Differences in the Test Results of Two Test Rubrics in the Yes/No Vocabulary Test

ความแตกต่างของผลสอบจากชุดคำสั่งที่แตกต่างกันในแบบทดสอบคำศัพท์ที่ให้ตอบ ใช่/ไม่ใช่

Wallapha Wongsirichan[*] and Jeremy Ward
วัลลภา วงศ์ศิริจันทร์[*] และ เจเรมี วอร์ด

School of Foreign Languages, Suranaree University of Technology, Thailand
สาขาวิชาภาษาต่างประเทศ มหาวิทยาลัยเทคโนโลยีสุรนารี จังหวัดนครราชสีมา

ABSTRACT

The Yes/No vocabulary test (YN test) is a receptive vocabulary size test that comes in a form of a checklist where learners answer yes to the words they know. However, an inconsistency of the test rubrics has been found. That is, rubric type 1 says: "Write Y (yes) if you **know the meaning** of the word" while rubric type 2 says: "Write Y (yes) if you **think that the word exists** in English". These 2 rubrics seem to require different levels of vocabulary knowledge from test takers. Rubric 1 may need more knowledge of the word meaning while Rubric 2 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, a study was conducted to compare these two rubrics by administering 2 YN test versions, i.e. Rubric 1 YN tests and Rubric 2 YN tests, to 600 first-year students of a university in Nakhon Ratchasima, followed by a translation test as a concurrent validity test. After that, a semi-structured interview of 72 students was conducted in order to gain some insight of how they did the tests. The findings revealed that the results of Rubric 1 YN tests correlated better with the results of the concurrent validity test than those of Rubric 2 YN tests.

ARTICLE INFO

---

[*] *Corresponding author*
E-mail address: aewallapha@yahoo.com

บทคัดย่อ

แบบทดสอบคำศัพท์ที่ให้ตอบใช่/ไม่ใช่ (แบบทดสอบ ใช่/ไม่ใช่) เป็นแบบทดสอบวัดจำนวนคำศัพท์เชิงรับรู้ ที่มาในรูปแบบของ การจัดเรียงคำศัพท์ให้ผู้ทำแบบทดสอบตอบว่า 'ใช่' หน้าคำศัพท์ที่เขารู้ความหมาย อย่างไรก็ตาม มีการพบความไม่สอดคล้องกัน ในแง่ของชุดคำสั่งในแบบทดสอบ กล่าวคือ ชุดคำสั่งแบบที่ 1 "ให้ตอบ 'ใช่' หากนักศึกษารู้ความหมายของคำศัพท์นี้" ส่วนชุดคำสั่ง แบบที่ 2 "ให้ตอบ 'ใช่' หากนักศึกษาคิดว่ามีคำศัพท์นี้ในภาษาอังกฤษ" ดูเหมือนว่าชุดคำสั่ง 2 ชุดนี้จะต้องการความรู้ทางด้านคำศัพท์ ของผู้ทำแบบทดสอบในระดับที่แตกต่างกัน นั่นก็คือ ชุดคำสั่ง 1 อาจต้องการความรู้ทางด้านความหมายของคำศัพท์มากกว่าชุดคำสั่ง 2 กล่าวคือชุดคำสั่ง 2 อาจต้องการแค่ความรู้ในระดับที่เคยเห็นคำศัพท์เท่านั้น หรืออาจจะในระดับที่ผู้ทำแบบทดสอบเดาเลยก็ได้ สิ่งนี้นำมาซึ่งความสงสัยว่าคำสั่ง 2 แบบนี้จะทำให้ผลสอบแตกต่างกันหรือไม่ ดังนั้นจึงดำเนินการศึกษาโดยให้นักศึกษาปีหนึ่ง จำนวน 600 คน จากมหาวิทยาลัยแห่งหนึ่งในจังหวัดนครราชสีมาทำแบบทดสอบ ใช่/ไม่ใช่ ที่มีชุดคำสั่งที่แตกต่างกัน คือแบบทดสอบ ใช่/ไม่ใช่ ที่ใช้ชุดคำสั่ง 1 และแบบทดสอบ ใช่/ไม่ใช่ ที่ใช้ชุดคำสั่ง 2 และตามด้วยการทำข้อสอบแปลคำศัพท์ ซึ่งใช้เป็นแบบทดสอบ เพื่อวัดความเที่ยงตรงตามสภาพ (concurrent validity test) จากนั้นดำเนินการ สัมภาษณ์แบบกึ่งโครงสร้างกับผู้ทำแบบทดสอบ 72 คน เพื่อให้ทราบรายละเอียดแบบเจาะลึกว่าผู้ทำแบบทดสอบนั้นมีวิธีการทำข้อสอบอย่างไร ผลการศึกษาพบว่าผลสอบของ แบบทดสอบ ใช่/ไม่ใช่ ที่ใช้ชุดคำสั่ง 1 มีความสัมพันธ์ที่ดีกว่ากับผลสอบของแบบทดสอบความเที่ยงตรงตามสภาพ เมื่อเทียบกับ ผลสอบของแบบทดสอบ ใช่/ไม่ใช่ ที่ใช้ชุดคำสั่ง 2

Introduction

For successful language learning, it is widely accepted that vocabulary plays a crucial role. A lot of research indicates that vocabulary knowledge correlates well with overall language proficiency (e.g. Saville-Troike, 1984; Laufer, 1997; Nation, 2001). Therefore, many scholars attempt to measure vocabulary knowledge of language learners in order to estimate their language proficiency. In research field, some researchers have estimated participants' L2 proficiency from their vocabulary knowledge and in some schools, learners' vocabulary knowledge has been used as a placement tool (Meara and Jones, 1988 and 1990; Harrington and Carrey, 2009).

The need for measuring vocabulary knowledge of a learner is crucial both for teachers and learners themselves. That is, 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. However, problems still remain that it is time-consuming to create and administer a vocabulary tests such as a commonly used multiple-choice test, which 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 used to measure vocabulary knowledge, is one interesting alternative for 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. This is because the

YN test only presents test takers with target words and requires them to only write Y (yes) in front of the words they know the meaning of. Therefore, to create this test, the writer does not need to be a trained item writer like a multiple-choice test writer. And without multiple choices that may confuse young test takers, the YN test is also easier for children to answer (Anderson and Freebody, 1983). And when creating and administering a test are no longer complicated tasks, the use of vocabulary tests will be more frequent, and L2 learners' vocabulary knowledge can be more accurately gauged and this can bring about the greater improvement of English language teaching and learning.

However, there are also some problems on the YN test. One of them is an inconsistency of the test rubric wording. That is, rubric type 1 says: "Write Y (yes) if you **know the meaning** of the word" while rubric type 2 says: "Write Y (yes) if you **think that the word exists** in the language". The 2 different wordings raised doubt to the researchers of this study. It called into question whether these 2 rubrics affect a test taker in the same ways. The reason is because the 2 rubrics seem to require different levels of word knowledge from the test takers. That is, the 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 the test takers only to have some phonological and/or morphological and perhaps semantic knowledge.

The purpose of this study, therefore, is to investigate whether the inconsistency of the test rubrics leads to different test results. If so, this knowledge will contribute to the improvement of the writing of the YN test rubric and means that the better quality of the YN test will be achieved.

This present study, therefore, aims to answer the research question: *Do the 2 different test rubrics, which use 2 different wordings: "knowing the meaning of the word" and "knowing whether the word exists", bring about differences in the test results?*

Review of literature

The YN test is a receptive vocabulary size test that presents the target words in a form of a checklist and asks test takers to answer yes to the word they know the meanings of (Read, 2000; Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., and Van de Velde, H., 2001; Nation, 2001; Eyckmans, J., Van de Velde, H., Van Hout, R., and Boers, F., 2007). The test may include nonwords, which are imaginary words added to the test to check whether the test takers does any guesswork. If the test takers claim that they know some nonwords, their scores will be adjusted downward. An example of a YN test containing the nonwords *metter, disploy, canbel, surchase,* and *factor* is provided below. A YN test can have 2 possible types of test rubrics, which could be Rubric 1: "Write Y (yes) if you

know the meaning of the word" or Rubric 2: "Write Y (yes) if you *think that the word exists* in English". Please note that, the example below is only the example of the YN test, which is not the same as the YN tests created for this present study. The full versions of the YN tests used in this study can be seen in the appendices while the full versions of the rubrics, which can be seen in the full versions of the YN tests, also appear in the methodology section .

Directions: 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. (Rubric 1)
Or
Directions: Look at the following word listed below. 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. (Rubric 2)

…..combine	…..future	…..enough	…..metter
…..disploy	…..escape	…..area	…..commerce
…..language	…..normal	…..canbel	…..audience
…..surchase	…..command	…..beauty	…..fictor

The YN test is very interesting because it does not require a lot of time to write and administer when compared with traditional vocabulary tests such as the multiple choice test (Meara and Buxton, 1987). It can test much more words than other kinds of tests in a shorter period of time, which means that the test can have a higher sampling rate than other test formats (Meara and Buxton, 1987; Read, 1988). Also, the writer of this test does not need to be a trained item writer because the test does not involve the complicated process of creating distractors (Anderson and Freebody, 1983). Moreover, this kind of test is easy for children to answer (Anderson and Freebody, 1983).

In addition, the YN test can be used as a placement test (Meara and Jones, 1990; and Harrington and Carrey, 2009) and can also be used to explore learners' vocabulary size for various purposes including research and diagnostic purposes (Meara, 1991; Read, 2007).

Despite many good characteristics, the YN test also has some problems regarding the discrepancy of its scoring methods, nonwords, and test rubric as well as the problems of different YN test results that come from different proficiency of the test takers (Meara, 1996) or from different L1 background (Meara, 1990)

Scoring the YN test

There are 4 types of answers involving in the scoring process:

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

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

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

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

At present, there are 4 main scoring formulae proposed to be used in the field, all of which have the same concept that if learners answer yes to false alarms, it means that they do some guesswork and their scores will be adjusted downward, while hits can roughly tell his receptive vocabulary knowledge. However, there are some differences in each scoring formula and there are no clear guidelines which formula is suitable for a YN test.

There have been some attempts to compare these different scoring methods. For example, Ward (2005) compared the four formulae and concluded that these 4 formulae yielded quite different results. He also mentioned that, with his personal communication with Meara, Meara suggested using the simple *hit minus false alarm* formula (i.e. H-F) because "All the empirical work we've done suggests that H-F is as good as anything, and it's easier for people to calculate and understand." (Ward, 2005, p. 30). Some recent researchers, e.g. Harrington and Carey (2009), also used the simple H-F formula in their studies.

In this present study, the H-F was also adopted as the scoring method for the YN tests.

Problems about YN test rubrics

A test rubric 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 have been only few studies on the rubrics of the YN test. Most of them addressed the inconsistency of the test rubrics. For example, Abels (1994) investigated whether telling participants that there are nonwords or not telling them made any differences and found that the participants had more careful behavior doing the YN test when they were informed that there were nonwords in the test. While Eyckmans (2004) compared a less strict test rubric: *"Tick the words you know. Some of the words in the list do not exist in Dutch"* with a more stringent one: *"Tick the words you know the*

meaning of. When in doubt, do not tick the item. Notice that some of the words in the list do not exist in Dutch. After completing this test, you will be asked to translate some of the words of the list." and found that the YN test results of these 2 rubrics were not significantly different.

Another inconsistency can be found in the 2 different types of rubrics. That is, rubric type 1 says: "Write Y (yes) if you *know the meaning* of the word" (e.g., Rubric A and B below), while rubric type 2 says: "Write Y (yes) if you think that the word *exists* in the language" (e.g., Rubric C). Nonetheless, rubric type 1 is more common in the YN test literature as can be seen in the examples of YN test rubrics below.

A). *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).

B). *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).

C). *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).

To the best of our knowledge, there has been no studies attempting to investigate whether the 2 different types of YN test rubrics mentioned above lead to any differences in the YN test results. This present study, therefore, was designed to answer this particular gap in the previous research.

Methodology

Instruments

The 2 Rubrics

The 2 test rubrics created for this present study were both in English and Thai as shown below. The sentence: *"There are some nonwords in this test."* was added to Rubric 1 on the assumption that the test takers would be more careful when doing the test if they know that nonwords were included. While Rubric 2 did not have this sentence because the rubric already implied the inclusion of nonwords.

1) 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.*

<u>คำสั่ง</u>:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาทราบความหมาย และเขียน N (No) หน้าคำที่นักศึกษา**<u>ไม่ทราบความหมาย</u>** ในข้อสอบนี้มีคำหลอกอยู่ด้วย*

2) 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.*

<u>คำสั่ง</u>:

*ให้นักศึกษาเขียน Y (Yes) หน้าคำที่นักศึกษาคิดว่าเป็น**<u>คำที่มีในภาษาอังกฤษ</u>** และเขียน N (No) หน้าคำที่นักศึกษา คิดว่าเป็น**<u>คำที่ไม่มีในภาษาอังกฤษ</u>***

YN tests

Actually, the data collected in this present study, which were 600 papers of the YN tests from 600 participants, were used to investigate 3 variables in a YN test: test rubrics, nonword types, and nonword proportions. The 2 test rubrics studied were Rubric 1 YN tests (knowing the meaning of the word) and Rubric 2 YN tests (knowing whether the word exists) while the 2 nonword types studied were Nonword type 1 (near-homophone nonwords) and Nonword type 2 (non-homophone nonwords). The 3 proportions were the proportion between real words and nonwords, which are Proportion 1 (50 real words : 50 nonwords); Proportion 2 (67 real words : 33 nonwords); and proportion 3 (90 real words : 10 nonwords).

To study these 3 variables at the same time, 12 YN test versions were created. These 12 test versions were the combination of the 3 variables and varied in terms of their rubric, nonword type, or nonword proportion. 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.

However, this present study focuses on the test rubric only, not nonword type or nonword proportion.

Real words

The 12 YN tests created for this present study comprise 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 words of their particular needs. Half of the real words in each YN test were the head words (i.e., words without any inflections (-*s*, -*ing*, -*ed*) or derivations (e.g., -*tion, -ism, -er, -ly*) drawn from the first thousand word frequency level and the other half drawn from the second thousand word frequency level.

Nonwords

For this study, the nonwords were created by changing one letter from the original words (i.e. real words used to create nonwords) such as *metter* from the original word *letter*. Two types of nonwords were employed, which were 1) near-homophone nonwords (e.g. *wictim* from the original word *victim*) and 2) non-homophone nonwords (e.g. *bictim* from the original word *victim*). The 2 nonword types 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.

Translation test

The translation test was designed to check the participants' guessing and the participants did it right after the YN test. In this study, the translation test was used as the criterion to measure the concurrent validity of the YN tests. 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 correlation between the 2 tests means that if the test takers get a high score in a YN test, we can predict that they will also get a high score in a translation test. The translation score is considered an accurate measure of the participants' vocabulary size in this present study. This means that if the test takers get a high translation score, it suggests that their actual vocabulary size is large. Also, the higher correlation between a YN test score and a translation score means that the particular YN test tended to be better in predicting the participant's actual vocabulary size.

The YN and translation tests were completed on the same paper. The participants did the YN test in the first period (10 minutes) and had to write the translation of the words they answered 'yes' on the same paper in the second period (10 minutes). The participants were not previously informed that they would have to translate the word they answered yes. In the first period (on the YN test), the participants had to write with the pink pens provided by the researcher while in the second period (on the translation test) they had to returned the pink pens and write with the green pens provided. They were not allowed 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 period. The translation scores were converted into percentages so that they could be compared with the YN test scores, which were also reported in a percentage form.

Semi-structured interview

The semi-structured interview was designed to draw deeper details about the participants' behavior when doing the YN test. The stratified random sampling method was employed 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 conducted in Thai and the question was *"Could you tell me frankly what criteria you used to answer Y (yes) for each word?"*

Data analysis

Data analysis of all 12 YN tests

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

Data analysis of the YN tests grouped for the rubric study

In this study, the 12 YN test versions were categorized into 2 main types according to the 2 types of the test rubrics: 1) Rubric 1 YN tests (R1 tests), which were Test 1-6, and 2) Rubric 2 YN tests (R2 tests), which were Test 7-12 (See Table 1).

Table 1. Grouping of the 12 YN test versions according to the 2 different test rubrics (See the information under this table for the explanation of the codes on the right of each test.)

Rubric 1 YN tests	Rubric 2 YN tests
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 $(R_1 \ N_1 \ P3)$	**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)

The variables investigated in this study were the 2 test rubrics. Therefore, the scores of the 2 rubric YN tests were compared to find out which set of scores (i.e., the scores of R1 tests (Test 1-6) and the scores of R2 tests (Test 7-12) had the higher correlation with the translation scores. This correlation was calculated using Pearson Correlation as a statistical tool. The higher correlation would mean that the particular rubric (either R1 or R2) could make the YN test scores more similar to the translation scores, which means that the particular rubric tended to be fitter for use in a YN test.

Three hundred papers were analyzed for R1 tests (Test 1-6) and the other 300 for R2 tests (Test 7-12). 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 alarms (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.

Examples of Rubric 1 and Rubric 2 YN tests can be seen in Appendix 1 and 2 respectively.

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 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 gave. One point was given to a 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.

Data analysis of the interview

The interview data was analyzed to gain an insight data as additional information to the quantitative data obtained from the test results. The audio interview data were transcribed and reported into the form of percentage.

Participants

The participants of this study were 600 first-year university students at a public university in Nakhon Ratchasima who were attending a mandatory basic English course of the university. 300 students did R1 tests while the other 300 did R2 tests.

The participants of the 2 groups were similar in their English proficiency. Each group of 300 students comprises the same proportion of higher (100 students), middle (100 students), and lower English proficiency students (100 students) according to their O-NET scores of English or O-NET (Eng). 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.

Please note that, 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 with other Thai students in general. What the researcher did was dividing the 600 participants into 3 groups of 200. The first 200 who scored the highest in 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 (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.

Procedures

1. The 2 types of YN tests were distributed to the 600 participants. That is, 300 students did R1 YN tests (Test 1-6) and the other 300 did R2 YN tests (Test 7-12).
2. Each participant did the translation test after the YN test.
3. Next, the semi-structured interview was conducted.
4. Then, the scores of the YN and translation tests and the interview data were analyzed.

Results and discussion

Reliability results of the 12 YN tests

In Cronbach's Alpha calculation, the acceptable reliability is at 0.700, and we can see from Table 2 below 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.

Table 2. 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

Mean difference and concurrent validity results of R1 and R2 YN 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 between YN and translation scores (i.e. concurrent validity) were used to find out which test would better predict participants' actual vocabulary size. This means that the correlation is the indicator of a YN test quality to predict the actual vocabulary size of learners while a YN score itself will not be as important as the correlation mentioned.

The mean difference and correlation results of R1 tests and R2 tests are reported in Table 3 and 5 respectively.

Table 3. Mean difference of R1 tests VS R2 tests (by t-test)

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. Concurrent validity results of R1 tests VS R2 tests

YN Test	Number of the participants	Pearson correlation	Sig. (2-tailed)	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 3 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, we can see that R1 tests yield a higher correlation coefficient between the YN and translation scores ($r = .890$) than that of R2 tests ($r =.812$). This suggests that R1 tests may have better tendency to predict the participants' actual vocabulary size than R2.

The reason why R1 tests tend to 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 2 (See Table 5 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. Summary of finding from the interview question: Could you tell me frankly what criteria you used to answer Y (yes) for each word?

Criteria for answering Y (yes) for each word	R1 test interviewees (out of 36)	percent	R2 test interviewees (out of 36)	percent
1) knowing the meaning of the word	36	100%	36	100%
2) having seen/ heard/ or felt familiar with the word	34	94.44%	36	100%
3) making a guess	6	16.67%	18	50%

From Table 5, 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 words they thought they had never seen/heard/or felt no familiarity with.

As can be seen, to answer Y (*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.

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 test scores of R1 (55.10%) and R2 (54.67%) are not significantly different ($p = .831$) (See Table 3). 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 3 shows the difference between the YN scores of R1 and R2 tests while Table 6 reports the difference between the translation scores of R1 and R2 tests; Table 7 illustrates the difference of the hit rate of R1 and R2 tests; and Table 8 presents the difference of the false alarm rate of R1 and R2 tests.

Table 6. 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 7. 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 8. 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 6, 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 vocabulary size. However, it can be seen from Table 7 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 8, 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. In any case, it appears that R2 tends to lead to more guessing than R1.

However, the calculation of YN test scores (hit rate minus false alarm rate) results in insignificantly different YN test scores of R1 and R2 as mentioned earlier because it can be seen 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 led to different reaction of the test takers to the YN tests as mentioned earlier.

Conclusion

The findings of this study showed that Rubric 1 (R1): "Write Y (yes) if you **know the meaning** of the word" tends to be better in predicting the actual vocabulary size of the test takers than Rubric 2 (R2): "Write Y (yes) if you **think that the word exists** in English". This is because R1 tests ($r = .890$) yield higher correlation between the YN scores and the translation scores than R2 ($r = .812$). This higher concurrent validity of R1 may come from the reason that R1 are likely to lead to less guessing than R2. This point is supported by the significantly higher rates of hit (R2 = 84.24%, R1 = 68.83%, $p = .000$) and false alarm (R2 = 29.56%, R1 = 13.73%, $p = .000$) of R2 tests than R1 tests, which suggests that R2 may lead to more guessing than R1. Moreover, the results of the semi-structured interview also support this point in that R2 test interviewees tended to make more guesses than R1 because they were likely to use Criterion 2 (answering Y [yes] to the words they had seen/heard/or felt familiar with but did not know the meaning of) and Criterion 3 (answering Y [yes] to the words they simply guessed about) more than R1 interviewees. This suggests that the wording "Write Y (yes) if you **think that the word exists in English**" in R2 may pave the way to more guessing than the wording "Write Y (yes) if you **know the meaning** of the word" in R1 because the wording "exist" possibly failed to lead the test takers to care about the meaning of the words they answer Y (yes) to.

Like most research, this present study also has some limitations so readers should be cautious about making generalizations of the results. The limitation of this study is that it was conducted with 600 first-year university students of a public university in Nakhon Ratchasima; thus, they may not represent students of other levels or in other areas. Further studies may be carried out with participants from different educational levels, areas, or L1 background.

Implications

The results of this study suggest that R1 could be a better choice to be used in a YN test if a test writer wants more accurate YN test results. R2 could also be a good alternative but its weakness seems to be that it could lead to more guessing as mentioned earlier.

Acknowledgements

References

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

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.

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.

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.

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

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.

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.

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

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), **Vocabulary in a Second Language: Selection, Acquisition, and Testing** (pp. 3-13). Amsterdam: John Benjamins.

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

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

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.

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

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

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.

APPENDIX 1

(An Example of Rubric 1 YN Test)

Test 5 (R1N2P2)

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)

	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

APPENDIX 2

(An Example of Rubric 2 YN Test)

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