LEARN Journal: Language Education and Acquisition Research Network
(ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 15, No: 2, July – December 2022
Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Do Two Nonword Types Bring About Different YN Vocabulary Test Results?

Wallapha Wongsirichan[a*], Anchalee Wannaruk[b], Jeremy Ward[c]

[a] aewallapha@yahoo.com, School of Foreign Languages, Suranaree University of Technology, Thailand
[b] wannaruk@sut.ac.th, School of Foreign Languages, Suranaree University of Technology, Thailand
[c] wardosan6@gmail.com, School of Foreign Languages, Suranaree University of Technology, Thailand
[*] Corresponding author, aewallapha@yahoo.com

APA Citation:
Wongsirichan, W., Wannaruk, A., & Ward, J. (2022). Do two nonword types bring about different YN vocabulary test results? *LEARN Journal: Language Education and Acquisition Research Network*, *15*(2), 237-262.

Received
18/12/2021

Received in revised form
31/03/2022

Accepted
28/04/2022

Keywords
YN vocabulary test; nonword; pseudoword

Abstract

The YN vocabulary test (YN test) is a vocabulary size test that presents test takers with a list of words and asks them to answer *yes* to the words they know. This test may include nonwords, i.e., imaginary words added to the test to check for guessing. However, there have been no conclusive guidelines about nonword construction. One way of creating a nonword is to change one or two letters from the original word such as *dactor* from *doctor.* This approach suggests one important question: To what extent should nonwords be different from their original words? This study, therefore, aims to compare 2 nonword types: one type which is phonologically similar to real words (N1) and the other which is less similar (N2) in order to check whether they lead to different YN test results. An example of N1 would be *willage*, which is how some Thai speakers pronounce *village* while N2 would be *cillage*. Two YN test types (N1 and N2 tests) were administered to 600 university students, followed by

	translation tests and semi-structured interviews. The results suggest that N1 tests tend to be better in predicting the actual vocabulary size of the participants than N2.

Introduction

It is undeniable that vocabulary is essential for successful language learning. Many studies reveal that vocabulary knowledge correlates well with overall language proficiency (e.g., Laufer, 1997; Nation, 2001; Qian, 2002). Thus, there have been attempts to use vocabulary knowledge as a measure of language proficiency. For example, some researchers employ learners' L2 vocabulary knowledge as a placement tool (Harrington & Carrey, 2009; Nation, 2001).

Measuring the vocabulary knowledge of a learner is necessary both for teachers and learners because the learners will have a better understanding of their English ability and how to improve it according to their vocabulary knowledge levels while the teachers can adjust the lessons they teach to meet the learners' needs according to the learners' vocabulary knowledge.

An interesting alternative to traditional vocabulary tests might be the YN test because it does not require much time to write and administer (Meara & Buxton, 1987). It does not need a trained item writer because the test does not involve the complicated process of creating distractors (Anderson & Freebody, 1983). In addition, it is also easy for children to answer because there are no multiple choices to confuse them (Anderson & Freebody, 1983). Also, when the administration time is compared with that of a multiple-choice test, the YN test may take only 10 minutes for 100 items while the multiple-choice may take 50 minutes. This means that the YN test can test many more words than multiple-choice or other kinds of tests in a shorter period of time, which is the reason why the test has a higher sampling rate than other forms of tests (Eyckmans, 2004).

Moreover, it has been found that the YN test correlates highly with other standard tests of vocabulary knowledge (e.g. Beeckmans et al., 2001; Eyckmans, 2004; Meara & Buxton, 1987; Mochida & Harrington, 2006) and it is also a good measure of L2 proficiency (Pellicer-Sánchez & Schmitt, 2012; Roche & Harrington, 2013) so it can also be used as a placement test (Harrington & Carrey, 2009; Nation, 2001) as well as a tool to measure the vocabulary size of learners for diagnostic purposes and for use in research (Read, 2007).

Because it is easier to write, less expensive, and quicker both to write and administer when compared with many traditional vocabulary tests, the YN test is a good choice for schools, universities and research use. If creating and administering a vocabulary test is no longer a complicated task, the test can be used more often, as a placement, a pretest, a posttest, or in any step of assessment in any English course or research work. This means that the more frequently the learners' vocabulary is tested, teachers can track the progress of their students more often and they can adjust their courses to suit their students while researchers do not have to use expensive and time-consuming standardized tests to measure their participants' vocabulary size or even their English proficiency.

Nevertheless, there is still a problem about creating nonwords, which are imaginary words added to the YN test to check whether learners are guessing their answers. In other words, there are no clear guidelines on how to create nonwords. It seems that people in the YN test field agree that nonwords should be constructed according to the orthographical and phonological rules of the language (Beeckmans et al, 2001). Anderson & Freebody (1983), who were the first to introduce nonwords in a YN test, created nonwords according to the following two principles: 1) changing one or two letters in a real word (e.g., *tand* from *land*; *sancire* from *sincere*) and 2) adding wrong suffixes to a base word (e.g., *suggestment, adjustion*).

Principle 1 raises the question as to what extent a nonword should differ from its original word. This point interested the researchers and led to the creation of nonwords that were very similar to the original words (i.e., near-homophone nonwords or N1) and to those that were different (i.e., non-homophone nonwords or N2) in order to find out whether they would make any differences to the YN test results.

Therefore, this study aims to explore whether nonwords that are very similar to the original words and those that are not so similar will lead to different test results. If so, this knowledge will contribute to the improvement of nonword creation and it will result in YN tests of a higher quality.

Theoretical Background

The YN Vocabulary Test

The YN vocabulary test (YN test) is a receptive vocabulary size test that presents test takers with a list of individual words and the test takers have to answer *yes* or tick the words they know (Beeckmans et al., 2001; Eyckmans et al., 2007; Nation, 2001; Read, 2000). This kind of test may contain nonwords, which are imaginary words put into the test to check whether a test taker is guessing the correct answers. If test takers answer *yes* to a nonword, their scores will be adjusted downward. Below is an example of a YN test comprising the nonwords *deloy, morder, carcle,* and *gertain*. The possible test instruction could be "Write Y (yes) if you ***know the meaning*** of the word" or "Write Y (yes) if you ***think that the word exists*** in English" as follows:

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

.....announce	.....area	.....deloy	.....escape
.....morder	.....adult	.....damage	.....combine
.....remind	.....tidy	.....pleasure	.....carcle
.....heavy	.....gertain	.....reply	.....clever

YN Test Scoring

There have been four main scoring formulae proposed for use in the field but it is still not clear which one is the most appropriate for the YN test. Each scoring formula varies in the calculation of the test scores and there are four types of answers involved in the process as follows.

1) *hit* => the answer *yes* to a real word
2) *miss* => the answer *no* to a real word
3) *false alarm* => the answer *yes* to a nonword

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

The same concept of the four formulae is that false alarms indicate guesswork and are used to adjust the score downward, while hits roughly represent learners' receptive vocabulary size. The four methods have been compared by some researchers such as Ward (2005), who compared the four formulae and found that these four formulae led to rather different results. He also addressed Meara's suggestion (from his personal communication with Meara) that the simple *hit minus false alarm* formula was appropriate because "All the empirical work we've done suggests that *hit minus false alarm* is as good as anything, and it's easier for people to calculate and understand." (Ward, 2005, p. 30). Some recent studies on YN tests, e.g., Harrington & Carey (2009), also used the simple *hit minus false alarm* formula. In this study, the *hit minus false alarm* was also employed as the YN test scoring method.

Nonwords in the YN Test

Normally, the nonwords used in the YN test are 1) English-like nonwords created by changing one or two letters from their original words such as *dactor* from *doctor* and 2) pseudoderivatives or nonwords that are created by adding wrong suffixes such as *suggestment* or *entertainism.*

There has been no report of using pseudohomophones in the YN test (Thoma, 2011). Pseudohomophones are a type of nonword widely used in psycholinguistics field. They are created to have the same sounds as their original words such as *cerum* from *serum* and the reason why they are not employed in the YN test may be because many researchers might feel that pseudohomophones are too close to real words. For example, Thoma (2011), in his YN 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 the similarity or difference in the sound of a nonword to its original word is a source of variance in a YN test. This leads to the question "To what extent should a nonword be different from a real word?" Thus, two types of nonwords, which are N1 and N2 as mentioned earlier, were created to serve this purpose.

Methods

Instruments

YN tests

The data in this study were the YN tests collected from 600 students. They included real words taken from BNC first- and second-thousand word frequency lists created by Nation (2004). High frequency words are very important because they cover most of the words in spoken and written texts and occur in all kinds of uses of the language (Nation, 2001). High frequency words are also essential for people who learn English for specific purposes because they are the necessary basic words before moving on to more specific words (Nation & Hwang, 1995). To create a vocabulary size test, using words from frequency lists is more systematic because they can tell us which set of words the learners actually know (e.g. the first 1000, or 2000 most frequent words). For L2 learners, who know a narrower range of words than native speakers, sampling of words should come from frequently used words rather than rarely used ones and word frequency lists allow us to do this.

In the present study, however, the purpose is to compare whether N1 and N2 lead to different YN test scores of the test takers, not to measure the test takers' total vocabulary size. Therefore, their vocabulary size, which is their translation scores in this study, is used only as an indicator as to whether each YN test version (i.e., N1 and N2 tests) would yield more accurate scores for their real vocabulary size (i.e., their translation scores).

For this reason, the words used in the YN tests of this study were drawn from the general language use frequency list which is from the BNC first- and second-thousand word frequency lists as mentioned earlier. Nation & Hwang (1995) suggested that the first 2000 high frequency words are suitable for general language use and this level of language use is necessary for L2 learners before they move to other specific types of words that serve their particular needs. Therefore, the vocabulary size of the participants in this study is only the vocabulary size of their general language use (2000 high frequency words), not their total vocabulary size.

The real words in each YN test were only head words, which means words without any inflections (*-s, -ed, -ing*) or derivations (e.g., *-ness,* -

ment, -tion, -ly). All of them are content words; function words are not included in the YN tests created for this study. The researchers randomly drew 50 content words from the first-thousand word frequency level and the other 50 from the second thousand word frequency level.

The YN tests in this study employed both types of test rubrics (i.e., Rubric 1. *yes = know the meaning* and Rubric 2. *yes = the word exists*). That is, half of the N1 YN test version employed Rubric 1 while the other half employed Rubric 2. This is also the same with the N2 YN tests, with half of them using Rubric 1 and the other half using Rubric 2. The reason that both rubrics were employed was to eliminate the possibility that the two different rubrics may be a source of variance that leads to different YN test results. Both test rubrics were 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) หน้าคำที่นักศึกษาทราบความหมาย และเขียน N (No) หน้าคำที่นักศึกษาไม่ทราบความหมาย ในข้อสอบนี้มีคำหลอกอยู่ด้วย

2) Test Rubric 2:

Directions:
Write Y (Yes) in front of the word if you think that the word exists in the 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) หน้าคำที่นักศึกษาคิดว่าเป็นคำที่มีในภาษาอังกฤษ และเขียน N (No) หน้าคำที่นักศึกษาคิดว่าเป็นคำที่ไม่มีในภาษาอังกฤษ

Nonwords

As mentioned earlier, two types of nonwords were created for this study, which were near-homophone nonwords (N1) and non-homophone

nonwords (N2). The term *near-homophone nonwords* (N1) was defined operationally in this study to cover:

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

2) nonwords deemed to be homophonous for most Thai learners with the original words (e.g., *serect* from *select*)

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. This idea comes from the confusion of some English phonemes (i.e., sounds) among Thai students which is the result of the fact that some English phonemes do not exist in the Thai language. The phonemes of Thai that are the same as those of English are: /b/ /d/ /f/ /s/ /h/ /m/ /n/ /l/ /w/ /y/; while the English phonemes that do not exist in Thai are /g/ /v/ /o/ /ð/ /z/ /ʃ/ /ʒ/ /tʃ/ /dʒ/ /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 are /c/ /ch/ /ʔ/ /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 /dʒ/ in English but is pronounced less forcefully, while the phoneme /ʔ/ 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 by these 3 sounds. This is possibly because the /ʃ/ and /tʃ/ sounds do not exist in Thai, so most Thai learners pronounce the sound /ch/, which is the nearest sound they can produce, instead of the actual pronunciation of the /ʃ/ and /tʃ/ sounds. Therefore, the researchers used this confusion between some English and Thai sounds to create near-homophone nonwords by changing one letter from the original word to a letter that might cause confusion; for example, the confusion between /v/ and /w/ (e.g. *vorry* and *worry*); the confusion between /r/(retroflex), /r/ (trill), and /l/ (e.g. *retter* and *letter*); the confusion between /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 (N2), 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*.

The two nonword types also came from the same original words. For example, the near-homophone nonword *lepair* and the non-homophone nonword *depair* were from the same original word *repair*. The

50 original words were also drawn from the first and second-thousand word frequency levels. All the nonwords created for this study can be seen in the appendix A.

Translation Test

The translation test was designed to check whether the participants really knew the items on the YN test. Knowing a word is defined as having the ability to supply one correct L1 translation of the word (Read, 2000). The translation test was used as the criterion to measure the YN tests' concurrent validity. The concurrent validity can be found when comparing one test to another test that is considered reliable. The correlation between the two tests means that if test takers obtain high scores in a YN test, we can predict that they will also get high scores in a translation test. The translation score is considered to be the participants' actual vocabulary size in this present study. This means that if test takers obtain high translation scores, it suggests that their actual vocabulary size is large.

The participants did the YN test first, then the translation test. These two tests were on the same paper. The participants did the YN test in the first period (10 minutes) and then wrote the translation of the words they answered 'yes' to on the same paper in the second period (10 minutes). The translation test was unannounced so the participants did not know at first that they had to translate the words they answered *yes* to. In the first period (of the YN test), the participants had to write with pink pens while in the second period (the translation test) they wrote with green pens. The pink and green pens were provided by the researchers. Erasers or wipeout were not allowed in this test. The participants could only cross out the answers if they wanted to correct them. The researchers employed the pink and green pens in this study because the different colors of the pens would show whether or not the participants returned to correct their answers from the first period. The translation and YN test scores were converted into percentages so that the two types of scores could be compared.

Semi-structured Interview

The semi-structured interview was used to elicit deeper details about the participants' behavior when doing the YN test. The researchers used the stratified random sampling method to select 12 higher-, 12 middle-, and 12 lower-scored students (36 interviewees) from the N1 test takers, and the same proportion from the N2 test takers, so there were 72 interviewees altogether from both the N1 and N2 groups. The interviews were conducted individually in Thai and they were audio-recorded.

There was one interview question and an additional question which analyzed the written data from the 600 translation papers as described below.

Interview Question

The interview question: "*What were the reasons you answered Y (yes) to some real words that you could not translate afterwards?"* was designed to elicit information about the reasons the participants could not translate some real words for which they had answered *yes.* Their answers could provide an insight into their overestimation of their performance. The term *overestimation* (or *overestimate* as used in Mochida & Harrington (2006) and Zhang et al. (2020) means answering *yes* to a real word that a test taker did not really know for which he or she made a guess.

Question analyzed from the written data from 600 translation papers

The question: "*What were the reasons the participants answered Y (yes) to some nonwords?*" aimed to elicit information about why the test takers decided to answer *yes* to some nonwords. This question could also provide deeper details about the participants' guesswork concerning the nonwords. The data were obtained from the translation test. The participants were told that they had to translate all the words (and nonwords) on the translation paper to which they had answered *yes* in the YN test. If they could not translate them, they had to write, for each word, the reason why they could not translate them.

Participants

In this study, the participants were 600 students at a public university in Nakhon Ratchasima, Thailand. They were first-year undergraduate students attending a mandatory basic English course of the university. Three hundred of them did N1 tests, and the other 300 did N2 tests. The English proficiency of the two groups of the participants was similar. That is, each group of 300 consisted of the same proportion of higher (100 students), middle (100 students), and lower English proficiency students (100 students) as measured by their O-NET scores of English, which is a national test used to assess the knowledge of all Thai school students which was written by Thailand's National Institute of Educational Testing Services.

It was possible that the proficiency levels of the participants would affect the test results so the researchers controlled this by mixing participants in each group from the three proficiency levels. The reason for this was a report that some types of nonwords presented problems to some groups of learners. For example, nonwords that were created by adding incorrect suffixes to real words such as *suggestment* (these are called pseudoderivatives) caused some problems to higher proficiency test takers. That is, these test takers tended to choose more pseudoderivatives than other groups of test takers (Beeckmans et al., 2001). This may be because higher L2 proficiency learners were likely to have better word formation knowledge (involving forming words by adding prefixes and suffixes) than those of a lower proficiency and they may guess their answers based on their greater knowledge.

Procedures

The two types of the YN tests were administered to the 600 participants: 300 students took the N1 tests, while the other 300 took the N2 tests. Then all of them did the translation tests. After that, the semi-structured interview was conducted, and the YN tests, the translation tests, and the interview data were analyzed.

Data Analysis

Analysis of N1 and N2 tests' Reliability

In this study, the reliability (the internal consistency) of the N1 and N2 YN tests was explored using Cronbach's Alpha as a statistical tool.

Analysis of N1 and N2 tests' Concurrent Validity

With regard to test validity, N1 and N2 tests were validated by establishing their concurrent validity. That is, the scores of the YN tests were compared with the scores of the translation tests by finding their correlation. As mentioned earlier, the variables investigated in this study were the two nonword types. For this reason, the scores of the two nonword type YN tests (i.e., N1 and N2 tests) were compared to find out which tests had a higher correlation with the translation scores. This correlation was calculated using Pearson Correlation as a statistical tool. A higher correlation would mean that the particular nonword type (either N1 or N2) could make the YN test scores more similar to the actual vocabulary knowledge of the participants because the translation scores indicated the real vocabulary size of the test takers (see 3.1.3). This also meant that the particular nonword type tended to be more appropriate for use in a YN test.

To score the YN tests, the scoring method called *hit minus false alarm* was employed (See section 2.2 for more details). One point is given to a hit, which is the *yes* answer to a real word, and then the number of hits was adjusted downward by subtracting it from the number of false alarms (i.e., the *yes* answer to the nonword). One false alarm means -1 point. After that, the final scores were converted to percentages so that the YN test scores and the translation scores could be compared.

Analysis of the Translation Test

To calculate the translation scores, one point was given to a correct translation of a real word (including all possible senses of the real word). There was only 1 point or 0 points rating (not 0.5) in order to make the points equal to the scoring of the YN test, which was also based on 1 or 0 points. Then these points were converted into percentages.

Analysis of the Interview and Written Data from 600 Translation Papers

As mentioned in 3.1.4, the audio interview data were transcribed and reported in the form of percentages. Also, the written data from the 600 translation papers were gathered and reported in the form of percentages and Thai translations.

Results and Discussion

The results from Cronbach's Alpha calculation showed that both N1(.935) and N2 (.925) tests were reliable because they yielded higher reliability coefficients than .700, which is the acceptable reliability value suggested by Nunnally (1978). See Table 1 below.

Table 1

Reliability Results of N1 And N2 Tests and The Correlation Between N1/N2 Tests and the Translation Test

YN Test	Number of participants	Number of test items	Cronbach's Alpha	Pearson correlation
N1 tests	300	100	.935	.858**
N2 tests	300	100	.925	.858**

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

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

For the concurrent validity results, it was found that the N1 and N2 tests had the same correlation coefficient (.858**) between the YN and translation scores (see Table 1), but the N1 tests had closer scores to the translation test (N1=52.81/Trans.=47.60) than the N2 tests (N2=56.96/Trans.=46.20) (see Table 2). As mentioned earlier, the translation scores indicate the actual vocabulary size of the participants in this study.
This suggests that the N1 and N2 tests are equal in predicting the relative vocabulary size of the participants in terms of ranking, but with regard to predicting exact vocabulary size, the N1 tests tend to be better because the N1 test scores are closer to the translation scores than those of N2. From Table 2, the difference in the scores of the N1 tests and the translation tests are 5.21% while the N2 and the translation tests are 10.76%.

Table 2

Mean Difference Between YN Test Scores and Translation Scores of N1 and N2 Tests

YN tests	N	Mean YN test score	Sig. (2- tailed) (N1 YN tests vs N2 YN tests)	Mean translation score	Mean difference between YN and translation scores
N1 tests	300	52.81	.037*	47.60	5.21
N2 tests	300	56.96		46.20	10.76

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

Actually, for the purpose of placement, the correlation coefficient between YN and translation scores should be a better indicator of the quality of a YN test in predicting a test taker's actual vocabulary size than how close a YN score is to a translation score. Apart from a good correlation coefficient between YN and translation scores, it is also ideal if the scores of a YN test are very close to the translation scores, which means that the YN test version could closely predict the actual vocabulary size of a test taker. However, there have been no reports of any YN test version that can demonstrate the exact vocabulary size of a test taker. For example, Barrow et al. (1999) reported a relatively high rate of vocabulary size overestimation 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 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 vocabulary size 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, firstly, the correlation coefficient between the YN and translation scores and, secondly, the closeness of the YN scores to the translation scores. When the results show that the correlation coefficient between the YN and translation scores of N1 and N2 tests 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 better results than N2 tests because they produce closer scores to the translation scores than N2.

The reason why the N1 test yields closer scores to the translation test (5.21% higher than the translation test) while N2 test scores are comparatively higher than the translation scores (10.76% higher) could be because the N1 tests produced a significantly higher number of false alarms (i.e., the answer *yes* to a nonword) than N2 (23.22% and 20.08% respectively) (p=.036*), so the scores of the N1 tests were adjusted downward more than

those of N2. Table 3 below illustrates the difference between the false alarm rate of N1 and N2 tests.

Table 3

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 the participants answered *yes* to N1 more than N2 is assumed to be because the nonwords in N1 were created to have the same sounds as real words (NB for Thai L1 learner). Therefore, the participants may have confused them with real words, while the N2 words were created to have different sounds which may have reduced confusion.

The results of this study also showed that the main reason for most participants to overestimate their vocabulary size was because they had partial knowledge of some words which means they were not guessing blindly. This can be seen in the summary of findings from the interview question "What were the reasons you answered Y (yes) to some real words that you could not translate or wrongly translated afterwards?" (See Table 4), which shows that almost all of the interviewees (70 out of 72 or 97.22%) said that they answered *yes* to some real words they could not translate afterwards because they had seen, heard, or felt those words were familiar (partial knowledge overestimation), while only 24 out of 72 (33.33%) reported that they had made a blind guess.

Table 4

Summary of Findings from The Interview Question "What Were The Reasons You Answered Y (Yes) to Some Real Words that You Could Not Translate or Wrongly Translated Afterwards?"

Reasons	Total interviewees (out of 72)	Percentages
1. Partial knowledge overestimation (i.e. having seen, heard, or felt the nonwords were familiar)	70	97.22%
2. Guessing	24	33.33%

Note: Participants could have had more than one reason for answering yes to some real words that he or she could not translate or wrongly translated afterwards.

Partial word knowledge means learners have some knowledge of a word but their knowledge is phonologically, orthographically, and/or semantically incomplete. Word knowledge is a gradual process starting from totally unknown, to different extents of partial knowledge, to complete knowledge (Durso & Shore, 1991; Nation, 2001). In their studies, words that the participants claimed that they had seen or heard even if they had no idea of their meanings were considered partially known words. Similarly, in this present study, the definition of partial knowledge is also that the participants who have partial knowledge of a word means that they had seen, heard, or felt familiar with the word. Thus, the term *partial knowledge overestimation* in this study means the participants did not have precise knowledge of a word, and therefore, answered *yes* to some words that sounded familiar to them. The term *guessing*, on the other hand, means they had no knowledge of the word but guessed at random.

Some examples of the participants mistakenly using their partial knowledge can be seen in the wrong translations the participants gave to some real words. These wrong translations come from many kinds of partial knowledge overestimation such as confusing *commerce* with *finance* (close in 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). (See Table 5)

Table 5

Examples of Wrong Translations of Some Real Words

Student No.	O–NET score of English (out of 100)	Words	Wrong Translation
1	85	pleasure	ความดี / น้ำใจ (merit)
		commerce	การเงิน (finance)
2	82	-	-
24	26	active	ท่าทาง (gesture)
25	26	become	กลับมา (come back)
		formal	สูตร (formula)
		active	กระตุ้น (activate)
		marry	สุขสันต์ (merry)
		minus	น้อย (little or few)
48	16	minus	น้อย (little or few)

This kind of overestimation was also found in the work of Anderson & Freebody (1983). They called this phenomenon "mock" hits (i.e., the answer *yes* to a real word when the test takers think they know the meaning of a word, but actually they confuse the word with another word or have only partial knowledge of that word). For example, in their study, *sham* was confused with *shame.* They also found that the YN test scores were inflated because of this mock hit phenomenon. Ward (2005) also noted the occurrence of mock hits. 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*.

With regard to nonwords, it was also found that the participants answered *yes* to some nonwords because of the 2 reasons mentioned above: partial knowledge overestimation and guessing. The findings from the question "What were the reasons the participants answered *yes* to some nonwords?" (See Table 6 below) revealed that the participants answered *yes* to N1 because they had partial knowledge of them (64.25%) rather than that they were blind guessing (35.75%) while they answered *yes* to N2 mainly because they just guessed (75.57%) rather than making a partial knowledge overestimation (24.43%). We do not claim that the distinction between random guessing and partial knowledge is absolute or that the difference is

clear in all the participants' minds. But the large difference in the responses, allied to the fact that the interviews were conducted in Thai, do suggest that the participants were aware of the distinction.

Table 6

Summary of Findings from the Question: What Were the Reasons the Participants Answered Y (Yes) to Some Nonwords?

Answers	N1	N2
1. Partial knowledge overestimation (i.e. having seen, heard, or felt familiar with the nonword)	64.25%	24.43%
2. Guessing	35.75%	75.57%

As mentioned earlier, the participants in this study tended to overestimate their vocabulary size because they have some partial knowledge of those words rather than guessing at random. Therefore, this partial knowledge overestimation should be adjusted downward by the nonwords that can also attract those who have this kind of overestimation. As reported above, N1 can attract those who use their partial knowledge more than blind guessing, so the N1 false alarm rate is higher than N2, which means that N1 can adjust downward the overestimation rate better than N2, and this explains why the N1 test scores are closer to the translation scores while the N2 test scores are considerably higher than the translation scores.

The question "To what extent should nonwords be different from real words?", which was raised by Beeckmans et al. (2001) and may also be questioned by many scholars in the YN test field, could possibly be answered by the results of this present study, which suggest that nonwords that are created to be very close to real words like N1 tend to make a YN test better in predicting the actual vocabulary size of a test taker than those created with different sounds like N2.

The reason why N1 tends to lead to better YN test results could be that N1 reflects the problems of the participants' real language use. That is, those who tend to have some problems with real words are likely to choose nonwords that reflect those 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 mastered only one or two aspects of the 3 aspects of word knowledge: the spelling, the sound, or the meaning. If they answer *yes* to real words because they have this kind of partial knowledge, they will, therefore, be likely to answer *yes* to a nonword for the same reason. On the other hand, those who have more precise knowledge of a real word (i.e., its spelling, sound, and meaning) tend to be able to reject N1 because they recognize that the N1 examples are similar to real words but have spelling mistakes.

Table 7

Mean Difference of False Alarm Rate Between Real Homophone and Near Homophone Nonwords

Nonword type	Number of nonwords	Mean false alarm rate (%)	SD	Sig. (2-tailed)
Real homophone nonwords	32	26.35	15.80	.239
Near homophone nonwords	18	21.02	13.90	

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

As for the N1 test, which was a combination of both real homophone nonwords (e.g. *persent* from the real word *percent*) and near homophone nonwords (e.g. *pergent* from the real word *percent*), the comparison of these two nonword types was used to determine which type attracted the test takers more. It can be seen from Table 7 that the participants tended to answer *yes* to the real homophone nonwords (26.35%) more than the near homophone nonwords (21.02%) although they were not significantly different (p=.239). This means that the real homophone nonwords are likely to cause more confusion than the near homophone ones because they sound completely the same as the original real words while there were still some test takers who were able to differentiate between the near homophone nonwords and their original real words because there were still some differences in the sounds even though they were pronounced nearly the same.

It is noted that the concurrent validity of the N1 and N2 tests of this study (.858) was relatively high compared to the YN test validity of other studies such as *r* = .84 of Anderson & Freebody (1983); *r* = .703 of Meara & Buxton (1987); *r* = .389-.596 of Eyckmans (2004, Experiment 2); *r* = .348-.862 of Eyckmans (2004, Experiment 3); *r* = .244-.666 of Eyckmans (2004, Experiment 4); *r* = .384-.594 of Eyckmans (2004, Experiment 7); *r* = .8 of Mochida & Harrington (2006); and *r* = .57-.62 (the correlation between YN test scores and the MC VST test scores) or *r* = .70-.77 (the correlation between YN test scores and the translation scores) of Zhang et al. (2020). The relatively strong correlation of the YN tests in this present study could possibly be explained by certain factors such as the well-managed proficiency levels of the participants in this study (i.e., each YN test version was taken by the same proportion of higher-, middle-, and lower-scored students), which could have led to a more accurate result if the proficiency levels of the participants proved a source of variance for the YN test results (See 3.2 for more information).

Another factor might be that the YN test validity of each study was obtained differently. That is, some studies compared the YN test scores with the translation test scores while others compared the YN test scores with other kinds of vocabulary size tests such as the MC VST, which may have led to different validity results. This can be seen in the study of Zhang et al. (2020), who compared the YN test scores with both a translation and the MC VST test and found that the validity was higher when the YN test scores were compared with the translation scores (.70-.77) than when compared with the MC VST scores (.57-.62). They mentioned that the vocabulary size scores in the MC VST could have been affected by blind guessing while the use of guessing could have been less in a translation test (Stewart, 2014). In this present study, the YN test validity came from a comparison of the YN test scores and the translation scores, which might explain the relatively strong correlation.

Conclusion and Recommendations

The results of this study suggest that, although the N1 and N2 tests are equally able to predict the participants' vocabulary knowledge in terms of ranking, an N1 test is likely to be better than an N2 test because it can better predict the actual vocabulary size of the participants. This can be seen from the closer N1 test scores compared to the translation scores. The reason why

the N1 test scores were closer to the translation scores may be because the N1 tests produced significantly higher false alarms than N2, so the scores of N1 tests were adjusted downward more than those of N2 and, therefore, were closer to the translation scores. This suggests that N1 could have caused more confusion with the real words because the N1 nonwords were created to have the same sounds as real words while the N2 nonwords were created to have different sounds.

It was also found that the N1 test would have caused more confusion to those who overestimated their vocabulary size because of their partial knowledge. That is, the reason the participants answered *yes* to the N1 nonwords was because they had some partial knowledge of them rather than because they just blindly guessed, while the answer *yes* to the N2 nonwords were mainly due to random guessing more than partial knowledge overestimation. The results of the interview question also revealed that the overestimation rate in this present study mainly derived from partial knowledge overestimation more than random guessing.

However, creating an N1 test is not always easy. This is because the N1 nonwords were created to have the same (or almost the same) sound as real words, so a test writer has to find an appropriate original real word that can be changed to a nonword with the same sound. For instance, in a YN test of English target words, to create a near-homophone nonword, a test writer needs 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 maintains the same sound such as the original word *serum*, which can then be changed to the nonword *cerum*. A word like *depend* may not be a good choice because it is not easy to find a homophone for this word.

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

As regards the N2 test, it is very easy to create when compared to N1. That is, an N2 test can be created by changing only one letter from any real word. For example, the 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 as in N1. If test writers are not concerned about the inflation of the N2 scores compared to the N1 scores, N2 examples area good alternative because they

are easy to create and can lead to equally good YN test results in terms of ranking.

As for future research, it is possible that test takers from different L1 backgrounds might react differently to nonwords. That is, a nonword may be more attractive to a learner of a particular L1 background than to those of a different L1 background. For example, Meara and Buxton (1987) gave an example of the nonword *observement*, which is similar to a real word in Italian or French but not in German. This means that it is easier for German speaking test takers to reject this nonword when compared with French or Italian test takers. Also, Cobb (2000) mentioned that the YN test did not work well with Arabic-speaking learners, who answered *yes* to a large number of nonwords because they were likely to be blind to the vowels. Therefore, they tended to confuse *tilt* and *toilet* or *mascarate* and *miscreate* (Ryan & Meara 1991). In this present study, near homophone nonwords seem to be attractive for Thai speaking learners of English, but it is possible that results may be different with learners of other L1 backgrounds. Therefore, further studies are needed to give a clearer picture as to whether the near homophone nonwords would also work well with other L1 background test takers.

As for the limitations of this study, readers should be cautious about making generalizations from these results because the participants of this study were 600 first-year university students of a public university in Nakhon Ratchasima, which means that they may not be representative of other levels, areas, or L1 backgrounds. In addition, there might also be other factors such as familiarity with the learners' previous syllabuses or familiarity with this YN test format that could lead to different results from those of the present study.

Finally, the YN test format has some advantages over other vocabulary size tests because it is easy to write and administer and can test a large number of words in a short period of time, which is very practical for teachers and ideal for research. Although there are some inconsistencies on how to create a YN test such as which kinds of nonwords or scoring methods should be used, the YN test still remains an interesting topic for researchers to explore because of its potential and convenience as a vocabulary size test.

Acknowledgements

This work would not have been completed without financial support from the Thailand Research Fund (TRF) and Suranaree University of

Technology through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0349/2551 to Wallapha Wongsirichan and Jeremy Ward). We also wish to acknowledge the lecturers and staff of the Department of Foreign Languages, Suranaree University of Technology for all their valuable guidance and support.

About the Authors

Wallapha Wongsirichan: A Ph.D. student in the School of Foreign Languages, Suranaree University of Technology, Thailand. Her research interests include vocabulary learning, teaching, and testing.

Anchalee Wannaruk: An associate professor in Second Language Acquisition and Teacher Education, School of Foreign Languages, Suranaree University of Technology, Thailand. Her main interests include English for specific purposes, discourse analysis, pragmatics and corpus linguistics.

Jeremy Ward: An associate professor in ELT and Applied Linguistics, School of Foreign Languages, Suranaree University of Technology, Thailand. His research interests mainly lie in vocabulary, English for engineering, English for academic and specific purposes.

References

Anderson, R. C., & 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, JAI Press.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System, 27*(2), 223-247. http://doi:10.1016/S0346-251X(99)00018-4

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the YN vocabulary test: some methodological issues in theory and practice. *Language Testing*, 18, 235-274.

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review, 57*(2), 295-324.

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, *120*, 190-202.

Eyckmans, J. (2004). Measuring receptive vocabulary size. Unpublished Doctoral Dissertation. LOT.

Eyckmans, J., Van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behaviour in YN vocabulary tests. In H. Daller, J. Milton and J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 59-76). Cambridge University Press.

Harrington, M. & Carey, M. (2009). The on-line YN 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 University Press.

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

Mochida, K., & Harrington, M. (2006). The YN test as a measure of receptive vocabulary knowledge. *Language Testing, 23*(1), 73–98.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* 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). John Benjamins.

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes‐No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489‐509. https://doi.org/10.1177/0265532212438053

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513–536.

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

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7*, 105-125.

Roche, T. & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance. *Language Testing in Asia*, *3*(12), 1-13.

Ryan, A. & Meara, P. (1991) The case of the invisible vowels: Arabic speakers reading English words. *Reading in a Foreign Language*, *7*(2), 531-540.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271‑282.

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in yes–no vocabulary tests change with Japanese university students' ability levels? *Language Testing, 29*(4), 471–488.

Thoma, D. (2011). *Strategic attention in language testing: Metacognition in a YN business English vocabulary test.* 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.

Wongkositkul, A. (1993). A comparison of English and Thai phonological systems for teaching Thai to foreigners. *Journal of Language and Culture, 12*(1), 86–93. (in Thai).

Zhang, X., Liu, J., & Al, H. (2020). Pseudowords and guessing in the YN format vocabulary test. *Language Testing, 37*(1), 6–30. http://doi: 10.1177/0265532219862265

Appendix A

50 Type-1 Nonwords (N1) and 50 Type-2 Nonwords (N2)

Item	Original real words	N1	N2
1	advance	advanse	adhance
2	begin	bigin	bugin
3	biscuit	bisguit	discuit
4	cancel	cansel	canbel
5	certain	sertain	gertain
6	circle	cercle	carcle
7	channel	shannel	chanbel
8	chapter	shapter	chaster
9	confirm	conferm	confarm
10	corner	cornor	cirner
11	degree	digree	dogree
12	delay	deley	deloy

13	display	displey	disploy
14	factor	factur	fictor
15	finish	finich	rinish
16	label	rabel	pabel
17	letter	retter	metter
18	machine	mashine	mathine
19	market	marget	mardet
20	maybe	meybe	moybe
21	mirror	mirrer	morror
22	murder	merder	morder
23	music	musig	musin
24	office	offise	offine
25	percent	persent	pergent
26	period	piriod	pariod
27	person	purson	porson
28	prefer	prefur	profer
29	purchase	purshase	surchase
30	repair	lepair	depair
31	return	retern	retorn
32	secret	sicret	socret
33	select	serect	sedect
34	serious	selious	sedious
35	service	serwice	serdice
36	shoulder	shouldir	shoilder
37	signal	sicnal	sinnal
38	sister	sistor	soster
39	surface	surfase	purface
40	survey	survay	survoy
41	system	sistem	sustem
42	table	teble	toble
43	ticket	tigket	ricket
44	traffic	traffig	traffin
45	trousers	trousors	troisers
46	victim	wictim	bictim
47	village	willage	cillage
48	western	westurn	wostern
49	window	vindow	lindow
50	worry	vorry	porry