Language Testing

Citation

Mizumoto, A., Sasao, Y., Webb, S. (in press). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*. doi:10.1177/0265532217725776 Retrieved from http://mizumot.com/files/cat-wplt.pdf

Introduction

Vocabulary knowledge is no doubt crucial in the development of second or foreign language learning. Although it was once referred as a neglected aspect of language learning (Meara, 1980), vocabulary has been much researched in the field of applied linguistics over the last few decades (Nation, 2013). Among a wide range of topics in vocabulary research, the scope in teaching vocabulary remains a focus for many researchers and practitioners. According to Nation (2008), in a well-designed vocabulary development program, the teacher's jobs "in order of importance are planning, strategy training, testing, and teaching vocabulary" (p. 1). Unexpectedly, teaching is listed as the least important aspect because direct vocabulary teaching tends to be inefficient considering that there are simply too many words to deal with (Nation, 2008, p. 5). Thus, a teacher is tasked with strategy training, which is considered essential to inculcate learner independence and autonomy in their vocabulary learning, outside the classroom, where most of the vocabulary learning actually takes place.

Among the vocabulary learning strategies that can be applied explicitly by the teacher in instruction, Nation (2008) suggested using word parts (the other strategies include guessing from context and using a dictionary). A word part consists of prefix, root, and suffix. For example, the word, *transportable*, is composed of *trans* (prefix), *port* (root), and *able* (suffix). Several studies have found that knowledge of word parts has a positive relationship with vocabulary size (Ishii & Schmitt, 2009; Qian, 1999; Schmitt & Meara, 1997), suggesting that both are presumed to grow reciprocally (Mochizuki & Aizawa, 2000). As such, knowledge of word parts is considered to be an "essential part of overall word knowledge" and “play[s] a role in vocabulary acquisition” (Mäntylä & Huhta, 2014, p. 45).

Previous intervention studies on utilizing learners’ word part knowledge for vocabulary development have shown positive results (e.g., Wei, 2015). As learners may not automatically learn derivational knowledge through exposure (Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002), and their L1 (i.e., Latin-based or not) is known to greatly influence the knowledge of English word parts (Bellomo, 2009), explicit attention should be paid to derivative word forms (Nation & Webb, 2011), and learners’ initial derivational knowledge, prior to teaching, should be tested and assessed for diagnostic and intervention purposes (e.g., identify the type of word forms that need focus). Despite such a need for a test of word parts, there has been no standardized instrument, except for tests developed for specific studies (Mäntylä & Huhta, 2014; Schmitt & Zimmerman, 2002). However, recently Sasao (2013) and Sasao and Webb (2017) developed the Word Part Levels Test (WPLT), which measures written receptive knowledge of affixes, and provided initial evidence to substantiate its validity. Because the WPLT was developed to respond to pedagogical concerns, this testing tool can provide teachers and learners with diagnostic information on learners’ strengths and weaknesses in affix knowledge.

Nonetheless, to make the WPLT more useful and accessible as a diagnostic test, we developed a computerized adaptive testing (CAT) version of the WPLT, and evaluated the accuracy of the CAT version of WPLT against Sasao and Webb’s study (2017) that presented the test and data about its use. The following sections provide overviews of the WPLT and computerized adaptive testing (CAT) in order to describe the rationale behind the creation of the CAT version of the WPLT. We then present the development and trialing of the CAT version of the WPLT, followed by a discussion of the results and pedagogical implications.

Word Part Levels Test

Although the amount of vocabulary research has greatly increased over a few decades, new vocabulary tests have rarely been developed, with the Vocabulary Levels Test (Nation, 1983; Schmitt, Schmitt, & Clapham, 2001) and the Word Associates Test (Read, 1993) being the most frequently used and cited tests (Webb & Sasao, 2013). The paucity of tests measuring different aspects of vocabulary knowledge is quite surprising, given that word knowledge involves different degrees of knowing a range of the word's characteristics (Nation, 2013). A testing instrument for assessing aspects of vocabulary knowledge other than size and depth would be of great value to researchers, teachers, and learners because such a test would inform them of what learners have learned and what they are lacking in integral components of vocabulary knowledge (Webb & Sasao, 2013).

Addressing this void, Sasao and Webb (2017) developed the WPLT. Affix knowledge has long been regarded as one of the key components of vocabulary development (Nation, 2013), but no comprehensive measure of affix knowledge existed prior to the WPLT. The WPLT was designed to measure the form, meaning, and use of the different affixes for the purpose of aiding test users in current and future vocabulary learning and teaching. To prioritize users' focus on the most useful affixes, Sasao and Webb (2017) selected 118 derivative affixes for the WPLT, all of which appear in the most frequent 10,000 word families in Nation's (2004) BNC word lists.

The WPLT comprises three sections (i.e., form, meaning, and use) with each section measuring one aspect of receptive affix knowledge. The three sections in the WPLT correspond to the three aspects of receptive knowledge of affixes proposed by the literature (Bauer & Nation, 1993; Nation, 2013; Tyler & Nagy, 1989). The test employs a multiple-choice format so that it would be easy to grade and rewrite poor-performing items based on item analysis.

The first section of the WPLT, the form section, measures knowledge of the written forms of affixes. Two example items for this section are shown below. Example 1 is for the prefix *dis*-, and Example 2 is for the suffix *-ful* (Notice that all the options are written with the same number of letters). Prefixes and suffixes are presented in separate items. Test-takers are presented with four options: one is a real affix form while the other three distractors are real strings of letters of English words but are not affixes.

Example 1.	(1) sal-	(2) cau-	(3) lin-	(4) dis-
Example 2.	(1) -rse	(2) -ack	(3) -ful	(4) -uin

The second section of the WPLT, the meaning section, measures knowledge of the relationships between affix forms and their meanings. Examples 3 and 4 (*re-* and *-able*) are given for this section below. As with the form section, prefixes and suffixes are presented in separate items. Test-takers are presented with a target affix with two example words to direct toward one correct answer and prevent underestimation of the test-takers' knowledge (see Sasao & Webb, 2017, for further discussion). Four options are given: one correct and three distractors. Test-takers are required to choose the correct option with the meaning of the affix represented in the two example words. The three distractors carry the meanings of other randomly selected affixes. The four options are written within the levels of the most frequent 2,000 word families of the BNC word lists so that lack of vocabulary knowledge should not affect the selection of an answer.

Example 3. re- (replay; rebuild)
(1) person
(2) again
(3) female
(4) before

Example 4. -able (acceptable; predictable)
(1) person
(2) not
(3) can be
(4) one

The third section of the WPLT, the use section, measures knowledge of the grammatical functions (i.e., the part of speech) of affixes. Examples 5 and 6 (*en-* and *-al*) for this section are provided below. The same, or similar, item format has been used in previous studies (e.g., Leontjev, Huhta, & Mäntylä, 2016; Mochizuki & Aizawa, 2000) to measure this aspect of L2 learners' knowledge of affixes. Similar to the preceding two sections, prefixes and suffixes are presented in separate items. Test-takers are given a target affix with two example words. The test-takers must choose the correct part of speech from the four fixed options throughout the use section: Noun, Verb, Adjective, and Adverb.

Example 5. en- (ensure; enable)

(1) Noun

(2) Verb

(3) Adjective

(4) Adverb

Example 6. -al (personal; traditional)

(1) Noun

(2) Verb

(3) Adjective

(4) Adverb

Some may argue that, in the use section, test-takers need to know grammatical terms such as "noun," "verb," "adjective," and "adverb," which are supposedly a qualitatively different type of knowledge from what is measured by the items in the other two sections. On this point, Sasao and Webb (2017) reported that the reliability estimate was high and the correlations were moderately high with the other two sections: These results suggest that the items in the use section contributed reliably to the overall score despite the fact that certain metalinguistic knowledge may have contributed to performance on this section but not on the others.

Descriptions of the WPLT development (Sasao & Webb, 2017) suggest that great care was taken to avoid a potential confounding effect in measuring knowledge of derivative affixes. For example, no context is provided in the WPLT, as it is not the intention of the test to measure sentence comprehension. In the meaning and the use sections, two example words were provided for each item to help the test-takers demonstrate their affix knowledge; otherwise, they have to recall a word containing the target affix themselves, which is beyond what a receptive test such as the WPLT intends to measure.

Sasao and Webb (2017) made an initial attempt to validate the WPLT. First, poor-performing items were revised through Rasch analysis based on the data from 417 Japanese university students. Next, the item difficulties were estimated from the data of 1,348 participants from over 100 countries with varied L1 backgrounds to eliminate the effect of a particular L1 knowledge. By including participants with different L1 backgrounds, Sasao and Webb (2017) were able to argued, "any advantages or disadvantages from cognates and loan words for one native language over another were less likely to influence results" (p. 22). Reliability coefficients for all three sections in the WPLT were high. Rasch item difficulty was estimated for each section. Based on the results, the 118 affixes were classified into three difficulty levels (i.e., beginner, intermediate, and advanced). The resulting number of affixes and items in the three levels are shown in Table 1.

Bauer and Nation (1993) proposed the teaching–learning order for seven levels of affixes, based on the theoretical argument that "once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort" (p. 253). Given the fact that the number of affixes in each level differs greatly in Bauer and Nation's classification, and that their order was more or less similar to the difficulty estimates in the WPLT, Sasao and Webb (2017) chose to have approximately the same number of affixes (39 or 40 affixes) in the WPLT. Therefore, the learning burden for each level in the WPLT is intended to be balanced, which obviously gives Sasao and Webb's (2017) affix classification practical advantages in providing diagnostic feedback.

Table 1

Number of Affixes and Items in the WPLT (Sasao & Webb, 2017)

Level	No. of affixes	No. of items in each section			Total items in the form
		Form	Meaning	Use	
Beginner	40	40	34	13	87
Intermediate	39	37	21	21	79
Advanced	39	38	18	22	78

For the WPLT scoring criterion, the scores are calculated for each section, rather than for the test as a whole, so feedback can be provided to the test-takers on each aspect of word part knowledge. Based on the diagnostic results, students, assisted by teachers, can then work on improving their knowledge of word parts, particularly in the sections in which they have performed poorly. Doing so is intended to help students learn, and teachers can teach unknown words that contain the targeted word parts in future instruction. Specifically, the list of all 118 affixes included in the WPLT (http://ysasaojp.info/VocabTests/WPLT/Affix_list.pdf) can be given to students, provided the test is not used again for measuring gains in their knowledge of word parts after instruction.

Furthermore, the WPLT is intended to offer the pedagogical value of raising teachers' and learners' awareness of important aspects of affix knowledge, because the WPLT has three sections that measure knowledge of the form, the meaning, and the use of affixes, respectively. By isolating and measuring different aspects of word part knowledge, detailed feedback can be given with respect to an individual's strengths and weaknesses in affix knowledge.

Despite the pedagogical value of the WPLT, it has at least two limitations. First, teachers need to have an estimate of their students' level of affix knowledge in order to choose from the three levels in the WPLT (i.e., beginner, intermediate, and advanced). More often than not, teachers do not have a clear understanding of students' affix knowledge; thus, the teachers may not always be able to make an appropriate selection of level for their students. The second limitation is the extensive number of test items in the WPLT that students are required to answer: Each level has about 80 items in total (Table 1), and this will take students between 20-30 minutes to complete. For diagnostic purposes, it would be desirable to have fewer items without compromising reliability.

These two limitations may be addressed by creating a computer-adaptive version of the WPLT. First, the computer-adaptive version can promptly diagnose the appropriate WPLT level (i.e., beginner, intermediate, and advanced) for each test-taker shortly after they begin the test. In addition, teachers are able to identify their learners' strengths and weaknesses in affix knowledge without having to guesstimate their levels. Second, in a computer-adaptive test, all test-takers answer different, individualized, and level-appropriate items depending on their ability, which results in a smaller number of items in a test. In addition, the precision of measurement of computer-adaptive tests is theoretically greater than that of the paper-and-pencil counterparts because the items assigned for each test-taker are level appropriate (Wainer et al., 2000). Furthermore, the online format of the test makes it more accessible to a wider audience outside the classroom.

For these reasons, this study was conceived and designed to develop the computer-adaptive WPLT so that a more accurate diagnosis and prompt feedback on test-takers' affix knowledge with a smaller number of items in the test can be generated.

Computerized Adaptive Testing

Computerized adaptive testing (CAT) is similar to computer-based testing (CBT) in their shared similarity in the use of computers, but CAT is differentiated from CBT by its adaptive selection of items to be administered based on the response for each item, whereas CBT normally refers to a fixed set of items administered on a computer. Figure 1 illustrates how the CAT estimates the test taker's ability and its corresponding standard error as more items are administered,. The x-axis shows the number of items administered in the test, and the y-axis is the test taker's estimated ability. The mid-point is the point ability estimate, while the error bars show interval estimation of the ability. The figure shows that correct responses increase the estimated ability, and incorrect responses decrease it and that successive confidence intervals have a decreasing size. In CAT, if the test-taker correctly answers the first item, a more difficult item will be administered next. If the test-taker gets the next item wrong, an easier item will be administered. Accordingly, the selection of the next test item in terms of its difficulty is linked to the result (i.e., correct or incorrect) of the previous item.

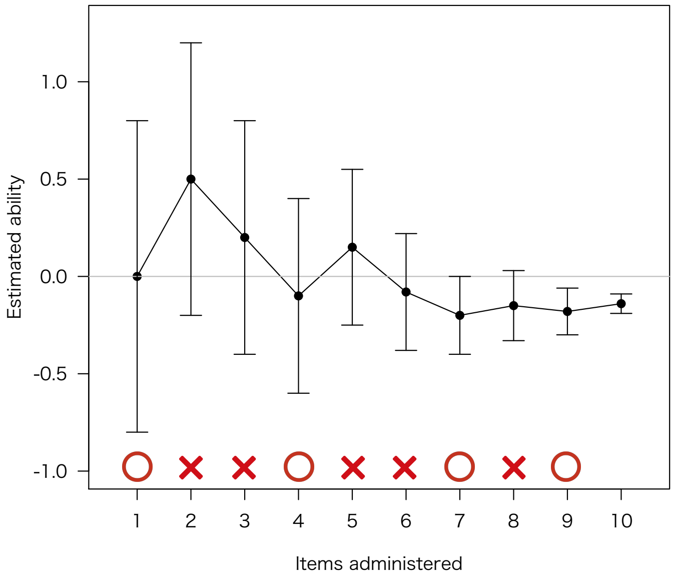


Figure 1. An example of how the CAT estimates the test taker's ability and its corresponding standard error, as more items are administered.

Figure 1 also highlights one of the advantages of CAT in that the range of standard error of ability estimate, which is expressed in logits, namely, a measurement unit used in the item response theory (IRT), progressively becomes smaller as the test-taker answers more items. This is achieved when each subsequent item that is more appropriate for the test-taker's ability is chosen adaptively. This estimation of the ability continues, until the test reaches the pre-determined number of items, or until the standard error becomes lower than the pre-determined threshold value.

The mechanisms behind CAT that create shorter but more accurate tests, as can be expected, attracted the attention of language testers from the 1980's throughout the 1990's (Chalhoub-Deville, 2001). Accordingly, during the 1990's, a number of studies reported the development of CATs (e.g., Brown & Iwashita, 1996; Young, Shermis, Brutten, & Perkins, 1996). The interest in CAT supposedly peaked around the turn of the century when two widely cited books on CAT (Chalhoub-Deville, 1999; Wainer et al., 2000) and their review articles (Fulcher, 2000; Norris, 2001) were published. These publications appeared simultaneously with the partial inclusion of CAT (i.e., the listening and grammar sections) in the Test of English as a Foreign Language (TOEFL) since 1998. With the rapid development of computer technology, the use of CAT gained momentum during the 2000's, and the application of CAT is reported to this day in language testing (e.g., Burston & Neophytou, 2014; Merrell & Tymms, 2007; Papadima-Sophocleous, 2008).

However, the overall prevalence of CAT in the L2 field has not reached the level of penetration, which was predicted in the 1990's (e.g., Dunkel, 1999). This may be due to the fact that it requires "expertise, time, money, and persistence to launch and sustain a CAT development project" (Dunkel, 1997, p. 3). More importantly, CAT is most appropriate for measuring knowledge and skills (Chalhoub-Deville & Deville, 1999), whereas the focus in the L2 field, that is, communicative language testing emphasizes measuring test-takers' performance (Fulcher, 2010).

However, assessment of specific knowledge and skills is useful in diagnostic testing. Alderson (2005), for example, listed the desirable features of diagnostic tests, mentioning that such tests are likely to focus on specific skills rather than communicative performance and that they are also likely to be enhanced by being computerized. In addition, Alderson (2005) pointed out that diagnostic tests are likely to use more discrete point items than performance tasks. In the conversion of WPLT to CAT, we expected that we could maximize on CAT's positive characteristics because WPLT employs multiple-choice, discrete-point items testing linguistic (i.e., affix) knowledge, in contrast to performance-based items intended to test communicative skills, which are not well suited to testing by CAT. Thus, it was assumed that the CAT version of the WPLT would provide further benefits to testing or assessment over the paper-and-pencil version of WPLT in terms of shortening the test length, increasing measurement accuracy, and giving prompt feedback on test-takers' affix knowledge.

By developing and investigating test takers' performance on the CAT version of the WPLT (henceforth CAT-WPLT), this study addressed the following research question:

> How accurate and efficient is the CAT-WPLT, in comparison with the fixed-item version of WPLT?

Method

The project consisted of two stages: development of the CAT-WPLT and trialing. The development stage comprised the typical steps in a test design for a CAT as described below. The trialing was done by engaging students with a typical profile with that of the intended test takers for the prototype test. The results were then compared against those of Sasao and Webb (2017) to address the research question.

Development of the CAT-WPLT

The CAT-WPLT was designed to match its paper-and-pencil counterpart (Sasao & Webb, 2017). First, an item bank was constructed. After adding all the questions in the three levels (i.e., beginner, intermediate, and advanced) in Sasao and Webb's WPLT, the form section had 115 items; the meaning section, 73 items; and the use section, 56 items (See Table 1). We used these items in our creation of an item bank.

Using Sasao and Webb's (2017) data (*N* = 1,348), item calibration (i.e., estimating the parameters related to items such as difficulty) was conducted for each section respectively, using the R package ltm for latent variable modeling and item response theory analyses (Rizopoulos, 2006). The 2-parameter IRT model was chosen for item calibration (i.e., item difficulty and item discrimination) because the items had been administered to a sufficient number of participants by Sasao and Webb (2017) to employ the 2PL IRT model. We chose the 2-parameter IRT model over the 1-parameter IRT model (or Rasch model) employed in Sasao and Webb's (2017) study because the standard error obtained from the result will be theoretically smaller for the 2-parameter IRT model. We checked the model fit with the Standardized Root Mean Square Residual (SRMSR) and confirmed that the SRMSRs for all sections (i.e., form section, .044; meaning section, .046; use section, .048) were smaller than the suggested threshold of .050 (Maydeu-Olivares, 2015), thereby suggesting that the model fits the data quite well. Following this item calibration, the item bank was ready for the CAT program. As we had three different sections, we split the CAT into three different parts by using three different item banks.

It should be acknowledged here that, by employing a multidimensional IRT (MIRT), we could have possibly modeled the underlying latent construct of the WPLT more precisely. As it is assumed that each section of the WPLT measures a somewhat different latent trait; thus, the three traits are correlated under the overarching theme of measuring affix knowledge. The MIRT can deal with this multidimensional nature of language tests (e.g., Min & He, 2014). In this study, however, we applied a 2-parameter IRT model because the CAT program we used (see below) could not implement the MIRT yet.

In CAT the stopping rule (i.e., when the test ends) is mostly based on either the length criterion or the precision criterion. If the length criterion is used in the stopping rule for a CAT program, the test ends when a predetermined number of questions have been administered. If the precision criterion is chosen, items will be given until the standard error (SE) of the ability estimate reaches the pre-specified level of precision; that is, each test-taker has answered a sufficient number of items to construct a score with the desired level of reliability. While the precision criterion could produce a very small SE, it is possible that a large number of items would be needed for reaching such a level of SE. Thus, we chose the length criterion based on the simulation described below as a stopping rule in this study. With the same number of items included in the CAT-WPLT for each test taker, the direct comparison of the reliability of the paper-based WPLT and the CAT-WPLT scores was achieved. In order to determine the number of items administered in each section (i.e., form, meaning, and use) of the CAT-WPLT, we conducted a simulation using the catR, an R package for computerized adaptive testing (Magis & Raîche, 2011, 2012). Figure 2 displays an example of the CAT simulation for the form section. As with Figure 1, the x-axis shows the number of items in the test, and the y-axis is a test taker's estimated ability. The mid-point is the point ability estimate, while the error bars show the interval estimation of the ability. Based on this simulation, we found that, if the true ability (i.e., theta) is 0, then the condition to reach the SE level of 0.3, the form section needs 20 items.

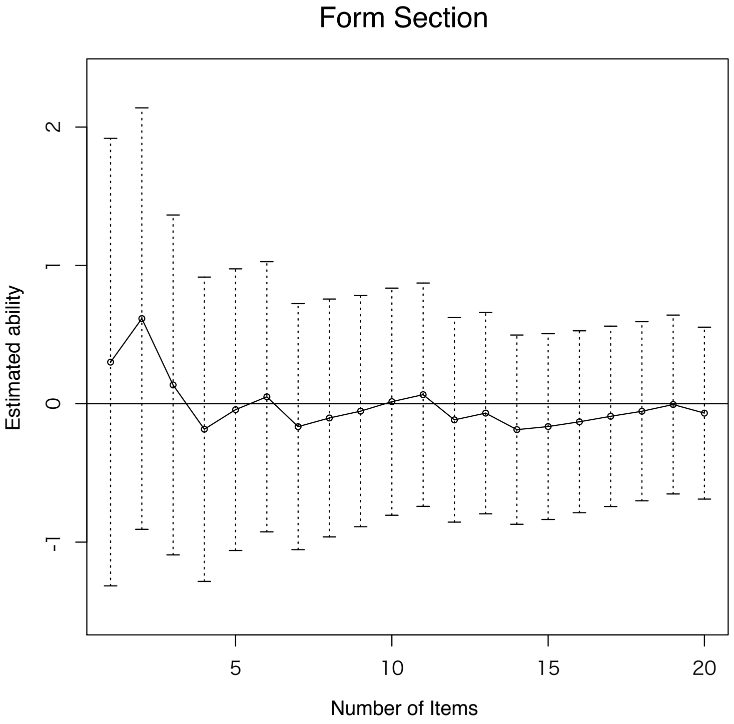


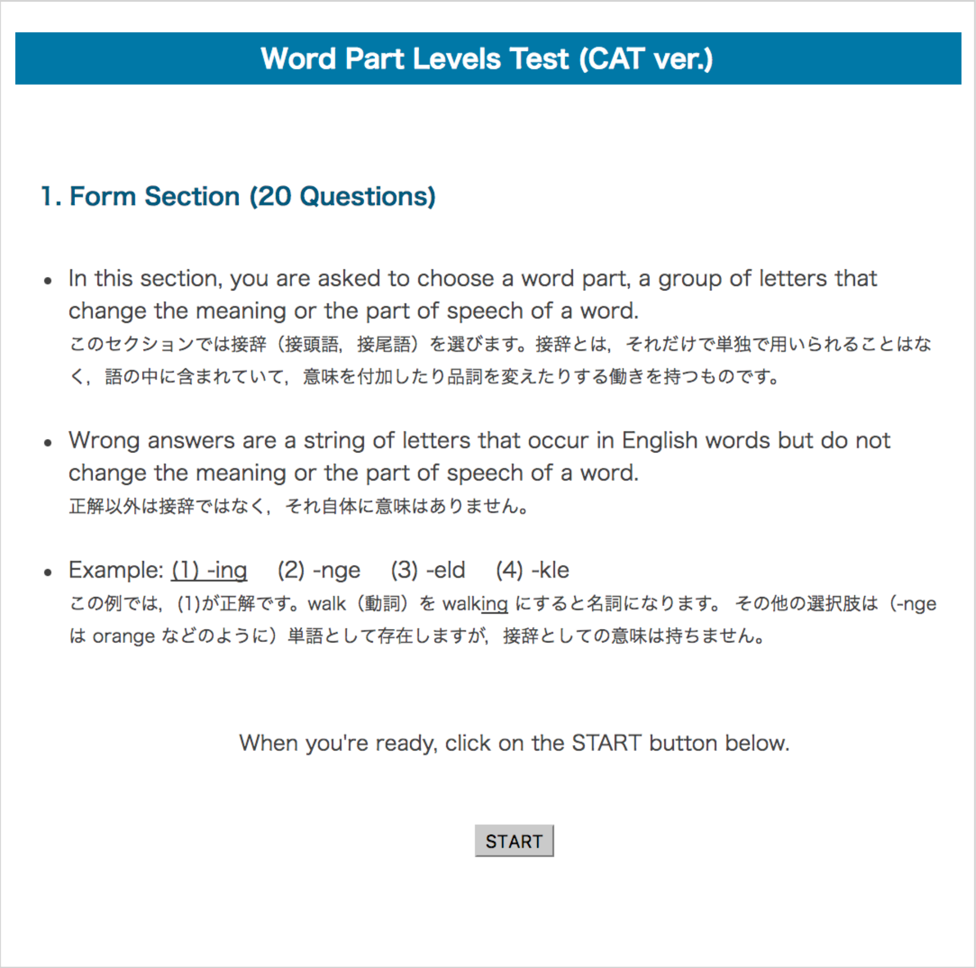
Figure 2. A plot of an example of the CAT simulation for the CAT-WPLT form section. Error bars show 95% confidence intervals.

As a standard error of 0.33 is equivalent to a test reliability of 0.90 in classical test theory (Rudner, 1998), given the ability scores' mean is 0 and the standard deviation is 1 (which is common in IRT), we set the target SE for each section of CAT-WPLT as 0.33 and ran a Monte Carlo simulation with 1,000 repetitions (i.e., SE is estimated each time for 1,000 times, and the mean of all SE values is then calculated). As a result, the number of items in each section in CAT-WPLT was determined as follows: form section, 20 items; meaning section, 15 items; use section, 10 items. Therefore, 45 items were administered to each test-taker, a total that equates to half of the number of items in the paper version of WPLT, as shown in Table 1: beginner level (*k* = 87), the intermediate level (*k* = 79), and the advanced level (*k* = 78).

In addition to the setting of the length criterion as the stopping rule, other steps of the CAT-WPLT were decided. First, the initial item of each section of the test was automatically selected by randomly generating one value from -0.3 to 0.3 to serve as the targeted level of difficulty close to 0 (i.e., the appropriate level for the average ability level test-takers). Selection of the next item was based on maximum Fisher information, whereas for the ability estimation method, including the final ability estimate and its SE, the Bayes modal estimator was used: Both are default settings of the catR package (see Magis & Raîche, 2012 for details).

The integration of the item bank and all the steps described above were performed using a platform called "Concerto" (http://www.psychometrics.cam.ac.uk/newconcerto), which is an online free and open-source adaptive testing platform that provides the flexibility for test developers to combine and use the R language, HTML, and the MySQL database. Since its release in 2011, this testing platform has piqued the interest of people who have longed for a CAT platform as flexible and versatile as Concerto (Scalise & Allen, 2015).

The CAT-WPLT was created using the R code and HTML on Concerto and upon the upload of the item bank. For the present study, all the instructions were also written in Japanese, and examples at the beginning of each section were given, as in the paper version of WPLT, to ensure that test takers in the study would understand the task requirement in the test. Figure 3 illustrates an example of how instructions, examples, and items are presented in the three sections of CAT-WPLT.

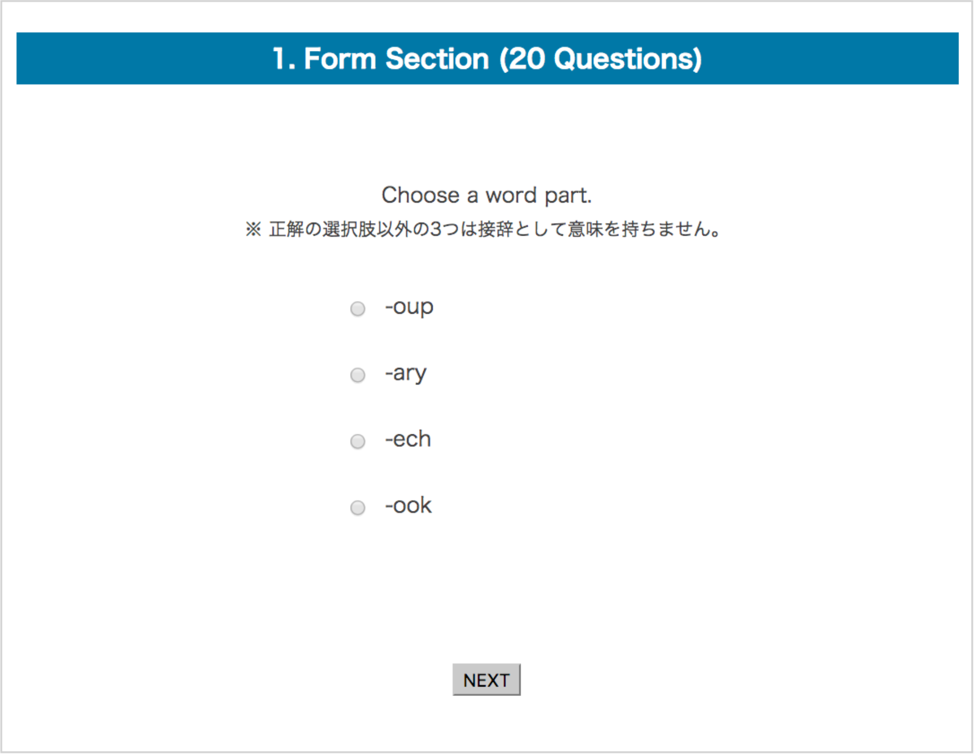


Word Part Levels Test (CAT ver.)

1. Form Section (20 Questions)

- In this section, you are asked to choose a word part, a group of letters that change the meaning or the part of speech of a word.
  このセクションでは接辞（接頭語，接尾語）を選びます。接辞とは，それだけで単独で用いられることはなく，語の中に含まれていて，意味を付加したり品詞を変えたりする働きを持つものです。
- Wrong answers are a string of letters that occur in English words but do not change the meaning or the part of speech of a word.
  正解以外は接辞ではなく，それ自体に意味はありません。
- Example: (1) -ing (2) -nge (3) -eld (4) -kle
  この例では，(1)が正解です。walk（動詞）を walking にすると名詞になります。 その他の選択肢は（-nge は orange などのように）単語として存在しますが，接辞としての意味は持ちません。

When you're ready, click on the START button below.



1. Form Section (20 Questions)

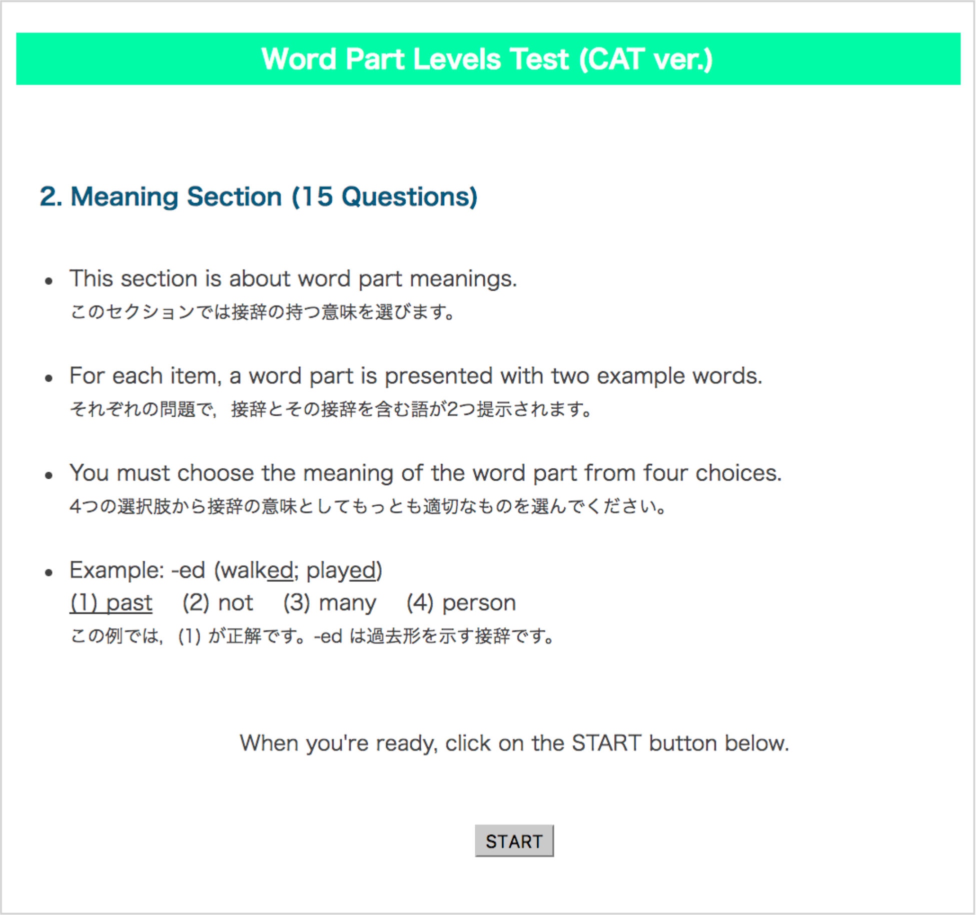
Choose a word part.
※ 正解の選択肢以外の3つは接辞として意味を持ちません。

○ -oup

○ -ary

○ -ech

○ -ook



Word Part Levels Test (CAT ver.)

2. Meaning Section (15 Questions)

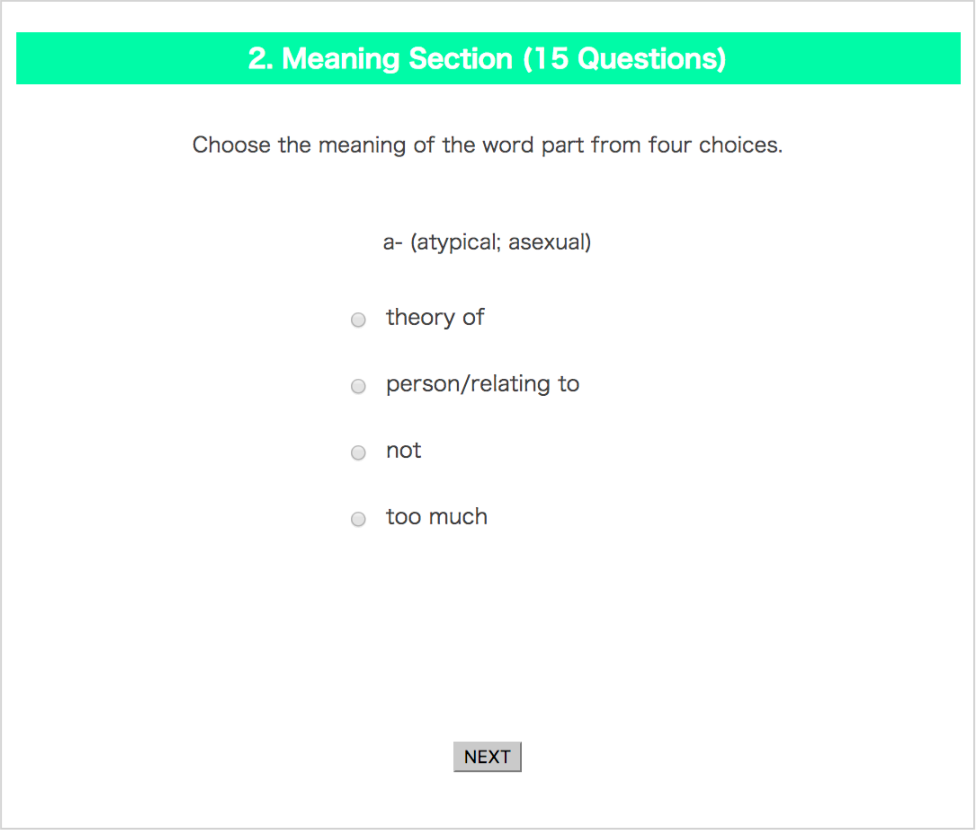
- This section is about word part meanings.
  このセクションでは接辞の持つ意味を選びます。

- For each item, a word part is presented with two example words.
  それぞれの問題で，接辞とその接辞を含む語が2つ提示されます。

- You must choose the meaning of the word part from four choices.
  4つの選択肢から接辞の意味としてもっとも適切なものを選んでください。

- Example: -ed (walked; played)
  (1) past (2) not (3) many (4) person
  この例では，(1) が正解です。-ed は過去形を示す接辞です。

When you're ready, click on the START button below.



2. Meaning Section (15 Questions)

Choose the meaning of the word part from four choices.

a- (atypical; asexual)

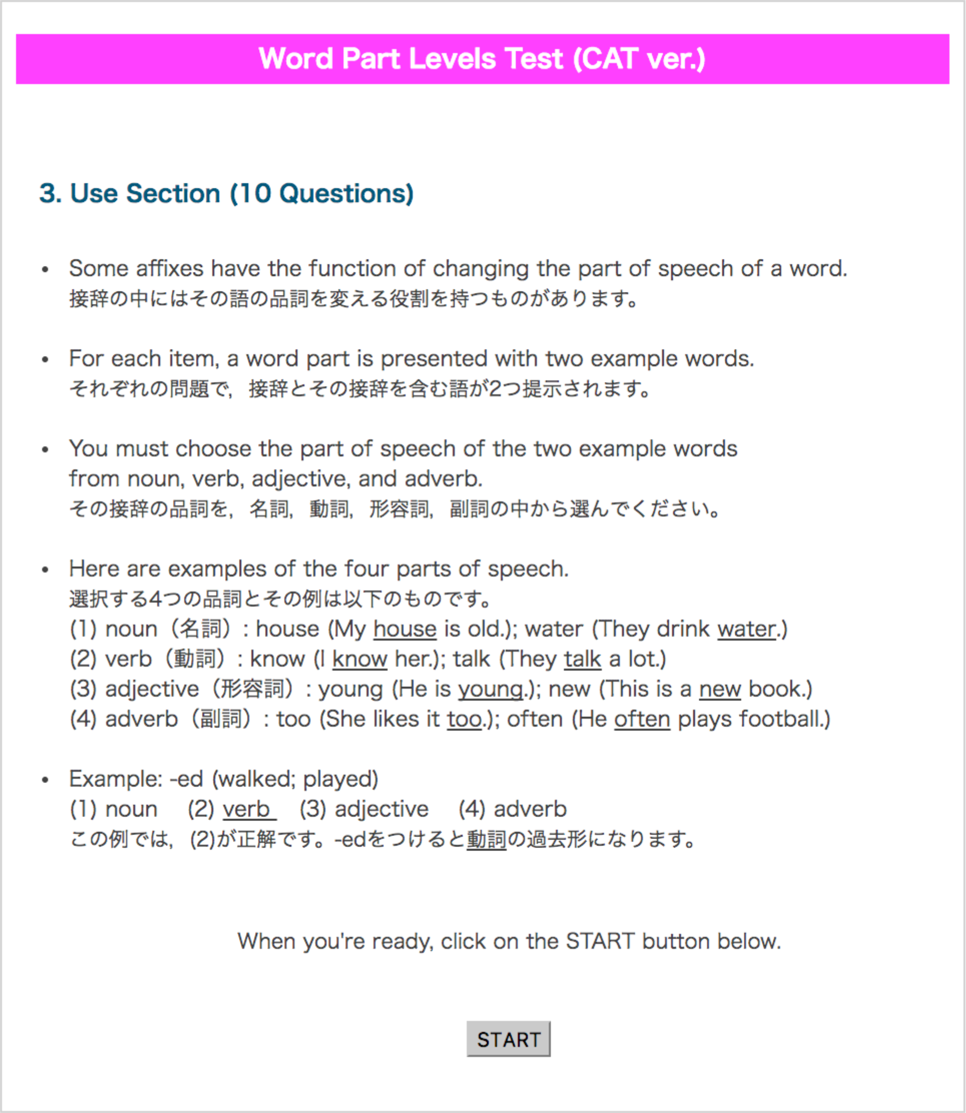
○ theory of

○ person/relating to

○ not

○ too much

E



Word Part Levels Test (CAT ver.)

3. Use Section (10 Questions)

- Some affixes have the function of changing the part of speech of a word.
  接辞の中にはその語の品詞を変える役割を持つものがあります。

- For each item, a word part is presented with two example words.
  それぞれの問題で，接辞とその接辞を含む語が2つ提示されます。

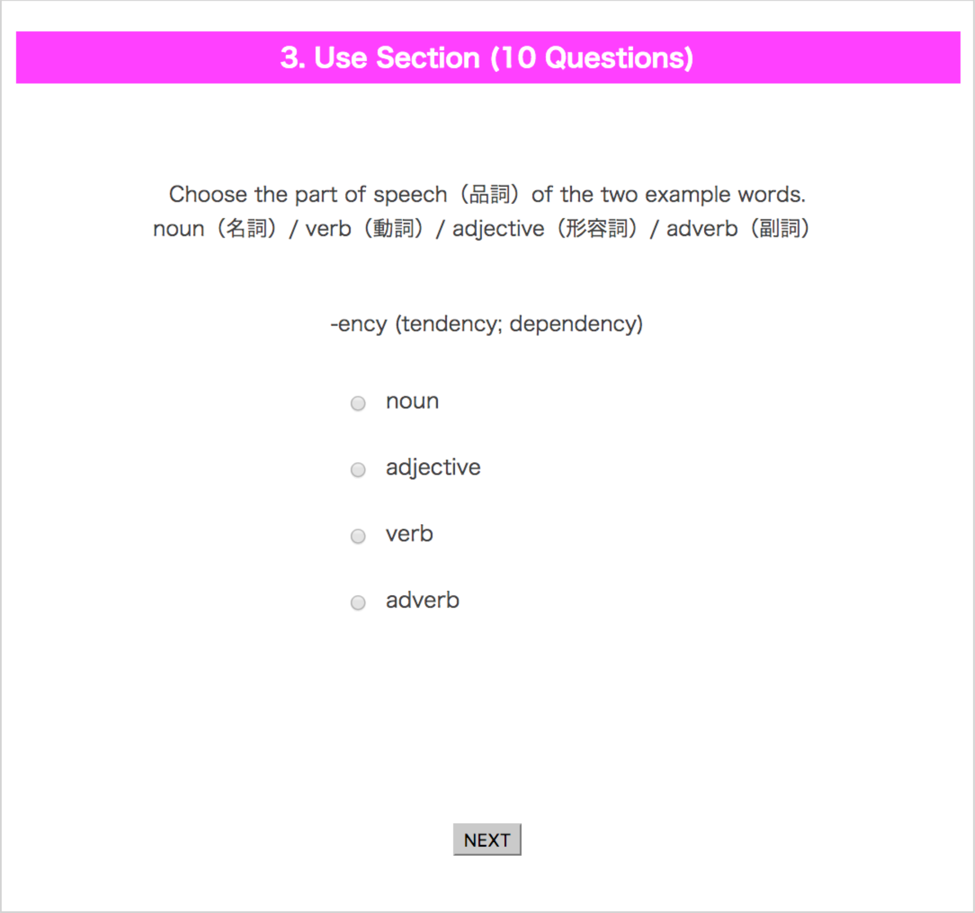
- You must choose the part of speech of the two example words from noun, verb, adjective, and adverb.
  その接辞の品詞を，名詞，動詞，形容詞，副詞の中から選んでください。

- Here are examples of the four parts of speech.
  選択する4つの品詞とその例は以下のものです。
  (1) noun（名詞）: house (My house is old.); water (They drink water.)
  (2) verb（動詞）: know (I know her.); talk (They talk a lot.)
  (3) adjective（形容詞）: young (He is young.); new (This is a new book.)
  (4) adverb（副詞）: too (She likes it too.); often (He often plays football.)

- Example: -ed (walked; played)
  (1) noun (2) verb (3) adjective (4) adverb
  この例では，(2)が正解です。-edをつけると動詞の過去形になります。

When you're ready, click on the START button below.

F



3. Use Section (10 Questions)

Choose the part of speech（品詞）of the two example words.
noun（名詞）/ verb（動詞）/ adjective（形容詞）/ adverb（副詞）

-ency (tendency; dependency)

○ noun

○ adjective

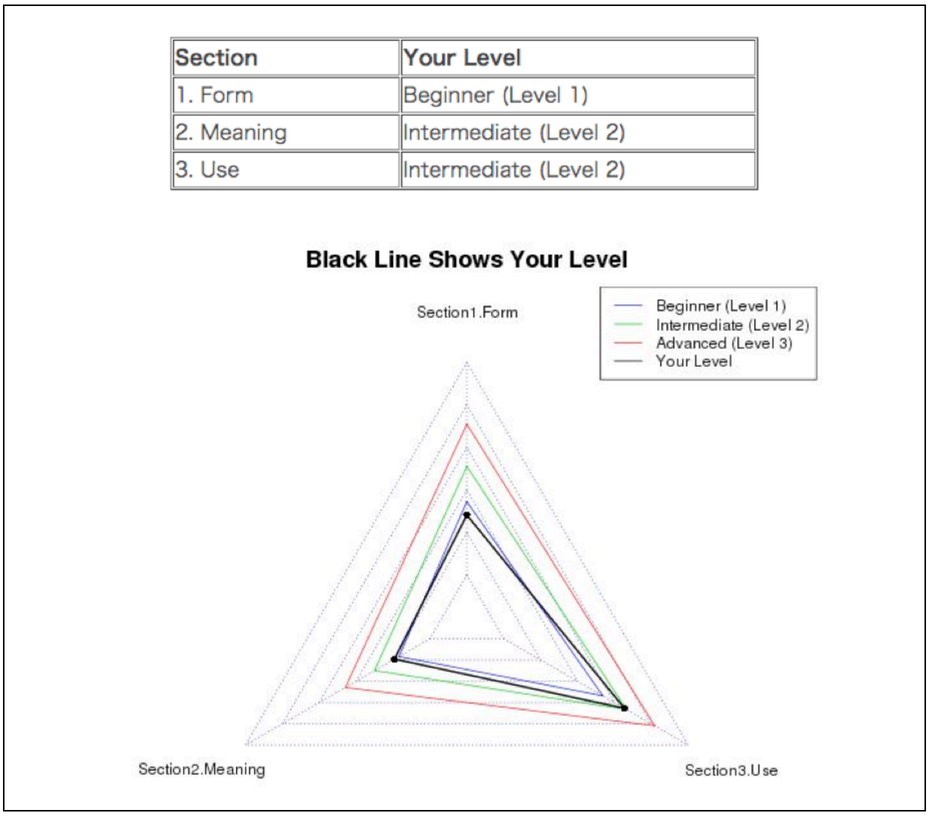
○ verb

○ adverb

Figure 3. Examples of instructions, examples, and items in the three sections of CAT-WPLT. Panels A, C, and E illustrate instructions and examples for the following respective sections: form, meaning, and use. Panels B, D, and F illustrate an example item for these three sections.

After designing each section of the CAT-WPLT, we planned and created the feedback page so that it can provide clear and accurate results of the estimated levels of a test-taker's word part knowledge for the three sections (Figure 4). The thresholds for each level in the three sections were decided based on Sasao and Webb's (2017) original data ($N =$ 1,348), which were used for the item bank. The thresholds for each level in the three sections correspond to the mean estimates of the three difficulty levels. That is, the proportion of correct answers in the level is 50%. As such, test-takers can still benefit from further study of the items in that level, and they are advised to begin their learning from the diagnosed level.

In addition to providing the diagnostic information in words (see table in Figure 4 below), we incorporated a radar chart to show the test-taker's word part knowledge levels visually (Figure 4 below). In Figure 4, the feedback page offers information on test-takers' level of knowledge in the three sections, in this example, knowledge of affix form is at the beginner level, knowledge of affix meaning is at the intermediate level, and knowledge of affix use is at the intermediate level. In addition, we also added an URL link to a file in the feedback page, which lists all the affixes used in the test, so that test-takers could review all the word parts included in the test. The CAT-WPLT is freely available at [URL; omitted for blind review].



Section	Your Level
1. Form	Beginner (Level 1)
2. Meaning	Intermediate (Level 2)
3. Use	Intermediate (Level 2)

Figure 4. The feedback page of CAT-WPLT.

Participants and Trialing

The participants in the trial of the prototype test were 760 university EFL (English as a foreign language) learners from universities in western Japan with the following demographics: first year students, humanities majors, 418 males and 342 females, and aged 18–20. The study was conducted as part of a compulsory English course. The participants provided their written consent to participate after they had been informed explicitly that their grades would not be affected by their test results because the data and findings were specifically for diagnostic and research purposes. Among the participants, 319 students had previously taken the TOEIC (*M* = 487.19, *SD* = 103.43). According to the Educational Testing Service (2013), learners with this range of proficiency are classified as "Basic User"

(A2) to "Independent User" (B1) in the Common European Framework of Reference for Languages (CEFR). Thus, the participants' English proficiency levels ranged from pre-intermediate to intermediate. Mochizuki and Aizawa (2000) reported that Japanese EFL learners need to have a certain vocabulary size to make use of their affix knowledge. Therefore, considering the correlation between learners' vocabulary size and their language proficiency (Beglar, 2010), we assumed that the participants of the present study would be the appropriate target sample because of the adequacy of their proficiency level relative to the difficulty levels of the test items.

Data Analyses

All analyses in this study were conducted using R version 3.2.3 (R Core Team, 2016). The research question concerned the comparison of precision of measurement in ability estimates between the CAT-WPLT and the fixed-item format of the original WPLT (Sasao & Webb, 2017). We, therefore, compared the standard errors of ability estimates of the test-takers in two formats to investigate if there were significant differences in the standard errors. As an index of measurement precision in IRT, the standard error of the ability estimate can be computed for each test-taker; thus, it is possible to state that the test-taker's ability is the estimated ability plus/minus the standard error, thus in sharp contrast with the classical test theory, which can only suggest the test reliability as a whole (Wainer et al., 2000). It was not possible, however, to directly compare the standard errors of the ability estimates in the two test formats because the standard error of ability estimates in item response theory (IRT) is calculated using the test information function, which is the sum of each item information function (Baker, 2001). Therefore, the fixed-item WPLT had a numerical advantage as it had 115, 73, and 56 items in the form, meaning, and use sections, respectively, whereas the CAT-WPLT had only 20, 15, and 10 items, correspondingly but in an individualized way.

Comparing the means between the standard errors was theoretically and mathematically not appropriate, given the standard errors are inversely related to the test information function; that is, the normal distribution cannot be expected for the standard errors unless the test has many items. For this reason, we compared the proportion of test-takers whose standard error of ability estimate was lower than 0.33 and those higher than 0.33, a threshold corresponding to a test reliability of 0.90 in the classical test theory (Rudner, 1998). It should be noted, with large sample sizes such as in the present study, it is easy to reject the null hypothesis (i.e., obtaining $p < .05$) even when there is no practical difference in the conventional null hypothesis statistical testing (NHST) (Plonsky & Oswald, 2014). Another problem with using NHST is that we cannot claim that “the two values are NOT different” even when the null hypothesis (that two values are the same) is accepted (i.e., $p > .05$) because of the logic underlying NHST (Larson-Hall, 2016, p. 319). In the current study, therefore, we chose the Bayesian estimation (Kruschke, 2013) over the traditional chi-square test to compare the proportion of test-takers under and over the threshold (i.e., 0.33) in standard errors. For the Bayesian estimation of the test of proportions, we used the R package, Bayesian First Aid (Bååth, 2014).

To ensure reproducibility and transparency in the data analysis (Larson-Hall & Plonsky, 2015; Marsden, Mackey, & Plonsky, 2016), all data and R codes used in this study are shared online (https://www.iris-database.org/iris/app/home/detail?id=york:932680).

Results

Table 2 shows the descriptive statistics and Pearson correlation coefficients between the three sections within each test format, the CAT-WPLT and the fixed-item WPLT. The standard deviation was larger in the fixed-item WPLT because the test-takers of this format were from more than 100 countries with different L1 backgrounds (Sasao & Webb, 2017). The high variability in their proficiency among test-takers also explains why the correlation coefficients were larger for the fixed-item WPLT than that for CAT-WPLT.

Table 2

Descriptive Statistics and Correlation Coefficients between the Three Sections within Each Test Format

Form	Section	*K*	Mean	*SD*	Correlation coefficient		
					Form	Meaning	Use
CAT (*N* = 760)	Form	20	-0.381	0.716	–	.401–.514	.273–.399
	Meaning	15	-0.775	0.579	.459	–	.473 –.576
	Use	10	-0.154	0.650	.337	.527	–
Fixed-item (*N* = 1,348)	Form	115	0.068	0.946	–	.737–.782	.622–.683
	Meaning	73	-0.116	0.964	.760	–	.656–.712
	Use	56	0.015	0.955	.654	.685	–

Note. k shows the number of items. Means and *SD*s are expressed in logits (log odds units). A logit is a unit used in IRT, and the higher the value means the more difficult (or able) the item (or the person) is. In correlation coefficients, the above diagonal shows 95% confidence intervals of correlation coefficients (from the lower limit to the upper limit).

The lower correlation coefficients for the CAT-WPLT than those for the fixed-item WPLT also indicate that, rather than reporting a total single score for the whole test, the scores for each section should be reported for practical use, if each sub-score is reliable, as suggested by Sasao and Webb (2017). In other words, the lower than expected correlation coefficients of the CAT-WPLT imply that there are learners who score high in one section but not in other sections (and vice versa). Specifically, the use section had a lower correlation with the form section, thus suggesting a possibility that this specific group of test-takers (i.e., Japanese university EFL learners) may lack the explicit metalinguistic knowledge of the parts of speech, which is necessary to answer the items in the use section correctly. Therefore, in terms of learners' word part knowledge, the section scores of the three aspects of receptive knowledge of affixes should be reported and interpreted separately for diagnostic purposes.

Next, in order to answer the research question of the current study, we compared the standard errors of ability estimates of the test-takers in two formats to examine the precision of measurement. Figure 5 displays the box plots of the standard errors of the CAT-WPLT and the fixed-item WPLT. The fixed-item WPLT had many data points in the lower area of the standard error due to its larger number of test items. Yet, the CAT-WPLT functioned well, with almost its 75th percentile (the upper end of the box) of the standard errors in the three sections under the threshold of 0.33. This result is noteworthy because, in contrast to the fixed-item format with a larger number of items, the CAT program succeeded in attaining this level of accuracy by administering items with appropriate difficulty, individually chosen for the test-takers with a much smaller number of items. This serves test-users with one the primary benefits of utilizing the CAT for diagnostic tests.

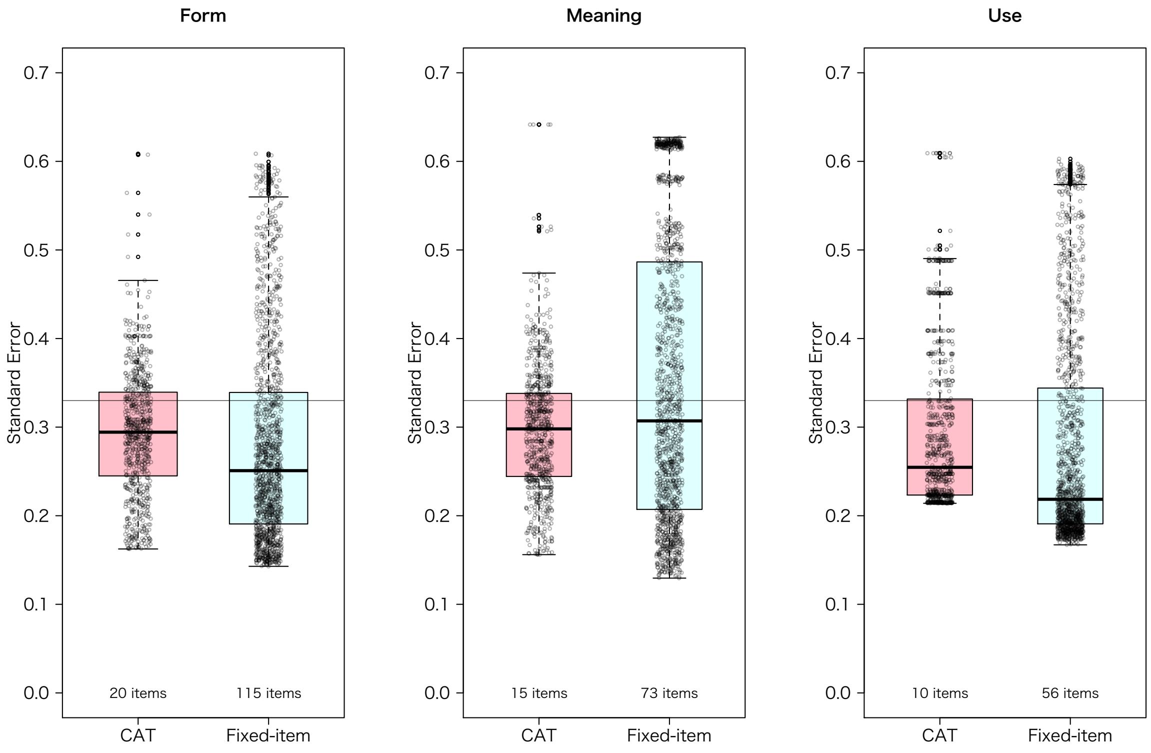


Figure 5. Box plots of the standard errors of the CAT-WPLT and the fixed-item WPLT. The overlaid dots show individual data points of the standard error. The horizontal line indicates the threshold (= 0.33) of the standard error.

We further investigated the same data (i.e., standard errors) from a different perspective. As stated earlier, the standard error of the ability estimate varies across individual test takers, depending on each test-taker's ability estimate. Figure 6 displays the correspondence between the ability estimates and standard errors in the two test formats. The figure shows the standard error (y-axis) of the test-takers with a certain ability estimate (x-axis). Although the fixed-item WPLT marked lower standard errors than CAT-WPLT for all sections in the test (which is only natural given its large number of items), the CAT-WPLT performed comparatively accurate with many of the standard errors under the threshold of 0.33.

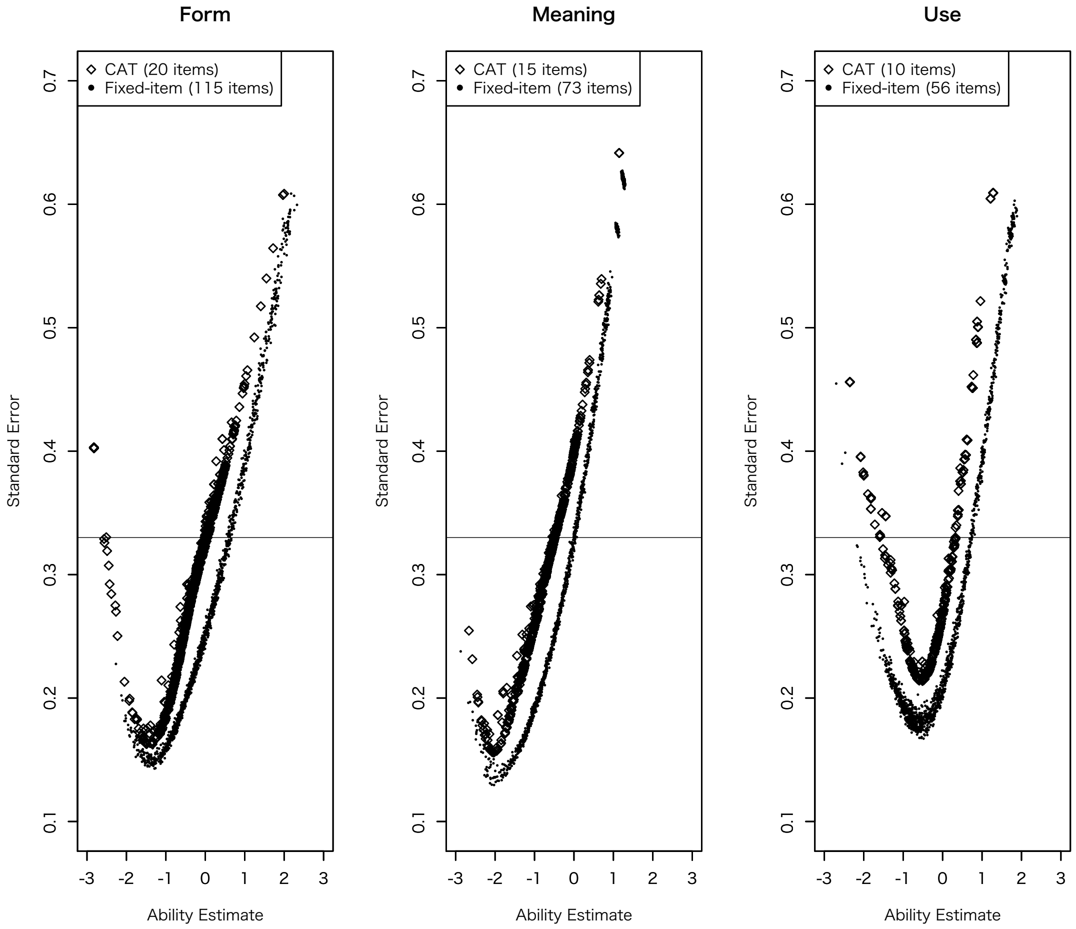


Figure 6. The correspondence between the ability estimate and standard error in the two test formats. The horizontal line indicates the threshold (= 0.33) of the standard error.

Table 3 provides the number and proportion of test-takers that is above and below the threshold (i.e., 0.33) in standard errors. It also shows the Bayesian estimation of the relative frequency and its 95% credible intervals. It is clear from the result that the estimated differences in terms of the number of standard errors under the threshold were almost the same for the form (estimated difference, 3%) and use sections (estimated difference, 1%) for both the CAT-WPLT and the fixed-item WPLT. However, for the meaning section, the estimated difference was large (estimated difference 15%) in favor of the CAT-WPLT. Using the posterior distribution, we calculated the probability that the difference between the numbers of standard errors under the threshold in the two test formats was practically equivalent. In the findings, the probability of the difference being less than five percentage points (i.e., estimated difference is ±0.05), which is small enough to be negligible, was 85.5% for the form section, 0% (smaller than .001) for the meaning section, and 98.1% for the use section. These results indicate that the precision of measurement was approximately the same for the form and use sections in the CAT-WPLT and the fixed-item WPLT, and that the CAT-WPLT performed better in the meaning section. It can thus be concluded that, in terms of the measurement precision, the CAT-WPLT with smaller number of items functioned as well as the fixed-item WPLT despite the larger number of items of the latter (Table 4).

Table 3

Number and Proportion of Test-takers Above and Below the Threshold in Standard Errors

Section	Format	Under 0.33	Over 0.33	Estimated relative frequency [95% credible intervals]	Estimated difference (CAT – Fixed item) [95% credible intervals]
Form	CAT	537 (70.66%)	223 (29.34%)	0.71 [0.67, 0.74]	-0.03 [-0.07, 0.01]
	Fixed-item	990 (73.44%)	358 (26.56%)	0.73 [0.71, 0.76]	
Meaning	CAT	529 (69.61%)	231 (30.39%)	0.70 [0.66, 0.73]	0.15 [0.11, 0.19]
	Fixed-item	734 (54.45%)	614 (45.55%)	0.54 [0.52, 0.57]	
Use	CAT	565 (74.34%)	195 (25.66%)	0.74 [0.71, 0.78]	0.01 [-0.03, 0.05]
	Fixed-item	991 (73.52%)	357 (26.48%)	0.73 [0.71, 0.76]	

Note. The number of test-takers: CAT (*N* = 760) and Fixed-item (*N* = 1,348). The 95% credible intervals cover 95% of the posterior probability distribution.

Table 4

Comparison of Numbers of Items in Paper-and-Pencil WPLT and CAT-WPLT

Format	Form	Section			Total
		Form	Meaning	Use	
Paper-and-pencil WPLT	Beginner	40	34	13	87
	Intermediate	37	21	21	79
	Advanced	38	18	22	78
CAT-WPLT	—	20	15	10	45

Discussion

This study addressed the possibility of creating a computerized adaptive version of the Word Part Levels Test developed by Sasao (2013) and Sasao and Webb (2017). Using an online adaptive testing platform, Concerto, the CAT-WPLT was developed, and its performance was evaluated after administering the CAT-WPLT to 760 Japanese university EFL learners. The research question of this study was, “How accurate and efficient is the CAT-WPLT in comparison with the fixed-item version of WPLT?” The findings suggest that the CAT-WPLT measured test-takers’ word part knowledge with many fewer items, administered in approximately in 10 minutes, than the paper-and-pencil version of WPLT and with similar or greater precision than the fixed-item counterpart. It was therefore able to provide a prompt diagnosis to the test users.

The results of the present study highlight the potential of utilizing CAT for an multiple-choice and discrete-point items as conventionally suggested in the literature on CAT (Alderson, 2000). CAT can indeed measure test-takers’ ability more efficiently than the paper-and-pencil counterpart. Specifically, the present study demonstrated that conversion of tests that have been shown to produce reliable scores, such as WPLT, into CAT through a rigorous development methodology (Sasao, 2013; Sasao & Webb, 2017), can greatly facilitate the process of developing a diagnostic test. Such diagnostic tests may provide valuable information about teaching and learning vocabulary in this context.

The same pedagogical implications suggested by Sasao and Webb (2017) for using

WPLT are also applicable to CAT-WPLT, but the CAT-WPLT has some added benefits as well. First, teachers and learners alike can identify the forms of affixes that they should focus on in their teaching and learning of vocabulary. This is because the interpretation of the results and identification of the proficiency levels of each section (i.e., form, meaning, and use) of CAT-WPLT are easy, succinct, yet comprehensive, as clearly reported in Figure 4, rather than having teachers or learners themselves manually calculate the percentage of correctly answered items for each section as recommended in the paper-and-pencil version of WPLT. This prompt and easy-to-understand diagnostic feedback may not only help learners raise their awareness of affixes but also allow them to undertake in improving their word part knowledge at an appropriate level immediately.

Second, with the CAT-WPLT, teachers can monitor their students' progress in developing the word part knowledge. The results for each learner are accessible in logits in the CAT-WPLT, in contrast to raw scores in WPLT, which require manual recording; hence, the monitoring of learners' development of affix knowledge before and after explicit instruction of word parts in a program can be done with increased ease and efficiency. In another case, CAT-WPLT could be used to measure the gains in affix knowledge before and after some treatment in order to detect factors that facilitate the development of word part knowledge.

One drawback of using the CAT-WPLT is its requirement for an Internet connection. Of course, the CAT-WPLT can be taken outside the classroom as long as the test-taker has an Internet connection. In an “unwired” classroom, however, the paper-and-pencil version of WPLT will be of great value. After all, the WPLT is a very practical diagnostic instrument for classroom teachers, and the CAT-WPLT is an alternative that added value to WPLT with the aid of information and communication technology.

Because the CAT-WPLT has proven to be equally or more reliable and efficient in comparison with the fixed-item version of WPLT, it would be worth investing time in some improvements. First, more items should be created for the item bank in the future enhancement of the CAT-WPLT to deal with content balancing and item exposure. Second, the meaningfulness of the test results might be improved by mapping the scores or levels according to CEFR standards as in the English Vocabulary Profile (http://www.englishprofile.org/wordlists) to reveal the correspondence between test-takers' levels of affix knowledge and proficiency. Third, computer-adaptive, individualized instruction, as reported in some studies (e.g., Fehr et al., 2012), could be incorporated by employing the concept of adaptivity in full scale. The development of the CAT-WPLT can thus be regarded as an initial step toward and an integral part of realizing such an ideal adaptivity.

Conclusion

The findings of this study have implications regarding adaptations to other tests of language knowledge that might be transformed into useful measures for learners, teachers, and researchers. This study has successfully demonstrated that by converting an established test into a CAT format, the test length experienced by any individual test-taker can be shortened and measurement accuracy can be increased at the same time. In general, tests for linguistic knowledge, such as vocabulary, tend to have many items (e.g., Ishii & Schmitt, 2009), and thus some practitioners might hesitate to use them for classroom teaching, but by utilizing the CAT system, those tests made in the past could be better used for diagnostic purposes. For promoting such use of CAT, collaboration among researchers, practitioners, and other related entities is necessary. Collaboration is needed to exploit computer adaptivity in instruction and assessment in the era where an open-source CAT platform, such as Concerto, is well within the reach of any individual language tester.

References

Alderson, J. C. (2000). Technology in testing: The present and the future. *System*, *28*, 593–603. doi:10.1016/S0346-251X(00)00040-3

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.

Bååth, R. (2014). Bayesian First Aid: A package that implements Bayesian alternatives to the classical *.test functions in R. *UseR! 2014 - the International R User Conference*. Retrieved from http://sumsar.net/files/academia/baath_user14_abstract.pdf

Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland. Retrieved from http://echo.edres.org:8080/irt/baker/

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*, 253–279. doi:10.1093/ijl/6.4.253

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*, 101–118. doi:10.1177/0265532209340194

Bellomo, T. S. (2009). Morphological analysis as a vocabulary strategy for L1 and L2 college preparatory students. *TESL-EJ*, *13*, 1–27. Retrieved from http://www.tesl-ej.org/wordpress/issues/volume13/ej51/ej51a1/

Brown, A., & Iwashita, N. (1996). Language background and item difficulty: The development of a computer-adaptive test of Japanese. *System*, *24*, 199–206. doi:10.1016/0346-251X(96)00004-8

Burston, J., & Neophytou, M. (2014). Lessons learned in designing and implementing a computer-adaptive test for English. *The EuroCALL Review*, *22*, 19–25. doi:10.4995/eurocall.2014.3632

Chalhoub-Deville, M. (Ed.). (1999). *Issues in computer-adaptive testing of reading proficiency*. Cambridge University Press.

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning & Technology*, *5*, 95–98. Retrieved from http://llt.msu.edu/vol5num2/deville/

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, *19*, 273–299. doi:10.1017/S0267190599190147

Dunkel, P. (1997). Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher*, *21*, 1–8.

Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, *2*, 77–93. Retrieved from http://llt.msu.edu/vol2num2/article4/

Educational Testing Service. (2013). Mapping the TOEIC and TOEIC Bridge tests on the Common European Framework of Reference for Languages. Retrieved from http://www.ets.org/s/toeic/pdf/toeic_cef_mapping_flyer.pdf

Fehr, C. N., Davison, M. L., Graves, M. F., Sales, G. C., Seipel, B., & Sekhran-Sharma, S. (2012). The effects of individualized, online vocabulary instruction on picture vocabulary scores: An efficacy study. *Computer Assisted Language Learning*, *25*, 87–102. doi:10.1080/09588221.2011.586640

Frey, A., & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*, 89–94. doi:10.1016/j.stueduc.2009.10.007

Fulcher, G. (2000). Book review: Issues in computer-adaptive testing of reading proficiency. *Language Testing*, *17*, 361–367. doi:10.1177/026553220001700305

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

Ishii, T., & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal*, *40*, 5–22. doi:10.1177/0033688208101452

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573–603. doi:10.1037/a0029146

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS* *and R* (2nd ed.). New York, NY: Routledge.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*, 127–159. doi:10.1111/lang.12115

Leontjev, D., Huhta, A., & Mäntylä, K. (2016). Word derivational knowledge and writing proficiency: How do they link? *System*, *59*, 73–89. doi:10.1016/j.system.2016.03.013

Magis, D., & Raîche, G. (2011). catR: An R package for computerized adaptive testing. *Applied Psychological Measurement*, *35*, 576–577.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R Package catR. *Journal of Statistical Software*, *48*. doi:10.18637/jss.v048.i08

Mäntylä, K., & Huhta, A. (2014). Knowledge of word parts. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 45–60). Basingstoke, England: Palgrave Macmillan.

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1–21). New York, NY: Routledge.

Maydeu-Olivares, A. (2015). Evaluating fit in IRT models. In Steve P. Reise & Dennis A. Revicki (Eds.). *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111–127). New York, NY: Routledge.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching*, *13*, 221. doi:10.1017/S0261444800008879

Merrell, C., & Tymms, P. (2007). Identifying reading problems with computer adaptive assessment. *Journal of Computer Assisted Learning*, *23*, 27–35.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, *31*, 453–477. doi:10.1177/0265532214527277

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: an exploratory study. *System*, *28*, 291–304. doi:10.1016/S0346-251X(00)00013-0

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, 12–25.

Nation, 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, the Netherlands: John Benjamins.

Nation, P. (2008). *Teaching vocabulary: Strategies and techniques*. New York, NY: Thomson Heinle.

Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage Learning.

Norris, J. M. (2001). Review of computerized adaptive testing: A primer (second edition). *Language Learning & Technology*, *5*, 23–27. Retrieved from http://llt.msu.edu/vol5num2/review2/default.html

Papadima-Sophocleous, S. (2008). A hybrid of a CBT- and a CAT-based New English Placement Test Online (NEPTON). *CALICO Journal*, *25*, 276–304. doi:10.1558/cj.v25i2.276-304

Plonsky, L., & Oswald, F. L. (2014). How big Is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. doi:10.1111/lang.12079

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*, 282–308. doi:10.3138/cmlr.56.2.282

R Core Team. (2016). R: A language and environment for statistical computing (Version 3.3.2) [Computer software]. Vienna, Austria. Retrieved from http://www.r-project.org/

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355–371. doi:10.1177/026553229301000308

Rizopoulos, D. (2006). ltm : An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*. doi:10.18637/jss.v017.i05

Rudner, L. M. (1998). An on-line, interactive, computer adaptive testing tutorial. Retrieved from http://edres.org/scripts/cat

Sasao, Y., & Webb, S. (2017). The Word Part Levels Test. *Language Teaching Research*, *21*, 12–30. doi:10.1177/1362168815586083

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British*

Journal of Mathematical and Statistical Psychology, *68*, 478–496. doi:10.1111/bmsp.12057

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17–36. doi:10.1017/S0272263197001022

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the bahviour of two versions of the vocabulary levels test. *Language Testing*, *18*, 55–88. doi:10.1177/026553220101800103

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*, 145–171. doi:10.2307/3588328

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. R., Mislevy, R. J., … Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Webb, S., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, *44*, 263–277. doi:10.1177/0033688213500582

Wei, Z. (2015). Does teaching mnemonics for vocabulary learning make a difference? Putting the keyword method and the word part technique to the test. *Language Teaching Research*, *19*, 43–69. doi:10.1177/1362168814541734

Young, R., Shermis, M. D., Brutten, S. R., & Perkins, K. (1996). From conventional to computer-adaptive testing of ESL reading comprehension. *System*, *24*, 23–40. doi:10.1016/0346-251X(95)00051-K