewees (out of 36)	%	R2 test interviewees (out of 36)	%
1) Answering Y (yes) when…..	1.1) knowing the meaning of the word	36	100	36	100
	1.2) having seen/ heard/ or felt familiar with the word	34	94.44	36	100
	1.3) making a guess	6	16.67	18	50.00
2) Answering N (no) when…..	2.1) not having seen/ heard/ or felt familiar with the word.	36	100	36	100
	2.2) being unsure whether the word existed.	9	25.00	1	2.78
	2.3) not knowing the meaning of the word	2	5.55	0	0

From Table 5.3, we can see that the interviewees of both R1 and R2 groups used 3 criteria to answer Y (yes) to some words. That is, 1) all of them answered Y to a word when they really know its meaning; 2) most of R1 interviewees (34 out of 36 or 94.44%) and all of R2 (36 out of 36 or 100%) answered Y to a word when they had seen/heard/or felt familiar with the words although they did not know its meaning; and 3) some of them (6 out of 36 or 16.67% from R1 group; 18 out of 36 or 50% from R2 group) also made a guess with the word they had never seen/heard/or felt familiar with.

As can be seen, to answer *yes* to some words, the interviewees from R2 group used Criterion 2 (a word has been seen/heard/or familiar with) and Criterion 3 (guessing) more than those from R1 group, which suggests that R2 tests may lead to more guessing behavior of the test takers than R1 as mentioned earlier.

It can also be seen that, to answer *no* to a word, R1 test interviewees used more of Criterion 5 (not being sure whether the word exists) and 6 (not knowing the meaning of the word although knowing that the word exists) than R2 test interviewees, which suggests that R1 may lead to less guessing behavior than R2. As can be seen, all the interviewees of both R1 and R2 groups have similar criteria of answering *no* to a word that they had not seen or heard (Criterion 4). However, as mentioned above, more interviewees of R1 group (9 out of 36 or 25%) than those of R2 group (1 out of 36 or 2.78%) rejected a word using Criterion 5 (not being sure whether the word exists). Moreover, some R1 group participants (2 out of 36 or 5.55%) use Criterion 6 (not knowing the meaning of the word although knowing that the word exists) to reject a word, while none of R2 group participants did (0%). It may be possible to say that the interviewees who used Criterion 5 and 6 have a more conservative response style (i.e. less guessing). Therefore, using more of these 2 criteria of R1 test interviewees suggests that R1 may lead to less guessing behavior than R2 as mentioned earlier.

5.2.2 Why do R1 and R2 tests lead to significantly different reaction of the test takers to the tests but insignificantly different YN test scores?

The results show that the mean YN scores of R1 (55.10%) and R2 (54.67%) are not significantly different (p = .831) (See Table 5.4 below). However, when looking at the significantly different hit rates (a *yes* answer to a real word) and false alarm rates (a *yes* answer to a nonword) between R1 and R2 tests, it suggests that the test takers have clearly different reaction to R1 and R2 tests in that R1 tends to lead to less guessing than R2. Table 5.4 below shows the difference between the YN scores of R1 and R2 tests while Table 5.5 reports the difference between the translation scores of R1 and R2 tests; Table 5.6 illustrates the difference of the hit rate of R1 and R2 tests; and Table 5.7 presents the difference of the false alarm rate of R1 and R2 tests.

Table 5.4: Mean difference of the YN scores of R1 VS R2 tests

YN tests	**Number of the participants**	**Mean YN scores %**	**SD**	**Sig. (2- tailed) (t-test)**
R1 tests	300	55.10	24.35	.831
R2 tests	300	54.67	24.51	

Table 5.5: Mean difference of the translation scores of R1 VS R2 tests

YN tests	**Number of the participants**	**Mean translation scores (%)**	**SD**	**Sig. (2- tailed) (t-test)**
R1 tests	300	47.67	23.05	.411
R2 tests	300	46.13	22.93	

Table 5.6: Mean difference of the hit rate of R1 VS R2 tests

YN tests	**Number of the participants**	**Mean hit rate %**	**SD**	**Sig. (2- tailed) (t-test)**
R1 tests	300	68.83	19.47	**.000**
R2 tests	300	84.24	12.08	

Table 5.7: Mean difference of the false alarm rate of R1 VS R2 tests

YN tests	**Number of the participants**	**Mean FA rate %**	**SD**	**Sig. (2- tailed) (t-test)**
R1 tests	300	13.73	13.10	**.000**
R2 tests	300	29.56	19.55	

From Table 5.5, we can see that the translation scores of R1 and R2 participants are not significantly different (47.67% and 46.13% respectively, p = .411), suggesting that the participants of the 2 test types are likely to have similar actual VS. However, it can be seen from Table 5.6 that the mean hit rate of R2 tests (84.24%) is significantly higher than that of R1 (68.83%) (p = .000), suggesting that R2 could lead to more *yes* answers to the real words than R1. Similarly, from Table 5.7, R2 participants' false alarm rate (29.56%) is significantly higher than that of R1 (13.73%) (p = .000), suggesting that R2 could also result in more *yes* answer to the nonwords than R1. Therefore, it might be possible to say that R2 tends to lead to more guessing than R1.

However, the calculation of YN scores (hit rate minus false alarm rate) results in insignificantly different YN scores of R1 and R2 as mentioned earlier because it is likely that R2 not only promotes the higher hit rate, but also the higher false alarm rate. That is to say, the similar YN test scores of R1 and R2 tests come from significantly different hit and false alarm rates, which suggests that these 2 rubrics could lead to different reaction of the test takers to the YN tests as mentioned earlier.

These results also may be of interest of those people (e.g. Shillaw, 1996 and Harrington and Carrey, 2009), who pointed out that it might not be necessary to include nonwords in a YN test. The results of this rubric study may be useful to them because these 2 different rubrics tend to lead to significantly different hit rate. The YN scores of a no-nonword YN test version will come from the calculation of hit rate only, and no false alarm rate involved. The significantly different hit rate means significantly different YN scores of a YN test version that do not include nonwords.

In conclusion, without nonwords, R1 and R2 are likely to lead to significantly different YN test scores. However, with the presence of nonwords, R1 and R2 tend not to be different in the YN test scores.

Nevertheless, the quality of these 2 test types are judged more on the basis of their correlation to the translation tests than that of their YN scores as mentioned earlier. Therefore, R1 are likely to have better quality in predicting the actual VS of the test takers than R2 because R1 YN test scores (.890) yield better correlation coefficient with the translation scores than R2 (.812), and R1 hit rate (.824) also has higher correlation with the translation scores than R2 (.754) (See Table 5.1 and 5.8 for more details). In case that a YN test version does not include nonwords, the hit rate will be used as the YN scores of that YN test version as mentioned earlier.

Table 5.8: Correlation between hit rate and translation scores of R1 VS R2 tests

YN Test	**Number of the participants**	**Pearson correlation**	**Sig. (2-tailed)**
R1 tests	300	.824**	**.000**
R2 tests	300	.754**	**.000**

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

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

5.3 Research Question 2: Do different types of nonwords bring about differences in the test results?

This research question comes from the question raised by Beeckmans *et al* (2001) to what extent a nonword should be different from a real word. The research question also comes from the information that there has been no report of using pseudohomophones in a YN test (Thoma, 2011). Pseudohomophones are nonwords created to have the same sounds as their original words such as *cerum* from *serum*. The reason why pseudohomophones have not been used in a YN test may be because researchers might feel that pseudohomophones are too close to real words. For instance, Thoma (2011), in his yes/no business English vocabulary size test, stated clearly that the nonwords he created were not pseudohomophones or any nonwords that looked too close to real words. This leads the researcher to explore whether or not pseudohomophones can be used in a YN test. In other words, the researcher is interested in investigating whether or not the similarity or difference in the sound of nonwords to their original words will lead to different YN test results.

As a consequence, in this present study, N1 (near-homophone nonwords) were created to be very close (in the aspect of the sound) to the original words while N2 (non-homophone nonwords) were created to be different. The 2 nonword types were created by changing one letter from their original words. N1 nonwords are nonwords which reflect how we judge a Thai L2 learner might pronounce the real word on which the nonword is based. An example is the nonword *willage*, which is how the Thai phonemic system (which lacks a /v/ sound) generally renders the real word *village*. N1 seeks to exploit this source of error through the use of what we call pseudohomophones. N2 nonwords, on the other hand, were created with different sounds-meaning differences which also exist in the Thai phonemic system) to the original words. An example would be *cillage* from *village.* The 2 nonword types created to be similar and different in the sounds from the original words may give us some ideas about to what extent a nonword should be different to a real word.

5.3.1 Why do N1 tests tend to be better in predicting actual vocabulary size of a test taker than N2 tests?

The results showed that although N1 and N2 tests have the same correlation coefficient (.858**) between the YN and translation scores (see Table 5.9 below), N1 tests have closer scores to the translation test (N1=52.81/Trans.=47.60) than N2 (N2=56.96/Trans.=46.20) (see Table 5.10 below). This suggests that N1 and N2 tests seems to have equal ability in predicting the actual VS of the participants in the aspect of ranking, but in the aspect of predicting exact VS, N1 tests tend to be better because N1 YN scores are closer to the translation scores than those of N2; from Table 5.10, we can see N1 scores are 5.21% different from the translation scores while the difference of N2 and translation scores are 10.76%.

Table 5.9 below shows the correlation coefficient between the YN and translation scores of N1 and N2 tests; Table 5.10 illustrates the difference between the YN and translation scores of N1 and N2 tests; and Table 5.11 reports the difference between the YN and translation scores of 6 different N1 tests and 6 different N2 tests.

Table 5.9: Correlation between YN and translation scores of N1 and N2 tests

YN Test	Number of participants	Pearson correlation	Sig. (2-tailed)
N1 tests	300	.858**	**.000**
N2 tests	300	.858**	**.000**

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

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

Table 5.10: Mean difference between YN test scores and translation scores of N1 and N2 tests.

YN tests	N	Mean YN test score	Mean translation score	Mean difference between YN and translation scores
N1 tests	300	52.81	47.60	**5.21**
N2 tests	300	56.96	46.20	**10.76**

Table 5.11: Mean difference between YN test scores and translation scores of six N1 tests and six N2 tests.

N1 tests					N2 tests				
Test	**N**	**Mean YN scores**	**Mean trans scores**	**Mean diff. (YN vs trans)**	**Mean diff. (YN vs trans)**	**Mean YN scores**	**Mean trans scores**	**N**	**Test**
1 (R1N1P1)	50	50.16	47.04	**3.12**	**11.28**	59.12	47.84	50	4 (R1N2P1)
2 (R1N1P2)	50	53.48	49.31	**4.17**	**12.47**	60.55	48.08	50	5 (R1N2P2)
3 (R1N1P3)	50	48.80	45.24	**3.56**	**9.96**	58.46	48.51	50	6 (R1N2P3)
7 (R2N1P1)	50	53.60	48.84	**4.76**	**5.20**	50.72	45.52	50	10 (R2N2P1)
8 (R2N1P2)	50	55.82	45.49	**10.33**	**14.04**	58.19	44.14	50	11 (R2N2P2)
9 (R2N1P3)	50	54.97	49.68	**5.29**	**11.62**	54.71	43.08	50	12 (R2N2P3)

However, it may seem confusing why N1 and N2 tests yield the same correlation coefficients (.858**) with the translation tests but have significantly different YN test scores (52.81 and 56.96 respectively) (p = .037*). See Table 5.12 below.

Table 5.12: Mean difference of N1 VS N2 tests

YN tests	Number of the participants	Mean YN test scores	SD	Sig. (2- tailed)
N1 tests	300	52.81	25.54	**.037***
N2 tests	300	56.96	23.09	

** The mean difference is significant at the 0.05 level*

Actually, the correlation coefficients between YN test scores and translation scores can give us information about to what extent the 2 sets of scores rank in the same way. The perfect correlation is 1.0, which means that a participant who did a YN test would rank exactly the same when he did the translation test. That is, if a participant ranks first in the YN test, he will also ranks first in the translation test, and this is so for the other participants. The mentioned correlation coefficient of .858 (i.e. high correlation) means that the YN test scores of N1 and N2 tests lead to nearly the same ranking of the translation scores.

However, the same ranking may means that the YN test scores and the translation scores differ either greatly or slightly from each other. For instance, suppose that the difference between YN scores of Test A and the translation scores was 5 marks and that Participant 1 scored 95 in the YN test and 90 in the translation test and he ranked first in both tests; Participant 2 scored 94 and 89 and ranked second; Participant 3 scored 93 and 88 and ranked third; and the same pattern also occurred to the other participants doing Test A, the correlation coefficient between Test A and the translation scores will be 1.0 (i.e. perfect correlation). And suppose that the difference of the YN scores of Test B and the translation scores were 10 marks, and that Participant 1 (of Test B group) scored 100 in the YN test and 90 in the translation test and he ranked first in both tests; Participant 2 scored 99 and 89 and ranked second, Participant 3 scored 98 and 88 and ranked third; and this was so to the other participants of the group, the correlation coefficient between YN and translation scores of this group would also be the same as those of Group A at 1.0 (perfect correlation).

The mentioned example means that although Test A and Test B have the same correlation coefficients of 1.0 between their YN and translation scores, it is possible that the YN scores of these 2 tests are different (e.g. the highest scoring participant of Group A scored 95 while of Group B scored 100 on the YN test although both had the same translation score of 90). And this example could be applied to the results of the comparison between N1 and N2 tests mentioned above in that N1 and N2 tests have the same correlation coefficient but significantly differ in their YN test scores.

Actually, for the purpose of placement, the correlation coefficient between YN and translation scores should be better indicator of the quality of a YN test in predicting a test taker's actual vocabulary size than how close a YN score can get to a translation score. For instance, suppose that Test A had the mean YN scores of 55, the translation scores of 50, and the correlation coefficient between the YN and translation scores of .80, while Test B had the mean YN scores of 60, the translation scores of 50, and the correlation coefficient between the YN and translation scores of .90. This means that Test A has closer YN scores to the translation scores than Test B but yields lower correlation coefficient than Test B, and it could be possible to conclude that Test B tend to has better quality in predicting a test taker's actual vocabulary size than Test A because of the higher correlation coefficient, which means the higher quality in ranking between YN and translation scores.

Nonetheless, apart from a good correlation coefficient between YN and translation scores, it is also ideal if the scores of a YN test can also get very close to the translation scores, which means that the YN test version could predict very close actual vocabulary size of a test taker. As mentioned earlier in 5.1, there have been no reports of any YN test version that can report exact vocabulary size of a test taker. For example, Barrow *et al* (1999) reported relatively high rate of vocabulary size overestimation (i.e. the rate that comes from the difference between the YN and translation scores) of 17%

in his study on Japanese college students while fairly low overestimation rates were reported by Stubbe (2012) of 3.24% and 5.67% in his study on higher ability Japanese university students and lower ability Japanese university students respectively. These low overestimation rates are quite ideal for any YN test version in order to predict a very close actual VS score of a test taker.

To summarize, in order to judge the quality of a YN test version (i.e. its ability to predict the actual vocabulary size of a test taker), we consider the correlation coefficient between the YN and translation scores first and then the closeness of the YN scores to the translation scores second. That is, the higher the correlation coefficient between the YN and translation scores is, the better the quality of that YN test version. And then, the second consideration, the closer the YN scores to the translation scores, the better the quality of that YN test version. In the same way, the quality of N1 and N2 tests are judged by the correlation coefficient between the YN and translation scores first and then the closeness of their scores to the translation scores second. When the results shows that the correlation coefficient between the YN and translation scores of these 2 test types are the same (.858**), the second criterion to judge the quality of the 2 tests is the closeness of their scores to the translation scores. As a result, N1 tests are likely to have more quality than N2 tests because they produce closer scores to the translation scores than N2 as mentioned earlier (5.21% and 10.76% difference from the translation scores respectively).

5.3.2 Why do N1 and N2 tests yield significantly different scores?

From Table 5.11 above, it can be seen that N1 tests yield significantly lower scores when compared with N2 tests (52.81% and 56.96% respectively) (p=.037*) because N1 tests produced many more false alarms (i.e. the answer *yes* to a nonword) than N2 (23.22% and 20.08% respectively) (p=.036*), so the scores of N1 tests were adjusted downward more than those of N2. This suggests that N1 may be more tempting than N2. Table 5.12 below illustrates the difference between the false alarm rate of N1 and N2 tests.

Table 5.13: Mean difference of false alarm rate between N1 and N2 tests

YN tests	Number of the participants	Mean false alarm rate (%)	SD	Sig. (2- tailed)
N1 tests	300	23.22	17.38	**.036***
N2 tests	300	20.08	19.29	

** The mean difference is significant at the 0.05 level*

The reason why N1 were more tempting than N2 may be because N1 were created to have the same sound as real words (nb for Thai L1 learners). So test takers may confuse them with real words, while N2 were created to have different sounds so less confusion might be found. Table 5.14 below reports reasons why the participants of this study answered *yes* to both N1 and N2.

Table 5.14: Summary of finding from Interview Question 7: What were the reasons you answered Y (yes) to some nonwords?

Answers	Details	N1 test participants (out of 300)	N2 test participants (out of 300)
1. Having some partial knowledge of the nonword	1.1 Confusing the nonword with another word by giving a wrong translation	43.24%	13.27%
	1.2 Having seen/heard/or felt familiar with the nonword.	21.01%	11.16%
	Total	**64.25%**	**24.43%**
2. Guessing	-	34.84%	75.28%
3. Misreading the nonword	-	0.90%	0.29%

This interview question aimed to draw the information about why the participants decided to answer 'yes' to some nonwords. Initially, it was intended to ask the 72 interviewees, but because the information on the translation test papers could be a good source to answer this question, the researcher decided to use the information on the paper instead. That is, on the translation test paper, the participants were told that they had to translate all the words (and nonwords) they answered Y. If they could not translate them, they had to write, on each word, the reason why they could not translate them. Therefore, the results in Table 5.14 come from 600 papers instead of the 72 interviewees.

From Table 5.14, we can see that the participants chose N1 because they had some partial knowledge of them (64.25%) rather than because they just guessed (34.84%) while N2 were mainly chosen by the reason of random guessing (75.28%) more than by partial knowledge overestimation (24.43%). The partial knowledge overestimation here means 1) the participants confused a nonword with another word such as *metter* with *meter*; or 2) they had seen, heard, or felt familiar with a nonword but could not translate it or were not sure about its spelling.

As mentioned above, the participants chose N1 nonwords because they had some partial knowledge (i.e. phonological knowledge, however flawed) of them rather than blindly guessing while N2 were mainly chosen by random guessing more than overestimating by partial knowledge. This suggests that N1 were chosen more on the basis of actual error than just guessing while N2 seems to cause less confusion so those who chose them tend to be those who guessed more than those who were confused. Thus, this finding from Interview Question 7 seems to support the assumption mentioned earlier that N1 are likely to cause more confusion than N2 because of their close similarity of the sound to real words and this could be the reason why N1 were chosen more than N2 and lead to the significant difference in N1 and N2 test scores as mentioned earlier.

5.3.3 Why are N1 test scores closer to the translation scores than N2?

As mentioned earlier in 5.3.1, N1 tests produce closer scores to the translation scores than N2. That is, N1 test scores are 5.21% different from the translation scores while N2 are 10.76%. (See Table 5.10). This suggests that N1 tests could predict closer actual vocabulary size of a test taker than N2.

The reason why N1 tests yield closer scores to the translation test than N2 is possibly because the participants in this study tend to use more of their partial knowledge overestimation than random guessing (see Table 5.15 below), and N1 tend to be more tempting to those who employed partial knowledge overestimation than N2 (as mentioned earlier in 5.3.2). Thus, it is possible to say that N1 tend to be chosen more than N2 on the basis of this partial knowledge overestimation and, therefore, N1 nonwords tend to adjust the scores downward by virtue of partial knowledge overestimation, as compared to N2. This reason comes from the findings of Interview Question 6 (What were the reasons you answered *yes* to some real words that you could not translate or wrongly translated afterwards?) (See Table 5.15 below) and Interview Question 7 (What were the reasons you answered *yes* to some nonwords?)(see Table 5.14 above). Table 5.15 below shows the summary of finding from Interview Question 6 as follows.

Table 5.15: Summary of finding from Interview Question 6: What were the reasons you answered Y (yes) to some real words that you could not translate or wrongly translated afterward?

Reasons	**N1 test interviewees (out of 36)**	**Percent**	**N2 test interviewees (out of 36)**	**Percent**
1. I had seen/heard/or felt familiar with the word.	35	97.22	35	97.22
2. I made a guess.	10	27.77	14	38.88

The results suggests that the main reason for most participants to overestimate their vocabulary size on the part of the real words could be because they have some partial knowledge of those words more than the reason of random guessing. From Table 5.15, it can be seen that almost all of the interviewees (70 out of 72 or 97.22%) said that they answered *yes* to the real words they could not translate afterwards (i.e. overestimating their vocabulary size) because they had seen, heard, or felt familiar with the words, while only 24 out of 72 (33.33%) reported that they made a guess. Partial knowledge in this study means they do not have precise knowledge of the words in the following aspects: 1) they might have seen the words but could not translate them; 2) they might have heard the words but could not translate them; 3) they might have heard the words, could translate them, but were not sure about their spellings; 4) they might know only part of the words such as *come* from the word *become*; or 5) they might confuse a word with another word such as *complain* with *explain*. This kind of overestimation or confusion due to having only partial knowledge of the words can be seen in the wrong translation the participants gave to some real words. Table 5.15 shows examples of partial knowledge leading to wrong translation from the interviewees of Test 1 (R1N1P1) and Test 4 (R1N2P1), which are the first 2 tests (from 12 tests) that yield the highest correlation coefficient between YN and translation scores (.947** and .931** respectively).

Table 5.16: Examples of wrong translation on some real words

Test No.	Student No.	O–NET score of English (out of 100)	Words	Wrong Translation
1	1	85	pleasure	ความดี / น้ำใจ (merit)
			commerce	การเงิน (finance)
1	2	82	-	-
1	24	26	active	ท่าทาง (gesture)
1	25	26	become	กลับมา (come back)
			formal	สูตร (formula)
			active	กระตุ้น (activate)
			marry	สุขสันต์ (merry)
			minus	น้อย (little or few)
1	48	16	minus	น้อย (little or few)
1	50	14	normal	ง่าย (easy)
			complex	พัง (broken)
			remind	ลิมิต (limit)
			complete	คอมพิวเตอร์ (computer)
4	151	75	command	ต้องการ (want, demand)
4	152	62	damage	ที่รุนแรง / ที่ทำร้าย (serious, harmful)
4	175	26	area	สมัย / ศตวรรษ (era, century)
			lady	เริ่มต้น (begin/ready)
			culture	ความจริง (fact, truth)
4	178	25	escape	ยกเลิก (cancel)
			command	ความคิดเห็น (opinion)
			become	กลับมา (come back)
4	199	16	lady	เริ่มต้น (begin/ready)
			common	เข้ามา (enter, come in)
			marry	มาก (many)
			destroy	ความเป็นมา / ตำนาน (history, legend)

Table 5.16: Examples of wrong translation on some real words (Cont.)

Test No.	Student No.	O–NET score of English (out of 100)	Words	Wrong Translation
4	200	15	enough	ยัง (not yet)
			believe	อาศัยอยู่ (live)
			damage	ส่งข้อมูล (transfer data)
			mistake	พบเจอ (find)
			correct	เปลี่ยน (change)
			forest	ที่สูง (highland)
			contain	งานใหญ่ / งานเลี้ยง (party)
			honest	อดทน (patient)
			adult	ด้วย (also)
			become	มา (come)
			destroy	ของสะสม (collection)
			complete	ตัดสินใจ (decide)

From Table 5.16, we can see that the wrong translation comes from many kinds of partial knowledge mention earlier such as confusing *pleasure* with *merit* or *minus* with *few* (close in the meaning); *become* with *come* (knowing part of the word ,i.e., *come*); *formal* with *formula*; *marry* with *merry* (close in spelling); and *lady* with *ready* (close in sound).

This kind of overestimation or confusion has also been found in the work of Anderson and Freebody (1983). They called this phenomenon resulting from the confusion mentioned earlier as “mock” hits (i.e. the answer *yes* to a real word a test taker thinks he know the meaning of but actually he confuses the word with another word or has only partial knowledge of that word). For example, in their study, *sham* was confused with *shame*. They also found that the YN test scores inflated because of this mock hit phenomenon. This phenomenon was also found in Ward (2005). In his study, he administered YN and translation tests to the university students from 2 universities in Thailand and found that many participants showed this kind of confusion such as *form* with *from*.

In summary, the main reason of the vocabulary size overestimation of the participants in this study tends to come from their partial knowledge of a word more than the reason of random guessing. Therefore, nonwords that can adjust scores downward should be created on the basis that they can be tempting to both those who use partial knowledge overestimation and those who randomly guess. The results from Interview Question 7 (see Table 5.14) suggest that although N1 and N2 tend to be tempting to both those who use partial knowledge overestimation and those who randomly guess, N1 are likely to be more tempting to those who had partial knowledge of them than N2. Therefore, the part of "mock" hits phenomenon (partial knowledge overestimation) could possibly be adjusted downward more by N1 than N2. As a result, N1 YN scores are closer to the translation scores than N2 (5.21% and 10.76% different from the translation scores respectively).

5.3.4 To what extent should nonwords be different from real words?

This question was raised by Beeckmans *et al* (2001), and may also be the question of many scholars in the YN test field because there has been no clear answer for this question until now. In the YN test literature, some researchers seem to avoid creating nonwords that tend to be too close to real words. For example, Abels (1994) suggested creating a nonword by changing more than one letter from an original word in order to prevent a test taker from misreading it with a real word.

However, the results of this present study suggests that N1, or near-homophone nonwords, could better predict the actual vocabulary size of a test taker than N2, which are non-homophone nonwords. This means that nonwords that are created to be very close to real words like N1 (i.e. created by changing only one letter from a real word and having the same or almost the same sound to that real word) tend to make a YN test more reliable than those created to be different (in the sound) like N2 (i.e. created by changing one letter from a real word but having different sound from that real words).

The reason why N1 tend to lead to more reliable YN test results could be that N1 reflect the problems of the participants' real language use. That is, those who tend to have some problems on real words are likely to choose nonwords that reflect that problems. As mentioned earlier, it was found that the main reason for the vocabulary size overestimation of the participants in this study seems to be partial knowledge overestimation, i.e., they may have only one or two aspects of the following knowledge of real words: the spelling, the sound, or the meaning. For instance, if a test taker answers *yes* to a real word because he has only partial knowledge of that word, for example, he has heard the real word's sound but does not know the word's meaning, he will be likely to choose a nonword for the same reason, i.e., the nonword he only knows parts of it such as only its sound but not its meaning or its spelling. On the other hand, those who have more precise knowledge of real words such as both their spelling and meaning tend to be able to reject N1 because they could recognize that N1 are similar to real words but have spelling mistakes.

The results that near-homophone nonwords (N1) could be used in a YN test and tend to make the test more reliable are supported by Thoma (2011)'s point of view that it is not necessary to have clearly defined rules about to what extent a nonword should differ from a real word because this issue also occurs among real words. That is, for real words, there are also no rules about to what extent they should differ from one another because some are even homonyms. This means that if a test taker has difficulty in recognizing 2 close real words such as confusing *form* with *from*, he will also choose nonwords that represent his confusion such as the nonword *persent*, which he confuses with *present*.

In short, for real words, a learner normally needs to have an ability to differentiate between 2 close real words because a lot of real words can be very close to one another in both their sound and spelling. Similarly, nonwords could be created to be very close to real words because those who could differentiate between 2 close real words would also be able to do so on the part of nonwords, and those who could not in the part of real words, would not be able to do so in the part of nonwords. So, a test taker who overestimates his vocabulary size because of this confusion would also choose nonwords that represent this confusion and so the scores would be adjusted downward properly and this is possibly the reason why near-homophone nonwords (N1) in this study can lead to relatively reliable YN test scores.

The mentioned point of view that N1 could better be used in a YN test possibly because they reflect problems of real language use of the participants in this present study on the aspect of partial knowledge overestimation may lead to the question whether we can put different types of nonwords together in one test in order to check for other kinds of their language problems. That is to say, different types of nonwords may fit different problems of a test takers' real language use. For example, the YN test in the study of Ward (2005) contained 3 types of nonwords within the same test: 1) “decoding”-type errors (which seem to be the same type as N2 in this present study) (such as *zabra*), 2) pseudoderivatives (such as *suggestment*), and 3) English-like nonwords such as *felinder*. He gave the reasons that the first type could be used to check whether a test taker overestimates his vocabulary size by confusing the spelling of this nonword type with his known words. The second type were created to check those who might guess by applying the knowledge of English derivation rules. And the third type were created to check those who might guess randomly. From this clear justification of employing different types of nonword, a test writer may gain more detailed insight the language problems a test taker may have. For instance, if a test taker chooses more of the “decoding”-error nonwords than the others, it possibly means that his main problem is spelling. And if he also chooses the English-like nonwords, it could be interpreted that he also makes a random guess. In short, the idea of mixing more than one type of nonwords within one test (with a clear justification for each type) might be a good way to increase validity and reliability of a YN test. However, in this present study, the difference on the justification of N1 and N2 seem not to be so clear-cut. That is, both N1 and N2 seems to be tempting to both those who overestimate their vocabulary size by partial knowledge (i.e. confusing the sound, the spelling, or the meaning of a word) and those who use random guessing, but at different levels. That is, as mentioned earlier in 5.3.2, 64.25% of those who chose N1 used partial knowledge overestimation while 34.84% reported that they made a guess. In contrast, 24.43% of those who chose N2 stated that they overestimated their vocabulary size by partial knowledge while 75.28% reported that they guess at random (See Table 5.14). It is possible to say that N1 tend to cause more confusion than N2 because of their close sound to real words, while N2 may also lead to confusion but to a lesser extent possibly because N2 were also created by changing only one letter (although the researcher's intention is to create N2 to have different sound from the original words). Therefore, some N2 may lead to the confusion to other real words that seem to have close spelling (e.g. the N2 nonword *metter* from the original word *letter* in this study was confused with the real word *meter*) while some may not. For example, there is no report of confusion on the N2 nonword *soster* from the original word *sister*, but the report of random guessing instead. In summary, both N1 and N2 could cause confusion to a test taker possibly because their creation of changing only one letter from the original word, which can lead to the confusion of word spelling. The difference of these 2 types of nonwords could come from the fact that N1 were created to be closer in the sound to real words while N2 were not. And this may be the reason why both N1 and N2 could cause confusion but to a different extent (i.e. N2 could cause less), so the justification for these 2 types of nonwords could not be clear cut. In other words, N1 and N2 are not completely different, so the purpose of using N1 or N2 in a YN test may be the same, i.e, to tempt both those who use partial knowledge overestimation and those who use random guessing. Therefore, if the purpose of a test writer is to check those who guess at random, using N1 or N2 in a YN test may not clearly explain a test taker's behavior. The reason could be that those who choose N1 or N2 tend to be both 2 types of people, i.e, either those who use partial knowledge overestimation or those who use random guessing. So, if a test writer really wants to find out those who only guess at random, adding the English-like nonwords like *felinder* might be better. This is possibly because the nonword *felinder* seems to be created without any model real words (i.e., not by changing one or two letters from a real word) so the possibility that it would be confused with another real word would be less than N1 or N2. So those who choose this kind of nonword (English-like) could be those who guess at random. Therefore, it could be concluded that this kind of nonword would be better for the purpose of checking random guessing of a test taker than N1 or N2. To summarize, the mixing of more than one kind of nonwords tend to be beneficial in that different types of nonwords may differently explain the reason of a test taker's overestimation. However, a clear justification for each nonword type is needed so the mixing of different nonword types will be of the greatest benefit.

As for the issue of nonword misreading, some might feel worry that N1 and N2, which were created by changing only one letter from the original word, would be misread because of the closeness in the sound or spelling to real words as mentioned earlier. Actually, the misreading of the nonwords did occur in this study, but of a low percentage of 0.90% for N1 and 0.29% for N2 (see Table 5.14), suggesting that the misreading of nonwords may have little effect, if any, to the results of this nonword study. The low percentage of this nonword misreading may result from the warning included in the YN test rubric that there are some nonwords in the test. This means that the awareness that there are some nonwords in the YN test may lead to more careful behavior of a test taker when doing this kind of test.

5.3.5 The plus and minus of N1 and N2

The plus of N1 is that N1 could trap a test taker who may have one aspect of partial knowledge, so they tend to make a YN test more reliable. This is possibly because N1 seems to be more tempting to those who overestimate their vocabulary size by their confusion than N2. As mentioned earlier in 5.3.3, the overestimation rate in this present study mainly comes from partial knowledge overestimation (i.e. confusion) (97.22%) more than random guessing (33.33%) (See Table 5.15). That is, the participants tend to claim that they know the real words on the basis that they have seen, heard, or felt familiar with the words without knowing their meaning or they thought they knew the words' meaning but actually they confused them with other words. This kind of partial knowledge overestimation can also be seen as confusion, so the nonwords that could adjust scores downward from those who overestimate their vocabulary size because of their confusion should be added in a YN test. It was found that N1 could be more tempting to those who overestimate their vocabulary size by their partial knowledge or confusion than N2. So, this may be the reason why N1 could result in more reliable scores than N2.

However, creating N1 is not an easy task. This is because N1 were created to have the same (or almost the same) sound to real words, so a test writer has to find an appropriate original word (real word used as a model for creating a nonword) that can be changed to be a nonword with the same sound. For instance, in case of a YN test of English target words, to create a near-homophone nonword, a test writer need to find an original word that contains, for example, the letter *S* in order to create a nonword by changing *S* to *C* so that it still maintain the same sound such as the original word *serum*, which can be changed to be a nonword *cerum*. The word like *depend* may not be a good choice because it is not easy to find a homophone for this word.

Nonetheless, creating N1 can be easy or difficult depending on the nature of the target language. In other words, one language may provide better opportunity for creating N1 than another because of different writing or pronunciation systems.

Also, test takers from different L1 backgrounds may react differently to nonwords, so it should be aware that N1 YN tests may work well with Thai speaking learners, but may not with other L1 background learners. An example of different L1 background reacting differently to nonwords can be seen in the work of Meara and Buxton (1987), who found that some nonwords are more attractive to the speakers of a particular L1 background than the speakers of other languages. They gave an example of the nonword *observement*, which is similar to a real word of Italian or French but not German. This means that it is easier for German speaking test takers to reject this nonword when compared with those who have latinate L1 background such as French or Italian test takers. Another example is the report that Arabic speaking learners are likely to be blind to the vowels in their L1 and, therefore, could misread the nonwords in the YN test. For example, they tend to confuse *tilt* with *toilet*, or *mascarate* with *miscreate* (Ryan and Meara 1991).

For the part of N2, N2 seems to cause less confusion than N1, so they seem to be less tempting for a test taker and tend to lead to less reliable YN test scores compared with N1.

However, if the purpose of a YN test is only for ranking (e.g. as a placement test), N2 tests could also be a good alternative of N1 because N2 tests yield the same correlation coefficient (.858**) as that of N1. This suggests that N1 and N2 tests tend to be equal in the ability to predict the actual vocabulary size of a test taker in the aspect of ranking. The less reliable N2 test scores comes from only the greater difference of N2 YN test scores to the translation scores when compared with N1 (i.e. more inflation of YN scores from the actual vocabulary size scores compared with N1), not from their correlation coefficient of the YN to the translation scores as mentioned above. That is, the mean N2 test scores are more different from the translation scores at 10.76% while those of N1 are different from the translation scores at 5.21%.

Another advantage of N2 is that they are very easy to create. That is, an N2 can be created by changing only one letter from any real words. For example, an original word *depend* can be changed to a nonword *repend* or *depind*. This means that a test writer will not have to spend time finding a proper word to be an original word for a near-homophone nonword like N1. This means that if a test writer do not care about the inflation of the scores N2 can cause more than N1, N2 could be very good alternatives to N1 because they are easy to create and can lead to comparatively reliable YN test scores in the ranking aspect.

5.4 Research Question 3: Do different proportions of nonwords to real words bring about differences in the test results?

This research question comes from the information that the proportion of nonwords are not the same from different studies and the appropriate proportion is still unclear (Beeckmans et al., 2001; Eyckmans et al., 2007). This lead to the doubt whether different proportions of nonwords lead to different YN test results. So this research question was formed to investigate 3 simple ranges of the proportion between nonwords and real words. The 3 proportions are as follows:

- Proportion 1 (P1) => 50 real words : 50 nonwords

  (The number of the nonwords is the same as that of the real words.)
- Proportion 2 (P2) => 67 real words : 33 nonwords

  (The number of nonwords is half of the number of the real words.)
- Proportion 3 (P3) => 90 real words : 10 nonwords

  (There are only few nonwords in the YN test.)

Table 5.17: Mean difference of P 1, P2, and P3 tests

YN tests	Number of the participants	Mean	SD	Sig.
P1 tests	200	53.40	26.18	.301
P2 tests	200	57.01	24.11	
P3 tests	200	54.24	22.82	

Table 5.18: Concurrent validity results of P1, P2, and P3 tests

YN Test	No. of the participants	Pearson correlation	Sig. (2-tailed)
P1 tests	200	.903**	**.000**
P2 tests	200	.871**	**.000**
P3 tests	200	.773**	**.000**

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

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

The results suggest that P1 tests seems to be the best in predicting the actual vocabulary size of a test taker, while P2 the second, and P3 the third. From Table 5.17, we can see that P1, P2, and P3 tests do not lead to significantly different YN test scores (53.40%, 57.01%, 54.24% respectively, P = .301). However, from Table 5.18, it can be seen that P1 tests yield the highest correlation to the translation test (.903**), while P2 the second (.871**), and P3 the third (.773**).

The reason why P1 lead to the most reliable YN test scores is possibly because it is commonly known that, all other things being equal, the more items in a test, the more reliable the tests are. That is, if a test contains a small number of items, how well a test taker performs on the test will somehow depend on their luck in knowing the right answers. In this study, P1 tests contain 50 nonwords, which is more than P2 (33 nonwords) and P3 (10 nonwords), so those who can identify the 10 nonwords from the 90 real words (P3 tests) may depend more on luck than those who can reject all 50 nonwords (P1 tests). Alternatively it may just be that the more opportunities a subject has to demonstrate the weaknesses of his vocabulary knowledge, the more reliable his scores are for predicting actual vocabulary size.

5.5 Combination of the 3 variables: rubric, nonword type, and nonword proportion.

As mentioned earlier, R1 tend to be better than R2 tests in predicting the actual VS of a test taker (r = .890** and .812** respectively). While N1 tend to be better than N2 tests in the aspect that N1 lead to closer scores to the translation test than N2. However, in the aspect of ranking, N1 and N2 tests seem to have equal ability because N1 and N2 tests yield the same correlation coefficient to the translation scores (.858**). For the proportion tests, P1 tests tend to be the most reliable tests (.903**), while P2 the second (.871**), and P3 the third (.773**).

Therefore, if a test writer wants to make use of a good quality (i.e. the ability to predict the actual VS of a test taker) YN test version, it would be best for him to choose the best 3 variables mentioned earlier, i.e., R1, N1 and P1, in order that the best quality YN test could be obtained.

As a result, if we look at the correlation coefficient between the YN and translation scores alone, the best combination could be R1, N1, and P1 or R1, N2, and P1. And the YN tests created in this study to represent these 2 best combinations are Test 1 (R1N1P1) and Test 4 (R1N2P1). The results also confirm that the combination of these 3 best variables yield the highest correlation coefficient with the translation scores at .947** and .931** respectively.

On the other hand, if we combine the worst variables (on the basis of correlation coefficient of YN and translation scores alone), which are R2, N1 or N2, and P3 together, we could expect the lowest correlation coefficient from this combination. Test 9 represents the combination of R2, N1, and P3 and yield the worst correlation coefficient of the 12 tests in this study (.687**) while Test 12 represents the combination of R2, N2, and P3 and yield the second worst correlation coefficient at (.745**)

5.6 Summary

All the information in this chapter is an attempt to provide discussion of the findings of this present study, starting from pointing out that a YN test available now should be used for the purpose of ranking because the YN scores could predict only tentative, not exact, vocabulary size. Then, the discussion of the 3 research questions follows. And finally, the combination of the 3 variables (i.e. rubric, nonword type, and nonword proportion) was discussed. The summary of this study will be presented in the following chapter.

CHAPTER 6

CONCLUSION

This chapter presents the summary of this study, implications, and suggestions for further study.

6.1 Summary of this present study

This study points out that a YN test should be used for the purpose of ranking now because the YN scores could predict only tentative, not exact, vocabulary size possibly because the present problems of inconsistency in its rubric, nonword type, nonword proportion, and calculation method, etc. Therefore, this present study pays more attention to the ability to rank test takers than the ability to tell the actual vocabulary size of a YN test. So, the correlation coefficient between the YN and translation scores will be the most important indicator to tell the quality of a YN test in predicting the actual vocabulary size of a test taker in the ranking aspect, while a YN score will not be as important in itself as the mentioned correlation.

As for the conclusion of Research Question 1 (Do different test rubrics bring about significant difference in the test results?) It might be possible to say that, though not significantly different on the part of their YN test scores, R1 tends to be better than R2 in predicting the actual vocabulary size of the test takers because R1 tests (.890) yield slightly higher correlation between the YN scores and the translation scores than R2 (.812). This higher correlation coefficient of R1 may come from the reason that R1 are likely to lead to less guessing than R2. This point is supported by the results of the semi-structured interview that R1 test interviewees tend to make fewer guesses than R2.

For Research Question 2 (Do different types of nonwords bring about significant difference in the test results?), it was found that although N1 and N2 yield the same correlation coefficient (.858**) of the YN and translation test, N1 tend to be better than N2 in predicting the actual vocabulary size of a test taker because N1 lead to closer YN scores to the translation scores than N2 (5.21% and 10.76% difference from the translation scores respectively). It was also found from the semi-structured interview that the participants in this study tend to overestimate their vocabulary size by partial knowledge more than by random guessing (97.22% and 33.33% respectively). In addition, the results from the semi-structured interview also reveal that N1 were chosen more on the reason of partial knowledge overestimation while N2 were chosen more on the reason of random guessing. This suggests that N1 and N2 tend to be tempting to both those who use partial knowledge overestimation and those who use random guessing but to a different extent. That is, N1 tend to be more tempting to those who used partial knowledge overestimation than N2, so N1 tests produce significantly higher false alarms than N2 because most of the participants in this study are likely to overestimate their vocabulary size through partial knowledge as mentioned earlier.

For Research Question 3 (Do different proportions of nonwords to real words bring about significant difference in the test results?), the results suggest that P1 tests (r = .903**) tend to be the best in predicting the actual vocabulary size of a test taker, with P2 the second (r = .871**), and P3 the third (r = .773**). The reason that P1 tests are likely to be the most reliable might be that the more items in the test, the more reliable the tests are. That is, if a test contains only small number of test items, the chance that a test taker may get the right answers may depend to some extent on his luck. In this study, P1 tests contain the most proportion of nonwords (50 real words : 50 nonwords), while P2 the second (67 real words : 33 nonwords) and P3 the third (90 real words: 10 nonwords), so the accuracy of the scores of those who can reject 50 nonwords should be better than those who can reject 33, or only 10 nonwords.

6.2 Implications

As mentioned earlier, in the aspect of ranking alone, R1 tests tend to be better in quality to predict a test taker's actual vocabulary size than R2, N1 and N2 seems to be equivalent (in the ranking aspect), and P1 are likely to be the best, with P2 the second, and P3 the third. Therefore, if a test writer wants a more accurate YN test, he could employ R1, N1 or N2, and P1 in the test. And the YN test versions in this study that represent these 2 best combination are Test 1 (R1N1P1) (r = .947**) and Test 4 (R1N2P1) (r = .931**), which also show the highest correlation coefficients with the translation test (of the12 YN tests). The results also confirm that the 2 worst combination tests in this study produce the lowest correlation coefficient between the YN and translation test. That is, Test 9, which is the combination of R2, N1, and P3, yields the lowest correlation coefficient (.687**) to the translation test, and Test 12, the combination of R2, N2, and P3 yields the second lowest correlation coefficient at .745**.

6.3 Suggestions for further study

As mentioned earlier in 5.3.5, test takers from different L1 backgrounds might react differently to nonwords. That is, a nonword may be tempting to a learner of a particular L1 background but not to others. For example, Meara and Buxton (1987) gave an example of the nonword *observement*, which is similar to a real word of Italian or French but not German. This means that it is easier for German speaking test takers to reject this nonword when compared with French or Italian test takers. In this present study, near homophone nonwords seem to be attractive for Thai speaking learners, but it is possible that results may be different with learners of other L1 backgrounds. Therefore, further studies are needed to give clearer picture whether the near homophone nonwords also work well with other L1 background test takers.

REFERENCES

Abels, M. (1994). **Ken ik dit woord?** M.S. thesis, University of Nijmegen, the Netherlands.

Adolphs, S., and Schmitt, N. (2003). Lexical coverage of spoken discourse. **Applied Linguistics**. 24: 425-438.

Aitchison, J., and Straf, M. (1981). Lexical storage and retrieval: a developing skill? **Linguistics**. 19: 751-796.

Anderson, R. C., and Freebody, P. (1981). Vocabulary knowledge. In Guthrie, J.T. (ed.), **Comprehension and teaching: Research reviews** (pp. 77-117). Newark, Delaware: International Reading Association.

Anderson, R. C., and Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Huxton (ed.). **Advances in Reading/Language Research** (pp. 231-256). Greenwich, CT: JAI Press.

Beck, I. L., Perfetti, C. A., and McKeown, M. G. (1982). Effects of long-term Vocabulary instruction on lexical access and reading comprehension. **Reading Research Quarterly**. 17: 462-481.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., and Van de Velde, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. **Language Testing**. 18: 235-274.

Beglar, D., and Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. **Language Testing**. 16: 131-162.

Berent, I., and Marom, M. (2005). Skeletal structure of printed words: Evidence from the stroop task. **Journal of Experimental Psychology Human Perception and Performance**. 31: 328-338.

Blachowicz, C., Fisher, P., Ogle, D., and Watts-Taffe, S. (2006). Vocabulary: Questions from the classroom. **Reading Research Quarterly**. 41: 524-539.

Bonk, W. (2000). **Testing ESL learners' knowledge of collocations**. Illinois. (ERIC Document Reproduction Service No. ED442309)

Borowsky, R., and Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. **Journal of Experimental Psychology: Learning, Memory and Cognition**. 22:63-85.

Brown, R., and McNeill, D. (1966). The "tip of the tongue" phenomenon. **Journal of Verbal Learning and Verbal Behavior**. 5: 325-337.

Carreiras, M., Dunabeitia, J. A., and Molinaro, N. (2009) Consonants and vowels contribute differently to visual word recognition: ERPs of relative position proming. **Cerebral cortex**. 19: 2659-2670.

Coady, J. (1993). Research on ESL/EFL vocabulary acquisition: Putting it in context. In T. Huckin, M. Haynes and J. Coady (eds.). **Second language reading and vocabulary learning** (pp. 3-23). Norwood, NJ: Ablex.

Cobb, T. (2000) One size fits all? Francophone learners and English vocabulary tests. **The Canadian Modern Language Review/ La Revue canadienne des langues vivantes**, 57: 295-324.

Coxhead, A. (2000). A new academic word list. **TESOL Quarterly**. 34: 213-238.

Eyckmans, J. (2004). **Measuring Receptive Vocabulary Size**. Utrecht: the Netherlands.

Eyckmans, J., Van de Velde, H., van Hout, R., and Boers, F. (2007). Learners'response behaviour in Yes/No vocabulary tests. In H. Daller, J. Milton and J. Treffers-Daller (eds.). **Modelling and Assessing Vocabulary Knowledge** (pp. 59-76). Cambridge: Cambridge University Press.

Gibbs, P., and Van Orden, G. C. (1998). Pathway selection's utility for control of word recognition. **Journal of Experimental Psychology: Human Perception and Performance**. 24: 1162–1187.

Harrington, M. and Carey, M. (2009). The on-line Yes/No test as a placement tool. **System**. 37, 614-626.

Henriksen, B. (1999). Three dimensions of vocabulary development. **Studies in Second Language Acquisition**. 21: 303-317.

Hirsch, D., and Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? **Reading in a Foreign Language**. 8: 689-696.

Hu, M., and Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. **Reading in a Foreign Language**. 13: 403-430.

Huibregtse, I., Admiraal, W., and Meara, P. (2002) Scores on a yes/no vocabulary test: correction for guessing and response style. **Language testing**. 18: 227-245.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud and H. Béjoint (eds.). **Vocabulary and Applied Linguistics** (pp. 126-132). London: Macmillan.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady and T. Huckin (eds.). **Second language vocabulary acquisition** (pp. 20-34). Cambridge: Cambridge University Press.

Laufer, B. (1998). The development of passive and active vocabulary: Same or different? **Applied Linguistics**. 19: 255-271.

Laufer, B., Elder, C., Hill, K., and Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? **Language Testing**. 21: 202-226.

Laufer, B., and Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. **Language Learning**. 54: 339-436.

Laufer, B., and Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. **Applied Linguistics**. 16: 307-322.

Laufer, B., and Paribakht, T. S. (1998). Relationship between passive and active vocabularies: Effect of language learning context. **Language Learning**. 48: 365-391.

Leech, G., Rayson, P., and Wilson, A. (2001). **Word frequencies in written and spoken English**. Harlow, England: Longman.

Levelt, W. J. M. (1993). The architecture of normal spoken language use. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall. C-W. Wallesch (eds.). **Linguistic Disorders and Pathologies: An International Handbook** (pp. 1-15). Berlin: de Gruyter.

Meara, P. (1980). Vocabulary acquisition: A neglected area of language learning. **Language Teaching and Linguistics**. 15: 221-246.

Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. In J. Tommola (ed.). **Foreign Language Comprehension and Production** (pp. 103-113). Turku : AFinLa

Meara, P. (1991). Vocabulary profiles. In S. Verlinde (Ed.), **Proceedings of the Symposium on differentiation in LS Learning and Teaching** (pp.7-10). Leuven: Instituut voor Levende Talen.

Meara, P. (1992). **EFL Vocabulary Tests**. Centre for Applied Language Studies, University of College.

Meara, P. (1996). The dimensions of Lexical Competence. In G. Brown, K. Malmkjaer and J. Williams (eds.). **Performance and Competence in Second Language Acquisition** (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P., and Buxton, B. (1987). An alternative to multiple choice vocabulary tests. **Language Testing**. 4: 142-151.

Meara, P., and Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (ed.). **Applied Linguistics in Society** (pp. 80-87). London: Centre for Information on Language Teaching and Research.

Meara, P., and Jones, G. (1990). Eurocentres Vocabulary Size Test (version E1.1/K10) [Computer software] Zurich: Eurocentres Learning Service.

Meara, P., Lightbown, P. M., and Halter, R. H. (1994). The effect of cognates on the applicability of yes/no vocabulary tests. **The Canadian modern language review/la revue canadienne des langues vivantes**. 50: 296-311.

Meara, P., and Milton, J.L. (2002). X_Lex: The swansea vocabulary levels test. Express, Newbury.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. **Review of Educational Research**. 53: 253-79.

Milton, J. (2009) **Measuring Second Language Vocabulary Acquisition**. Bristol: Multilingual Matters.

Mochida, A. and Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary. **Language Testing**. 23: 73–98.

Nakanishi, Y., and Shimamoto, T. (2003). Vocabulary and its assessment. In S. Kadota (ed.). **The English mental lexicon** (pp. 31-62). Tokyo: Shohakusha.

Nation, I. S. P. (1983). Testing and teaching vocabulary. **Guidelines**. 5: 12-25.

Nation, I.S.P. (1990). **Teaching and Learning Vocabulary**. New York: Heinle and Heinle.

Nation, I. S. P. (2001). **Learning Vocabulary in Another Language**. Cambridge: Cambridge University Press.

Nation, I. S. P. (2005). Teaching and learning vocabulary. In E. Hinkel (ed.). **Handbook of research in second language teaching and learning** (pp. 581-595). Mahwah, NJ: Lawrence Erlbaum.

Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton and J. Treffers-Daller (eds.). **Modelling and Assessing Vocabulary Knowledge** (pp. 33-43). Cambridge: Cambridge University Press

Nation, I. S. P., and Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? **System**. 23: 35-41.

Nunally, J.C. and Bernstein, I.H. (1994) **Psychometric Theory**. New York: McGraw-Hill.

Oscarson, M. (1997). Self-Assessment of foreign and second language proficiency. In C. Clapham, and D. Corson (eds.). **Language Testing and Assessment** (pp. 175-187). Dordrecht: Kluwer Academic Publishers.

Paribakht, T. M., and Wesche, M. (1993). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huckin (eds.). **Second Language Vocabulary Acquisition: A Rationale for Pedagogy** (pp. 174-200). Cambridge: Cambridge University Press.

Paribakht, T. M., and Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huchin (eds.). **Second language vocabulary acquisition** (pp. 174-200). Cambridge: Cambridge University Press.

Pellicer-Sanchez, A., and Schmitt, N. (2012, Oct). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches [On-line]. **Language Testing**. 1-21.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge. **Canadian Modern Language Review**. 56: 282-307.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. **RELC-Journal**. 19: 12-25.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. **Language Testing**. 10: 355-371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (ed.). **Validation in Language Assessment** (pp. 41-60). Mahwah, NJ: Lawrence.

Read, J. (2000). **Assessing Vocabulary**. Cambridge: Cambridge University Press.

Read, J. (2007). Second Language Vocabulary Assessment: Current Practices and New Directions. **International Journal of English Studies**. 7: 105-125.

Rost, H. (2005). Interpretation und Bewertung padagogisch-psychologischer studien. Eine Einfuhrung. **Zeitschrift fur Padagogische Psycholoogie**. 19: 271-273.

Ryan, A., and Meara, P. (1991). The case of the invisible vowels: Arabic speakers reading English words. **Reading in a Foreign Language**. 7: 531-540.

Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? **TESOL Quarterly**. 18: 199-219.

Schmitt, N. (2000). **Vocabulary in Language Teaching**. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D., and Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. **Language Testing**. 18: 55-88.

Shillaw, J. (1996). **The Application of Rasch modelling to yes/no vocabulary tests** [On-line]. Available: http://www.swan.ac.uk/cals/vlibrary/js96a.htm

Shohamy, E. (2001). **The power of tests: A critical perspective on the uses of language tests**. London: Longman.

Sims, V. M. (1929). The reliability and validity of four types of vocabulary test. **Journal of Educational Research**. 20: 91-96.

Stahl, S. A. (1990). **Beyond the instrumentalist hypothesis: Some relationships between word meanings and comprehension** [Technical report No. 505]. University of Illinois, Urbana: Center for the Study of Reading.

Takac, V. P. (2008). **Vocabulary learning strategies and foreign language acquisition**. Clevedon, UK: Multilingual Matters.

Thoma, D. (2011). **Strategic attention in language testing : metacognition in a yes/no business English vocabulary test**. Frankfurt am Main: Peter Lang.

Tilley, H.C. (1936). A technique for determining the relative difficulty of word meanings among elementary school children. **Journal of Experimental Education**. 5: 61-64.

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

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. **Studies in Second Language Acquisition**. 27: 33-52.

Wesche, M., and Paribaht, T.A. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. **The Canadian Modern Language Review**. 53: 13-40.

West, M. (1953). **A General Service List of English Words**. London: Longman.

Wilkins, D. A. (1972). **Linguistics in Language Teaching**. London: Arnold.

Ward, J. W. (2005). **The lexical aspect of reading English as a foreign language for engineering undergraduates.** Ph.D. thesis. The University of Birmingham, U.K.

Zimmerman, C. B. (1997). Historical trends in second language vocabulary instruction. In J. Coady and T. Huckin (eds.). **Second language vocabulary acquisition** (pp. 5–19). New York: Cambridge University Press.

Zimmerman, K. J. (2004). **The role of vocabulary size in assessing second language proficiency**. M.S. thesis. Brigham Young University, USA.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., and Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures and some correlates of word and non word recognition. **Intelligence**. 1: 5-13.

APPENDIX A

90 Real Word

…..pleasure	…..baby	…..commerce
…..enough	…..compare	…..avoid
…..combine	…..detail	…..control
…..active	…..equal	…..hundred
…..bottle	…..dinner	…..beauty
…..million	…..complete	…..happy
…..normal	…..minus	…..occur
…..comfort	…..honest	…..morning
…..career	…..adult	…..open
…..culture	…..future	…..garden
…..believe	…..damage	…..permit
…..common	…..escape	…..profit
…..mistake	…..forest	…..heavy
…..correct	…..language	…..complain
…..lady	…..reply	…..little
…..marry	…..audience	…..second
…..effort	…..contain	…..remind
…..balance	…..command	…..destroy
…..become	…..area	…..announce

…..confuse	…..injure	…..nature
…..tidy	…..pregnant	…..disease
…..subject	…..treasure	…..refuse
…..body	…..excite	…..exchange
…..forget	…..agree	…..invite
…..suffer	…..promise	…..address
…..order	…..coffee	…..formal
…..convince	…..discount	…..progress
…..complex	…..easy	…..disgust
…..deny	…..busy	…..gather
…..famous	…..country	…..answer

APPENDIX B

50 Near-homophone nonwords

1 …..sertain	from certain	21 …..cercle	from circle
2 …..survay	from survey	22 …..shouldir	from shoulder
3 …..bisguit	from biscuit	23 …..offise	from office
4 …..mirrer	from mirror	24 …..meybe	from maybe
5 …..purshase	from purchase	25 …..serect	from select
6 …..bersurk	from berserk	26 …..retter	from letter
7 …..vindow	from window	27 …..displey	from display
8 …..sicret	from secret	28 …..cansel	from cancel
9 …..sicnal	from signal	29 …..musig	from music
10 …..marget	from market	30 …..westurn	from western
11 …..rable	from label	31 …..conferm	from confirm
12 …..cornor	from corner	32 …..deley	from delay
13 …..digree	from degree	33 …..durvish	from dervish
14 …..wictim	from victim	34 …..merder	from murder
15 …..advanse	from advance	35 …..traffig	from traffic
16 …..factur	from factor	36 …..shapter	from chapter
17 …..purson	from person	37 …..sistem	from system
18 …..vorry	from worry	38 …..persent	from percent
19 …..serwice	from service	39 …..piriod	from period
20 …..bigin	from begin	40 …..retern	from return

41 …..selious	from serious	46 …..prefur	from prefer
42 …..trousors	from trousers	47 …..finich	from finish
43 …..lepair	from repair	48 …..mashine	from machine
44 …..teble	from table	49 …..sistor	from sister
45 …..shannel	from channel	50 …..willage	from village

APPENDIX C

50 non-homophone nonwords

1 …..gertain	from certain	20 …..bugin	from begin
2 …..survoy	from survey	21 …..carcle	from circle
3 …..discuit	from biscuit	22 …..shoilder	from shoulder
4 …..morror	from mirror	23 …..offine	from office
5 …..surchase	from purchase	24 …..moybe	from maybe
6 …..bersork	from berserk	25 …..sedect	from select
7 …..lindow	from window	26 …..metter	from letter
8 …..socret	from secret	27 …..disploy	from display
9 …..sinnal	from signal	28 …..canbel	from cancel
10 …..mardet	from market	29 …..musin	from music
11 …..pabel	from label	30 …..wostern	from western
12 …..cirner	from corner	31 …..confarm	from confirm
13 …..dogree	from degree	32 …..deloy	from delay
14 …..bictim	from victim	33 …..dervash	from dervish
15 …..adhance	from advance	34 …..morder	from murder
16 …..fictor	from factor	35 …..traffin	from traffic
17 …..porson	from person	36 …..chaster	from chapter
18 …..porry	from worry	37 …..sustem	from system
19 …..serdice	from service	38 …..pergent	from percent

39 …..pariod	from period	45 …..chanbel	from channel
40 …..retorn	from return	46 …..profer	from prefer
41 …..sedious	from serious	47 …..rinish	from finish
42 …..troisers	from trousers	48 …..mathine	from machine
43 …..depair	from repair	49 …..soster	from sister
44 …..toble	from table	50 …..cillage	from village

APPENDIX D

Test 1 (R1N1P1)

Directions:

*Write Y (Yes) in front of the word if you **know its meaning** and write N (No) in front of the word if you **don't know its meaning**. There are some **nonwords** in this test.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษา**ทราบความหมาย** และเขียน N (No) หน้าคำที่นักศึกษา **ไม่ทราบความหมาย** ในข้อสอบนี้มี**คำหลอก**อยู่ด้วย*

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐	become	☐	retter	☐	survay	☐	open
☐	sertain	☐	combine	☐	cansel	☐	honest
☐	bisguit	☐	escape	☐	purshase	☐	formal
☐	bottle	☐	mirrer	☐	beauty	☐	westurn
☐	sicret	☐	future	☐	marget	☐	durvish
☐	merder	☐	vindow	☐	shapter	☐	cornor

rable	conferm	persent	effort
permit	mistake	profit	piriod
retern	correct	destroy	trousors
purson	sicnal	serect	finich
vorry	traffig	lepair	prefur
area	morning	normal	million
complex	balance	shannel	displey
active	wictim	forest	hundred
bersurk	advanse	deley	occur
lady	remind	marry	damage
pleasure	selious	factur	minus

☐ audience	☐ reply	☐ commerce	☐ mashine
☐ serwice	☐ offise	☐ comfort	☐ musig
☐ language	☐ contain	☐ enough	☐ sistem
☐ digree	☐ command	☐ injure	☐ career
☐ culture	☐ cercle	☐ meybe	☐ adult
☐ happy	☐ announce	☐ complete	☐ sistor
☐ dinner	☐ shouldir	☐ teble	☐ baby
☐ bigin	☐ believe	☐ common	☐ willage

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ   ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =……………words) (I answered N (no) =………………words)

APPENDIX E

Test 2 (R1N1P2)

Directions:

Write Y (Yes) in front of the word if you ***know its meaning*** *and write N (No) in front of the word if you* ***don't know its meaning****. There are some* ***nonwords*** *in this test.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษา**ทราบความหมาย** และเขียน N (No) หน้าคำที่นักศึกษา **ไม่ทราบความหมาย** ในข้อสอบนี้มี**คำหลอก**อยู่ด้วย*

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐ contain	☐ pleasure	☐ bigin	☐ heavy
☐ command	☐ enough	☐ cercle	☐ complain
☐ area	☐ combine	☐ shouldir	☐ little
☐ minus	☐ marry	☐ finich	☐ injure
☐ honest	☐ effort	☐ bersurk	☐ formal

☐ adult	☐ balance	☐ wictim	☐ sertain
☐ future	☐ become	☐ purshase	☐ bisguit
☐ commerce	☐ active	☐ meybe	☐ second
☐ damage	☐ baby	☐ survay	☐ mirrer
☐ escape	☐ compare	☐ factur	☐ merder
☐ forest	☐ detail	☐ digree	☐ vindow
☐ language	☐ equal	☐ prefur	☐ sicret
☐ reply	☐ dinner	☐ vorry	☐ offise
☐ audience	☐ complete	☐ serwice	☐ marget
☐ avoid	☐ bottle	☐ retter	☐ remind
☐ control	☐ million	☐ musig	☐ destroy

☐ hundred	☐ normal	☐ durvish	☐ announce
☐ beauty	☐ comfort	☐ sistem	☐ confuse
☐ happy	☐ career	☐ persent	☐ tidy
☐ occur	☐ culture	☐ retern	☐ subject
☐ morning	☐ believe	☐ selious	☐ suffer
☐ open	☐ common	☐ trousors	☐ convince
☐ garden	☐ mistake	☐ lepair	☐ complex
☐ permit	☐ correct	☐ deley	☐ deny
☐ profit	☐ lady	☐ shannel	☐ famous

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ   ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =……………words) (I answered N (no) =………………words)

APPENDIX F

Test 3 (R1N1P3)

Directions:

Write Y (Yes) in front of the word if you ***know its meaning*** *and write N (No) in front of the word if you* ***don't know its meaning****. There are some* ***nonwords*** *in this test.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษา**ทราบความหมาย** และเขียน N (No) หน้าคำที่นักศึกษา **ไม่ทราบความหมาย** ในข้อสอบนี้มี**คำหลอก**อยู่ด้วย*

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐	beauty	☐	language	☐	minus	☐	heavy
☐	durvish	☐	normal	☐	forest	☐	announce
☐	subject	☐	bisguit	☐	profit	☐	coffee
☐	disease	☐	marry	☐	control	☐	complex
☐	nature	☐	lady	☐	avoid	☐	convince
☐	shannel	☐	discount	☐	escape	☐	destroy

exchange	balance	sertain	famous
invite	become	happy	injure
country	correct	commerce	order
gather	survay	permit	promise
digree	active	future	second
busy	mistake	area	suffer
formal	compare	morning	treasure
body	equal	garden	agree
complete	enough	honest	complain
refuse	effort	hundred	deny
easy	common	command	forget

☐ address	☐ baby	☐ occur	☐ pregnant
☐ vindow	☐ culture	☐ audience	☐ answer
☐ million	☐ believe	☐ contain	☐ disgust
☐ progress	☐ detail	☐ open	☐ excite
☐ trousors	☐ bottle	☐ damage	☐ remind
☐ dinner	☐ comfort	☐ pleasure	☐ confuse
☐ bersurk	☐ career	☐ reply	☐ tidy
☐ retter	☐ combine	☐ adult	☐ little

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX G

Test 4 (R1N2P1)

<u>Directions:</u>

Write Y (Yes) in front of the word if you ***<u>know its meaning</u>*** *and write N (No) in front of the word if you* ***<u>don't know its meaning</u>****. There are some* ***<u>nonwords</u>*** *in this test.*

<u>คำสั่ง:</u>

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษา****<u>ทราบความหมาย</u>*** *และเขียน N (No) หน้าคำที่นักศึกษา* ***<u>ไม่ทราบความหมาย</u>*** *ในข้อสอบนี้มี****<u>คำหลอก</u>****อยู่ด้วย*

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐	combine	☐	future	☐	enough	☐	metter
☐	disploy	☐	escape	☐	discuit	☐	canbel
☐	bersork	☐	language	☐	surchase	☐	bottle
☐	normal	☐	deloy	☐	lindow	☐	audience
☐	dervash	☐	command	☐	area	☐	gertain

☐ believe	☐ chaster	☐ pabel	☐ commerce
☐ bictim	☐ pergent	☐ dogree	☐ beauty
☐ lady	☐ occur	☐ adhance	☐ retorn
☐ effort	☐ open	☐ porson	☐ troisers
☐ morder	☐ survoy	☐ damage	☐ pleasure
☐ million	☐ wostern	☐ reply	☐ confarm
☐ culture	☐ sinnal	☐ mardet	☐ traffin
☐ correct	☐ fictor	☐ mistake	☐ pariod
☐ active	☐ morror	☐ forest	☐ musin
☐ cirner	☐ hundred	☐ common	☐ sustem
☐ socret	☐ contain	☐ comfort	☐ career
☐ depair	☐ sedious	☐ happy	☐ morning

☐	balance	☐	permit	☐	porry	☐	marry
☐	prafer	☐	remind	☐	bugin	☐	chanbel
☐	carcle	☐	shoilder	☐	announce	☐	rinish
☐	honest	☐	injure	☐	adult	☐	soster
☐	become	☐	serdice	☐	profit	☐	toble
☐	minus	☐	mathine	☐	offine	☐	complex
☐	dinner	☐	destroy	☐	complete	☐	baby
☐	moybe	☐	formal	☐	sedect	☐	cillage

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX H

Test 5 (R1N2P2)

<u>Directions:</u>

Write Y (Yes) in front of the word if you ***<u>know its meaning</u>*** *and write N (No) in front of the word if you* ***<u>don't know its meaning</u>***. *There are some* ***<u>nonwords</u>*** *in this test.*

<u>คำสั่ง:</u>

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษา**<u>ทราบความหมาย</u>** และเขียน N (No) หน้าคำที่นักศึกษา **<u>ไม่ทราบความหมาย</u>** ในข้อสอบนี้มี**<u>คำหลอก</u>**อยู่ด้วย*

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

	announce		area		deloy		escape
	morder		adult		morrer		combine
	remind		fictor		pleasure		carcle
	heavy		forest		offine		normal
	tidy		socret		damage		gertain

☐ suffer	☐ contain	☐ serdice	☐ believe
☐ dogree	☐ minus	☐ become	☐ mistake
☐ metter	☐ survoy	☐ little	☐ pergent
☐ formal	☐ hundred	☐ moybe	☐ effort
☐ complete	☐ sustem	☐ destroy	☐ complex
☐ rinish	☐ morning	☐ happy	☐ compare
☐ dinner	☐ garden	☐ retorn	☐ equal
☐ prafer	☐ avoid	☐ troisers	☐ reply
☐ complain	☐ discuit	☐ honest	☐ enough
☐ musin	☐ subject	☐ career	☐ shoilder
☐ lindow	☐ audience	☐ profit	☐ culture

☐ deny	☐ commerce	☐ bugin	☐ correct
☐ depair	☐ beauty	☐ famous	☐ balance
☐ open	☐ bersork	☐ lady	☐ sedious
☐ second	☐ future	☐ bottle	☐ active
☐ convince	☐ command	☐ porry	☐ common
☐ marry	☐ occur	☐ dervash	☐ baby
☐ confuse	☐ language	☐ mardet	☐ comfort
☐ surchase	☐ permit	☐ detail	☐ bictim
☐ injure	☐ control	☐ million	☐ chanbel

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX I

Test 6 (R1N2P3)

Directions:

Write Y (Yes) in front of the word if you ***know its meaning*** *and write N (No) in front of the word if you* ***don't know its meaning****. There are some* ***nonwords*** *in this test.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษา**ทราบความหมาย** และเขียน N (No) หน้าคำที่นักศึกษา **ไม่ทราบความหมาย** ในข้อสอบนี้มี**คำหลอก**อยู่ด้วย*

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐ survoy	☐ morning	☐ busy	☐ active
☐ adult	☐ combine	☐ normal	☐ little
☐ formal	☐ bottle	☐ disease	☐ comfort
☐ forest	☐ language	☐ effort	☐ contain
☐ reply	☐ career	☐ bersork	☐ become

promise	believe	pleasure	body
area	mistake	enough	suffer
avoid	announce	nature	common
hundred	dervash	refuse	commerce
progress	tidy	invite	injure
discount	exchange	coffee	damage
garden	equal	disgust	agree
profit	complete	answer	lady
honest	discuit	balance	excite
escape	million	remind	destroy
dinner	permit	treasure	subject

☐ metter	☐ correct	☐ country	☐ order
☐ gertain	☐ minus	☐ compare	☐ gather
☐ open	☐ detail	☐ happy	☐ complain
☐ future	☐ heavy	☐ dogree	☐ second
☐ command	☐ convince	☐ easy	☐ troisers
☐ occur	☐ baby	☐ address	☐ pregnant
☐ deny	☐ chanbel	☐ beauty	☐ confuse
☐ culture	☐ audience	☐ famous	☐ lindow
☐ control	☐ marry	☐ forget	☐ complex

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX J

Test 7 (R2N1P1)

<u>Directions:</u>

Write Y (Yes) in front of the word if you think that the word ***<u>exists</u>*** *in English language and write N (No) in front of the word if you think that the word* ***<u>does not exist</u>*** *in the language.*

<u>คำสั่ง:</u>

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น**<u>คำที่มีในภาษาอังกฤษ</u>** และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็น**<u>คำที่ไม่มีในภาษาอังกฤษ</u>***

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐	adult	☐	lepair	☐	factur	☐	sicnal
☐	normal	☐	correct	☐	purson	☐	enough
☐	retter	☐	shannel	☐	prefur	☐	combine
☐	deley	☐	finich	☐	survay	☐	active
☐	forest	☐	destroy	☐	musig	☐	bisguit

☐ language	☐ mashine	☐ mirrer	☐ million
☐ conferm	☐ merder	☐ purshase	☐ rable
☐ bersurk	☐ willage	☐ reply	☐ comfort
☐ contain	☐ damage	☐ vindow	☐ sertain
☐ serwice	☐ displey	☐ sicret	☐ culture
☐ area	☐ cansel	☐ pleasure	☐ sistor
☐ complex	☐ bottle	☐ marget	☐ commerce
☐ advanse	☐ westurn	☐ escape	☐ effort
☐ beauty	☐ audience	☐ cornor	☐ teble
☐ cercle	☐ future	☐ honest	☐ lady
☐ occur	☐ durvish	☐ wictim	☐ marry

☐ morning	☐ hundred	☐ complete	☐ happy
☐ become	☐ traffig	☐ vorry	☐ balance
☐ offise	☐ shapter	☐ persent	☐ open
☐ profit	☐ digree	☐ bigin	☐ baby
☐ remind	☐ believe	☐ mistake	☐ trousors
☐ career	☐ piriod	☐ shouldir	☐ retern
☐ announce	☐ command	☐ permit	☐ minus
☐ common	☐ selious	☐ meybe	☐ sistem
☐ injure	☐ dinner	☐ serect	☐ formal

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =…………….คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX K

Test 8 (R2N1P2)

Directions:

Write Y (Yes) in front of the word if you think that the word ***exists*** *in English language and write N (No) in front of the word if you think that the word* ***does not exist*** *in the language.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น**คำที่มีในภาษาอังกฤษ** และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็น**คำที่ไม่มีในภาษาอังกฤษ***

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐	wictim	☐	serwice	☐	bottle	☐	remind
☐	adult	☐	shannel	☐	combine	☐	hundred
☐	vindow	☐	lepair	☐	bersurk	☐	commerce
☐	second	☐	normal	☐	purshase	☐	confuse
☐	sertain	☐	permit	☐	heavy	☐	become

☐ complain	☐ believe	☐ little	☐ prefur
☐ area	☐ mirrer	☐ reply	☐ mistake
☐ avoid	☐ lady	☐ cercle	☐ announce
☐ digree	☐ effort	☐ baby	☐ comfort
☐ honest	☐ detail	☐ deley	☐ bisguit
☐ escape	☐ million	☐ pleasure	☐ sicret
☐ audience	☐ merder	☐ equal	☐ language
☐ marry	☐ correct	☐ injure	☐ contain
☐ future	☐ active	☐ damage	☐ culture
☐ factur	☐ common	☐ forest	☐ vorry
☐ minus	☐ meybe	☐ survay	☐ marget

control	finich	happy	shouldir
beauty	selious	subject	retter
tidy	command	musig	suffer
occur	offise	convince	garden
persent	compare	enough	sistem
open	complex	durvish	balance
deny	bigin	famous	retern
career	dinner	destroy	morning
profit	complete	formal	trousors

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX L

Test 9 (R2N1P3)

<u>Directions:</u>

Write Y (Yes) in front of the word if you think that the word ***<u>exists</u>*** *in English language and write N (No) in front of the word if you think that the word* ***<u>does not exist</u>*** *in the language.*

<u>คำสั่ง:</u>

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น**<u>คำที่มีในภาษาอังกฤษ</u>** และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็น**<u>คำที่ไม่มีในภาษาอังกฤษ</u>***

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐ digree	☐ disgust	☐ dinner	☐ heavy
☐ sertain	☐ honest	☐ minus	☐ complain
☐ complete	☐ address	☐ little	☐ bisguit
☐ adult	☐ comfort	☐ future	☐ permit
☐ career	☐ second	☐ profit	☐ remind
☐ culture	☐ damage	☐ destroy	☐ million

☐ escape	☐ forest	☐ believe	☐ pleasure
☐ announce	☐ common	☐ enough	☐ trousors
☐ language	☐ mistake	☐ combine	☐ confuse
☐ shannel	☐ correct	☐ durvish	☐ reply
☐ active	☐ tidy	☐ equal	☐ famous
☐ beauty	☐ detail	☐ deny	☐ hundred
☐ coffee	☐ compare	☐ control	☐ garden
☐ open	☐ promise	☐ morning	☐ complex
☐ baby	☐ lady	☐ survay	☐ audience
☐ bottle	☐ subject	☐ marry	☐ bersurk
☐ contain	☐ injure	☐ body	☐ effort
☐ vindow	☐ command	☐ pregnant	☐ forget

☐ balance	☐ area	☐ suffer	☐ order
☐ happy	☐ treasure	☐ become	☐ commerce
☐ agree	☐ avoid	☐ occur	☐ excite
☐ convince	☐ discount	☐ easy	☐ busy
☐ country	☐ nature	☐ disease	☐ refuse
☐ progress	☐ gather	☐ invite	☐ exchange
☐ answer	☐ normal	☐ formal	☐ retter

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =................คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX M

Test 10 (R2N2P1)

<u>Directions:</u>

Write Y (Yes) in front of the word if you think that the word ***<u>exists</u>*** *in English language and write N (No) in front of the word if you think that the word* ***<u>does not exist</u>*** *in the language.*

<u>คำสั่ง:</u>

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น**<u>คำที่มีในภาษาอังกฤษ</u>** และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็น**<u>คำที่ไม่มีในภาษาอังกฤษ</u>***

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

☐	porry	☐	porson	☐	complete	☐	happy
☐	canbel	☐	pleasure	☐	combine	☐	lindow
☐	bottle	☐	language	☐	wostern	☐	surchase
☐	correct	☐	escape	☐	normal	☐	survoy
☐	beauty	☐	commerce	☐	career	☐	musin

☐ adhance	☐ active	☐ cillage	☐ sinnal
☐ morning	☐ deloy	☐ comfort	☐ morder
☐ adult	☐ audience	☐ enough	☐ damage
☐ confarm	☐ bersork	☐ million	☐ reply
☐ traffin	☐ mardet	☐ culture	☐ area
☐ offine	☐ bictim	☐ dogree	☐ future
☐ command	☐ morror	☐ occur	☐ forest
☐ sustem	☐ sedect	☐ common	☐ hundred
☐ dervash	☐ socret	☐ disploy	☐ contain
☐ retorn	☐ profit	☐ marry	☐ pabel
☐ formal	☐ gertain	☐ complex	☐ open
☐ destroy	☐ serdice	☐ troisers	☐ chaster

☐	effort	☐	prafer	☐	bugin	☐	toble
☐	mathine	☐	honest	☐	pariod	☐	metter
☐	believe	☐	cirner	☐	lady	☐	sedious
☐	baby	☐	fictor	☐	carcle	☐	pergent
☐	chanbel	☐	discuit	☐	dinner	☐	remind
☐	soster	☐	moybe	☐	mistake	☐	injure
☐	depair	☐	balance	☐	become	☐	permit
☐	rinish	☐	shoilder	☐	minus	☐	announce

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX N

Test 11 (R2N2P2)

Directions:

Write Y (Yes) in front of the word if you think that the word ***exists*** *in English language and write N (No) in front of the word if you think that the word* ***does not exist*** *in the language.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น**คำที่มีในภาษาอังกฤษ** และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็น**คำที่ไม่มีในภาษาอังกฤษ***

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

	occur		announce		profit		shoilder
	convince		minus		bottle		command
	porry		adult		lindow		permit
	chanbel		normal		offine		pleasure
	reply		little		fictor		dervash

☐ morrer	☐ escape	☐ equal	☐ retorn
☐ area	☐ socret	☐ discuit	☐ language
☐ become	☐ forest	☐ carcle	☐ effort
☐ hundred	☐ combine	☐ depair	☐ formal
☐ happy	☐ bugin	☐ musin	☐ deloy
☐ bictim	☐ complex	☐ sustem	☐ common
☐ garden	☐ prafer	☐ beauty	☐ morning
☐ gertain	☐ complete	☐ troisers	☐ dogree
☐ marry	☐ lady	☐ serdice	☐ complain
☐ believe	☐ detail	☐ contain	☐ destroy
☐ audience	☐ culture	☐ survoy	☐ subject

☐ commerce	☐ correct	☐ avoid	☐ deny
☐ suffer	☐ balance	☐ metter	☐ control
☐ comfort	☐ million	☐ pergent	☐ second
☐ future	☐ active	☐ morder	☐ bersork
☐ career	☐ rinish	☐ damage	☐ mistake
☐ famous	☐ baby	☐ tidy	☐ enough
☐ compare	☐ open	☐ mardet	☐ confuse
☐ remind	☐ dinner	☐ sedious	☐ surchase
☐ moybe	☐ honest	☐ heavy	☐ injure

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =………………words) (I answered N (no) =……………words)

APPENDIX O

Test 12 (R2N2P3)

Directions:

Write Y (Yes) in front of the word if you think that the word ***exists*** *in English language and write N (No) in front of the word if you think that the word* ***does not exist*** *in the language.*

คำสั่ง:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น**คำที่มีในภาษาอังกฤษ** และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็น**คำที่ไม่มีในภาษาอังกฤษ***

ต้องตอบ Y หรือ N ทุกคำ โดยห้ามเว้นไว้ (Do not omit any box)

chanbel	discount	forget	dinner
little	busy	confuse	combine
remind	dogree	damage	bottle
announce	refuse	forest	normal
language	invite	enough	career

☐ become	☐ formal	☐ exchange	☐ promise
☐ suffer	☐ profit	☐ area	☐ mistake
☐ convince	☐ answer	☐ avoid	☐ order
☐ deny	☐ discuit	☐ complete	☐ effort
☐ injure	☐ easy	☐ happy	☐ body
☐ treasure	☐ heavy	☐ second	☐ metter
☐ agree	☐ adult	☐ garden	☐ equal
☐ coffee	☐ lindow	☐ disgust	☐ hundred
☐ subject	☐ tidy	☐ address	☐ reply
☐ destroy	☐ disease	☐ escape	☐ pregnant
☐ complain	☐ honest	☐ audience	☐ culture
☐ lady	☐ gather	☐ commerce	☐ correct
☐ famous	☐ compare	☐ beauty	☐ balance

☐ excite	☐ dervash	☐ open	☐ detail
☐ morning	☐ country	☐ future	☐ active
☐ minus	☐ progress	☐ troisers	☐ common
☐ million	☐ command	☐ occur	☐ baby
☐ survoy	☐ contain	☐ bersork	☐ comfort
☐ believe	☐ nature	☐ permit	☐ gertain
☐ complex	☐ pleasure	☐ control	☐ marry

ให้นักศึกษานับจำนวนคำดังนี้ (Please count the followings)

นักศึกษาตอบ Y (Yes) ทั้งหมด =……………..คำ ตอบ N (No) ทั้งหมด= ………………..คำ

(I answered Y (yes) =……………words) (I answered N (no) =………………words)

CURRICULUM VITAE

Ms. Wallapha Wongsirichan was born in Prachin Buri. She received a B.A. in English from Chulalongkorn University in 1995, and an M.A. in English Language Studies from Suranaree University of Technology in 2008. She obtained a Ph.D. in English Language Studies from Suranaree University of Technology in 2014.

She is currently an English tutor for KAPLAN (Thailand), a test preparation center in Bangkok. Her research interests mainly lie in vocabulary learning and teaching.