学校教育における第２言語の語彙測定の基本的問題
Fundamental Issues in L2 Vocabulary Assessment in Academic Settings

中村　優治 NAKAMURA, Yuji

● 慶應義塾大学
　Keio University

Keywords　語彙測定，語彙知識，語彙力の広さ，語彙力の深さ，言語テスト，英語読解力
vocabulary assessment, word knowledge, width, depth, language testing, reading comprehension

ABSTRACT

　語彙力は母国語においても，外国語においても読解力には欠かせないものである。語彙力をつけるための指導に関しては多くの議論がなされているが，語彙力の測定に関してはまだ十分とは言えない。読解力の測定が常に何らかの形で語彙力と強い結びつきを持っているとされながら，語彙力の測定の知識や技術はまだ限界があるようである。語彙力測定の際にはそもそも語の知識，語彙力がなんであるかを知る必要がある。良いテストの作成には測られているものの構成要素を確定することが求められる。本稿の目的はこれらのことを踏まえてよい語彙テストを作成するための基本事項をまず議論し，その中でも学校教育で実施可能な語彙テストづくりに必要な問題点を明らかにすることである。また，語彙テストの実施例としてプレイスメントテストの語彙部分の結果の検証も合わせて行う。

　Knowledge of vocabulary is essential to proficient reading in both L2 and L1. While teaching vocabulary has been discussed frequently in academic settings, testing vocabulary has received less focus. Whereas assessment of reading and reading comprehension has always been linked to some form of vocabulary assessment, knowledge and understanding of the principles and techniques of vocabulary assessment among teachers has generally been limited. When we test vocabulary, we should know what is meant by a word and how to use the word properly. It is essential for the construction of any good test that measures vocabulary knowledge to measure exactly what is intended. It is also important to know how vocabulary knowledge can be measured and what qualities a good test should have. The purpose of the present study is twofold: 1) to discuss important issues (a. discrete or embedded; b. validity and reliability; c. productive or receptive; d. breadth or depth; e. word or collocations) in L2 vocabulary

assessment in academic settings, and 2) to analyze the vocabulary section of a placement test.

I. Purpose of study

It can be argued that vocabulary knowledge is essential to measuring proficient reading levels in both L2 and L1. While there has been extensive discussion about the importance of teaching vocabulary in academic settings, testing vocabulary itself has received less attention. Currently, the validity of vocabulary testing as a distinct construct is a matter of debate, whereas the assessment of reading and reading comprehension has consistently been linked to some form of vocabulary measurement (Ekbatani, 2011).

Minton (2009) argues that when vocabulary is being tested, the essence of the words should be apparent to the examinee. When constructing a test that seeks to measure vocabulary knowledge, it is essential for the test to explicitly define what is being measured. Further, in designing the exam, a thorough understanding of how vocabulary knowledge can be effectively measured and what constitutes a good test should be investigated.

Read (2000) presents two opposing viewpoints on the role of vocabulary in language assessment. One view suggests that it is entirely reasonable to create tests that measure vocabulary meaning and the correct usage of a set of vocabulary items. Another view is that vocabulary is only assessed in the context of a language-use task, where vocabulary interacts in a natural way with other elements of language knowledge.

We need to consider these two viewpoints in our academic test preparations.

The purpose of the present study is two fold: 1) to discuss important issues (a. discrete or embedded; b. validity and reliability; c. productive or receptive; d. breadth or depth; e. word or collocations) in L2 vocabulary assessment in academic settings, and 2) to analyze the vocabulary section of a placement test.

II. Defining word knowledge

According to Minton (2009) one common convention is to divide word knowledge into receptive or passive knowledge and productive or active knowledge, which suggests that a learner's receptive knowledge (e.g. words that are recognized when heard or read) is greater than a learner's productive knowledge (e.g. words that can be called to mind and used in speech or writing).

Another convention is to distinguish between breadth of word knowledge and depth of word knowledge. Breadth of knowledge refers to the number of words a learner knows and depth of knowledge refers to what the learner knows about these words (Minton 2009).

As Minton (2009) suggests, simple binary divisions like receptive and productive, or breadth and depth do not really explain the complexity of word knowledge. Nation (2001) divides word knowledge into three areas: 1.knowledge of form; 2.knowledge of meaning; and 3. knowledge of use.

Knowledge of word is threefold: 1) the written form of a word; 2) the phonological form; and 3) the knowledge of word parts. Knowledge of word meaning has sub-divisions, such as form and meaning or concepts, referents and associations. Knowledge of word use has sub-divisions such as grammatical functions that concerns knowing what part of speech a word is and how it will link with other words as a consequence (Minton 2009).

The sub-division of collocations refers to the company words like to keep (Minton 2009). This is one of the challenging areas to deal with in vocabulary issues, because some words occur very

frequently alongside others and these words are said to collocate with one another.

Still another model contrasts breadth and depth against a quality of fluency. The assumption of this idea is that breadth and depth are aspects of passive word knowledge, while fluency refers to the productive word knowledge a user has (Minton 2009).

III. Measuring vocabulary knowledge

How can vocabulary tests that are both reliable and valid, and also gain greater face validity be constructed? Minton (2009) claims that there are two main issues to consider when constructing vocabulary tests: 1. selection of words for measurement, examination or counting; and 2. testing learners' knowledge of word usage. To deal with the first question, vocabulary tests and other assessments make use of word frequency data and test the most frequent vocabulary.

Concerning the second question, in order to test different aspects of word knowledge, multiple methods are needed. A test of a learner's receptive vocabulary knowledge will require the test writer to select words. A test of productive knowledge will require a technique that can elicit vocabulary (Minton 2009).

Schmitt (2010) maintains that in discussing vocabulary measurement, it is useful to explore ways in which measurement formats differ. He refers to the proposal suggested by Read (2000) that utilizes three different format dimensions of vocabulary assessments.

Schmitt (2010) compares embedded with discrete dimensions and suggests that tests which focus specifically on vocabulary knowledge are likely to be discrete in the sense that particular lexical items are highlighted. However, vocabulary measures can be a component of measures of broader linguistic proficiency, and in this case the test would be embedded. He then discusses the contrast between receptive and productive vocabulary measures. Receptive vocabulary measures are typically selective, because the test writer needs to select the lexical items to measure, determine their characteristics, and write test items for them. On the other hand, a measure of the complete vocabulary output of learners' speaking or writing production would be comprehensive. If this is free output, it poses difficulties for the tester, as there is no way to know in advance exactly what the will be included on the vocabulary exam.

In terms of context, vocabulary items can range from completely context-independent (e.g. an L2-L1 translation task) to completely context-dependent (e.g. define the target word according to the meaning sense used in a certain passage). As Schmitt (2010) claims, context-dependent formats should provide a potentially more thorough method of tapping into the 'contextualized ' facets of word knowledge, such as collocation and register.

IV. Discussion of Two-Dimensional Issues for a Receptive Test

The placement test that will be analyzed later in this paper is an example of a receptive test; therefore, two-dimensional issues will be taken and examined in order to broaden the general understanding of the role that vocabulary tests might play in language assessment: 1) discrete vs. embedded; and 2) context-independent vs. context-dependent.

The discrete vs. embedded dimension

Read (2000) asserts that the embedded/discrete distinction does not refer to the method of testing vocabulary; rather it differentiates between tests that treat vocabulary as an independent construct.

Ekbatani (2011) further states that discrete tests treat vocabulary knowledge as an independent construct, separated from other components of

language. An example might be a teacher who wants to monitor his or her student's progress by testing and retesting words that are repeatedly taught in class over a period of time. On the other hand, as Ekbatani (2011) claims, an embedded vocabulary measure test is one that measures vocabulary knowledge as a part of another skill, such as reading comprehension or writing. In the embedded format, vocabulary knowledge does not receive a separate score, but it is stated as one of the main features of a larger construct.

The Context-Independent vs. Context-Dependent Dimension

Traditionally, the contextualization of vocabulary referred to the way in which a word was introduced to test-takers, it differentiated between presenting sentence target words and presenting isolated target words. The current perspective on contextualization, however, has broadened the concept to include assessing the degree to which the learners can use the given context to correctly ascertain the meaning of unfamiliar word (Ekbatani 2011).

Context-dependent assessment of vocabulary is especially effective in testing words with multiple meanings or low frequency words occasionally appearing in unabridged texts (Ekbatani 2011). Knowledge of vocabulary is essential to the development and demonstration of linguistic skills, but that does not necessarily mean that it should be tested separately.

Valid assessment of vocabulary knowledge according to current thinking is based on the degree to which the learners can use a wide range of vocabulary words to express themselves both in writing and speaking, and to understand the main points in reading and listening. For this reason, many standardized tests such as the iBT and TOEFL exercise the embedded approach, where lexical ability is assessed as part of a larger construct (i.e. in general, reading) (Ekbatani, 2011).

Nevertheless, Hughes (2003) argues, in institutional testing the discrete point approach does play an integral role in achievement tests that seek to measure the extent of vocabulary learned throughout the course.

V. Reliability and validity of a vocabulary test

When constructing vocabulary tests there are two primary issues to consider: 1. reliability; and 2. validity.

Reliability is the ability of a test to measure something consistently and accurately (Minton 2009). It is whether the testing instruments give consistent (i.e. reliable) results. Because reliability is essential to valid testing, reliability should be determined for all our instruments, and reported (Schimitt 2010).

If a vocabulary test is administered twice in the same afternoon when the learner's vocabulary should not have changed significantly, then it should offer similar results on both exams. Judging the reliability of a test might also include equivalence estimates (e.g. whether different forms of a test compare well and produce equivalent results). As a general rule, so-called objective testing of vocabulary, multiple-choice and forced answer tests, has good reliability as measured by test and retest methods, and some methods can also produce good equivalence scores. Currently, there is a great deal of focus on creating objective measures for language produced by learners in written essays or oral examinations, using vocabulary richness scores (Minton 2009).

The reliability of most research instruments can be established using different methods. The most recognized conceptualization is the test-retest method (e.g. a test is given once and then again before any significant learning can occur). In practice, reliability is established via an internal consistency approach. Instead of giving a test twice

as in the test-retest method, it is given only once, but the split into smaller parts which can then be compared (Schmitt 2010). For the purposes of the current work, validity can be said to ascertain whether a test measures what it is supposed to measure, which can be a complex area with various issues to be examined (Minton 2009).

Content validity considers whether a test has the necessary and appropriate content to measure what it is intended to measure. Frequently, tests of vocabulary breadth make use of frequency information as the basis of word selection for testing. It is suggested that depth may not really exist as a separate construct and is an extension of breadth of knowledge, as tests of the two correlate closely. Individual aspects of vocabulary depth, such as knowledge of collocations or idioms, tend to be measured separately as a result (2009 Minton). This suggests that the validity of a vocabulary test will usually have to be demonstrated through its own development and performance. The development part starts with specifying the content. It is about what lexical items the test includes and what is being measured about those items. According to Schmitt (2010), some of the specifications will include:

1) whether the test measures only a specific set of lexical items, or whether the lexical items on the test are supposed to represent a wider population of vocabulary;
2) if the test items represent a wider population;
3) what word knowledge aspects are being addressed; and
4) whether the tests measures recall/recognition/receptive/productive levels of mastery.

It should be possible to develop detailed and focused specifications for new vocabulary tests, based on the literature in the field and on previous research. As Schmitt (2010) states, after test items have been written for these specifications, it is time to investigate how well the test captures the specified content. That is, how well examinees' scores represent this content when they take the test.

Construct validity, which is often closely associated with content validity, considers whether the test measures the construct or skill for which it was designed. Measures of productive knowledge also need a method of analyzing this output that fairly and accurately describes vocabulary knowledge (Minton 2009). One common way is through criterion validity. In this method, a new test is judged according to how closely it correlates with an already established measure. This can work well if an accepted standard measure already exists to compare against. However, in vocabulary, few such standards exist. Moreover, the complex nature of vocabulary knowledge dictates that any particular test would be severely limited as a criterion measure. A criterion validity approach has serious limitations at the moment (Schmitt 2010).

The question of face validity is whether the test is credible to users as a test of what it is supposed to measure. Vocabulary tests can excite surprising passions in users, and even tests with good construct and content validity can be challenged by learners. Learners can have very firm ideas as to what a language test should be like, and these tests do not always involve explicit vocabulary measurement. Where test writers have used frequency data and produced carefully targeted tests of vocabulary knowledge, the comparatively small scale and simplicity of the tests can often raise doubts in the minds of users. The potential benefits of a short and simple test can be lost on users (Minton 2009).

VI. An analysis of a placement test

Keio university's Faculty of Letters has administered our in-house placement test to incoming freshman students and new sophomore

students since the spring of 2006. This placement test measures students' reading ability in English and overall proficiency in order to provide streamed instruction appropriate to their proficiency levels. The test has four components: 1. grammar, 2. vocabulary, 3. cloze and 4. reading comprehension. Students' vocabulary knowledge is measured mainly in the vocabulary section as well as in reading comprehension or cloze sections.

Lexical knowledge is measured in the vocabulary test that should cover the basic and necessary words that students will encounter in their major study reading in English. This test aims to measure students' knowledge of words that refer to the ability to comprehend the meaning of words in an appropriate context for academic purposes.

Test Construction

Materials were selected via.

1) instructor teaching experience

2) reading sections of existing tests

3) linguistic theories (cf. Hughes, 2003)

4) Mita campus reading requirements

5) text books from students' major area of study

The vocabulary items were based on word frequency counts using as a benchmark the English-Japanese dictionaries available at bookstores. These included textbooks authorized by the Ministry of Education, Sports and Science.

Results and Discussion

1) Mean scores of each test (based on raw scores)

Table 1 shows the change of the mean scores of each test (vocabulary) for five years (PT1 through PT5). The test results are based on the raw scores. N.B. PT was conducted at the beginning of the academic year and CT was at the end of the academic year.

2) Vocabulary section of the placement test and the distribution of frequency of the logit scores

Figures V1, V2 and V3 show the distribution of frequency of the logit scores in the vocabulary

Table 1
Mean scores of vocabulary tests based on raw scores

Test Form	N	Vocabulary/10
PT1	853	5.24
CT1	790	6.08
PT2	856	6.03
CT2	830	4.85
PT3	841	5.56
CT3*	794	5.61
PT4	830	6.61
CT4	768	5.13
PT5	816	7.25

section of each placement test.

Figure V1

Figure V2

2010-4 PT5 VOCABULARY

2011-4 RPT1 VOCABULARY

2011-2 CT5 VOCABULARY
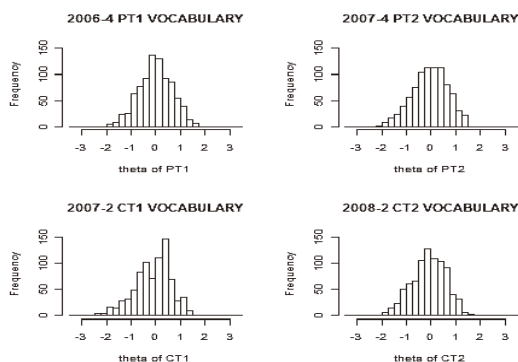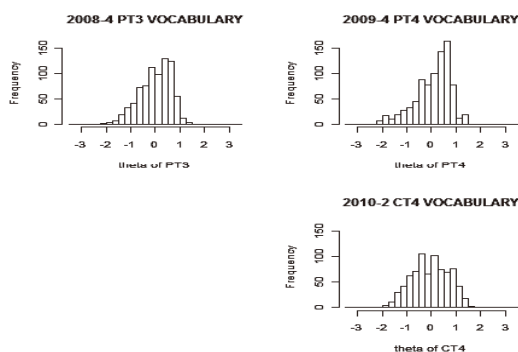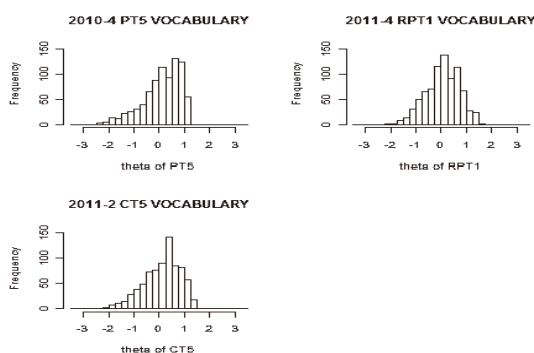
Figure V3

Figures V1, V2 and V3 indicate that there was no noticeable increase in the results of the vocabulary section when results were compared. Moreover, students' vocabulary abilities actually exhibited annual declines for each individual academic year. This trend suggests that because there was no specific vocabulary building courses during their university tenure, students' vocabulary knowledge peaked for the purposes of taking the entrance examination, but after entering university the retention of the learned vocabulary gradually faded. Instead, student vocabulary that focused on textbook reading improved even if the width of the vocabulary did not show a significant increase.

The data suggests that entering students' vocabulary knowledge has improved over the last five years.

Summary of the placement test results and suggestions for future research

We can summarize the test results as follows:

1. The analysis shows that there was no remarkable teaching or learning effect on students' grammar and vocabulary knowledge. This is probably because in university, unlike high school, there were no specific grammar courses or vocabulary building courses.

2. Vocabulary knowledge could be more precisely measured by using an appropriate corpus to establish an ideal or practical objective to master.

3. Communicative vocabulary knowledge should be investigated in performance tests (e.g. speaking and writing skills) where students are expected to express productive skill, which were not measured in the present tests.

VII. Suggestions for a better measurement of vocabulary knowledge

In conclusion, for our better measurement of vocabulary knowledge and ability, we can refer to Nation (2008 pp.153-154) where he suggests that a good vocabulary test has the following features:

1. It is reliable. That is, it continually offers similar results for the same person even if different people administer the test, it is taken under slightly different conditions, or different people mark it. Reliability is helped if:

a. the test contains at least 30 items or points of assessment

b. the test format is familiar to the learners because they have taken such a test before

c. the instructions and way of answering are the same in all versions of test, and

d. the marking uses a marking key and criteria that take account of most possible variations in answering

2. It is valid; that is, it measures what it is supposed to measure. Validity is helped if

a. the test is used for the purpose for which it was designed,

b. the knowledge and skills learners use to take the test are as close as possible to the knowledge and skills involved in the vocabulary knowledge being measured,

c. the test is suited to the level of the learners, and

d. the learners take the test seriously, try their best, and do not cheat.

3. It is practical and easy to use. Practicality is

helped if

    a. it does not take along time and a lot of skill to make the test, or the test is ready-made,

    b. it does not take a long time to take the test,

    c. it is easy to mark the test

    d. it is easy to interpret the score on the test and

    e. it is not expensive to make copies of the test for the learners to take.

References

Douglas, D. (2010). *Understanding language testing*. UK: Hodder Education.

Ekbatani, G. (2011). *Measurement and evaluation in post-secondary ESL.* New York: Routledge.

Fulcher, G. (2010). *Practical language testing*. UK: Hodder Education.

Hughes, A. (2003). *Testing for language teachers* (20.1[nd] ed.). Cambridge UK : Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol UK: Multilingual Matters.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge UK: Cambridge University Press.

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

Nation, I.S.P. (2008). Teaching vocabulary: strategies and techniques. Boston: Heinle

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

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146-161.

Schmitt, N. (2010). *Researching vocabulary: a vocabulary research manual.* Oxford: Palgrave Macmillan.