Computers & Education 97 (2016) 69–85

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Measuring English vocabulary size via computerized adaptive testing

Wen-Ta Tseng

English Department, National Taiwan Normal University, Taiwan

ARTICLE INFO

Article history:
Received 14 April 2015
Received in revised form 24 February 2016
Accepted 27 February 2016
Available online 3 March 2016

Keywords:
Computerized adaptive testing
Dynamic testing
English vocabulary size
Diagnostic testing

ABSTRACT

Measuring English vocabulary size in EFL contexts normally requires a large number of test items and relies on paper-and-pencil (P&P) formats. The aim of this study was to examine the feasibility and practicality of computerized adaptive testing (CAT) as an alternative to measuring English vocabulary size. Differing from the fixed, uniform item sequences in conventional P&P tests, CAT adopts a dynamic, adaptive item selection procedure to optimally target the interim ability estimate and reach the convergence, resulting in a shorter, putatively more efficient test-taking process. The study involved three phases. The first phase built up a vocabulary item bank using the Rasch model, which was used for administering the CAT study; the second phase undertook an experiment to compare various termination conditions in both the P&P and CAT contexts; the third phase examined the accuracy and efficiency of the two test modes in classifying test-takers into mastery and non-mastery groups. The results show that testing EFL learners' English vocabulary size with CAT requires only one third of the items in the item bank while still producing comparable vocabulary size estimates to the original test calibrated by all the 180 items in the item bank. The study also demonstrates that CAT can be more efficient and precise in classifying test-takers into mastery and non-mastery groups. These research findings suggest that CAT has great potential in efficiently and precisely measuring EFL learners' English vocabulary size. The relevant research and pedagogical implications are further discussed.

1. Introduction

The integration of computer technology into education has profoundly transformed teaching and learning in the 21st century. The ever increasing number of technology-enabled school environments offers a tremendous potential for both empowering learners and increasing their learning motivation. Modern technologies also provide new channels through which teachers can efficiently administer classroom tests to better monitor and more validly evaluate students' academic performance (Gusev & Armenski, 2014; Mayrath, Clarke-Midura, & Robinson, 2012; Wu, Kuo, Jen, & Hsu, 2015).

One way of taking advantage of technology for assessing learners' academic performance is computerized adaptive testing (CAT), which symbolizes the integration of computerized testing and adaptive testing (Chang, 2015). The beginnings of computerized testing in the 1970s were hindered because of under-developed computer technology, and it was not until the late 1990s that computerized testing fully matured with the invention of high-speed CPUs. The concept of adaptive administration of items was invented and used by the founder of the intelligence test, Alfred Binet, whose research dates back to the early 20th century. Yet adaptive testing lay dormant until the early 1950s due to technological constraints and an emphasis on classroom testing with time limits (Weiss, 1983). The goal of adaptive testing is to maximize information

http://dx.doi.org/10.1016/j.compedu.2016.02.018
0360-1315/

concerning test-takers by choosing informative items from a large item bank which should ideally be calibrated by item response theory (van der Linden & Pashley, 2010). This process of adaptive testing yields better reliability and achieves a higher efficiency of measurement (Thissen, 2000); and because of its promising psychometric features, there have been a variety of stand-alone CAT system developments over the past decade (Klinkenberg, Straatemeier, & van der Maas, 2011; Lilley, Barker, & Britton, 2004; Nirmalakhandan, 2007; Verschoor & Straetmans, 2010).

One field that has made extensive use of CAT is English language testing. The application of CAT for measuring English language proficiency has grown over the past two decades, being adopted in high stakes international English proficiency exams such as the TOEFL, GRE, and GMAT (Rudner, 2010). Systematic empirical studies based on CAT-applications have been undertaken for a range of language skills, including listening (Dunkel, 1999; Madsen, 1991), reading (Chalhoub-Deville, 1999; Kaya-Carton, Carton, & Dandonoli, 1991), and vocabulary (Laufer & Goldstein, 2004; Vispoel, 1993, 1998; Vispoel, Rocklin, & Wang, 1994) but less frequently in writing (Stevenson & Gross, 1991) and speaking (Malabonga, 2000; Malabonga & Kenyon, 1999).

Pioneering studies of CAT vocabulary tests (e.g., Vispoel, 1993, 1998; Vispoel et al., 1994) established that CAT could reach levels of reliability and validity equal to or higher than paper-and-pencil (P&P) tests using considerably fewer test items. Although these early comparability inquiries showed the advantage of CAT over P&P in reducing the overall number of test items, the P&P tests used in the early studies were not specifically designed as English vocabulary size instruments embedded in an English-as-a-foreign-language (EFL), or a second language (L2), context. The crucial difference between these two types of tests is that an L2 vocabulary size test must evaluate a greater range of theorized levels in both academic and non-academic domains. It is therefore still not clear whether CAT can be utilized to accurately measure English vocabulary *size*, the information of which is normally acquired through a P&P test procedure. In particular, it remains to be demonstrated whether CAT can be used to determine if learners have passed or failed in the acquisition of an English vocabulary size prescribed by a national curriculum-based wordlist. To operationalize vocabulary size, this study focuses on measuring the active recognition dimension of word knowledge (Laufer, Elder, Hill, & Congdon, 2004). More specifically, the vocabulary size test is designed to tap into EFL learners' fundamental ability to recognize the form-meaning connection of words. Taken together, the extent to which CAT can function as a mechanism to replace a normally lengthy vocabulary size test still remains largely unexplored and unverified in the literature.

2. Literature review

There has been an extensive amount of research on measuring EFL learners' English vocabulary size over the past three decades in the field of English language education. Among the different lines of this research, the issue of how to measure a learner's English vocabulary size accurately and reliably has been a focal point of investigation. Acquiring sufficient vocabulary size, after all, is critical for mastering a language, and ample research has indicated that English vocabulary size is a significant indicator of language ability (Laufer & Goldstein, 2004; Milton, 2009). Empirical evidence, for example, has shown a significant, positive relationship between vocabulary size and speaking (Milton, 2009), listening (Stæhr, 2009), and writing (Stæhr, 2008); vocabulary size, in particular, appears to be the strongest predictor for reading comprehension ability (Laufer & Ravenhorst-Kalovski, 2010). Clearly, mastering a language relies very much on the acquisition of a wide range of words, thus prompting Alderson (2005) to remark that "language ability is to quite a large extent a function of vocabulary size" (p.88).

Vocabulary size is one of the central criteria for evaluating one's vocabulary growth. Creating and developing valid and reliable measurements of English vocabulary size has therefore become an important task for most L2 vocabulary researchers (Nation & Webb, 2011). Typically, a fixed-length test format is used as a universal template to assess test-takers of various abilities, with all test takers being required to take an identical set of test items in the same order (Schultz, Whitney, & Zickar, 2014). According to the research done in this field, fixed-length tests usually fall into two categories: A peaked test design or a rectangular test design (Weiss, 1985). In describing the psychometric characteristics of the two conventional test designs, Weiss (1985) proposes a well-known thesis – "the bandwidth-fidelity dilemma" – to explain the predicaments that are likely to be encountered in developing the two types of conventional tests. A peaked conventional test typically contains numerous items with difficulties centering on a pre-determined level of difficulty, so that the test itself can differentiate very well for examinees whose ability levels are approximate to the chosen level of difficulty. As shown in Fig. 1, a peaked conventional test can procure a relatively high degree of measurement precision around the peaked region of ability levels, whereas it gradually fails to serve the aim along the two ends of an ability continuum due to an insufficient number of test items designed for lower and higher ability levels of examinees. That is, fidelity is held, but bandwidth is sacrificed in a peaked conventional test. On the contrary, a rectangular conventional test refers to designing a set of items with a wide range of difficulty levels suited for all levels of ability. Unless a very long test is allowed, only a few items can be selected for each ability level. As shown in Fig. 1, although a rectangular test can have a relatively equal level of measurement precision along the ability continuum, the overall test precision is low because only a few items can be included for each ability level. In other words, bandwidth is saved, but fidelity is lost in a rectangular test.

To avoid the psychometric predicaments that are likely to arise in fixed-length tests, Weiss (1983, 2004) suggests that a possible solution is to adopt a dynamic and flexible testing algorithm to select test items to fit ability estimations during the testing process. The main tenet underlying the adaptive algorithm is to select the next test item whose difficulty is tailored to a test taker's provisional ability estimate. (Thompson & Weiss, 2011). This adaptive feature makes it possible that not only can sufficient items be provided for all ability levels of examinees but also all examinees can receive sufficient items that are

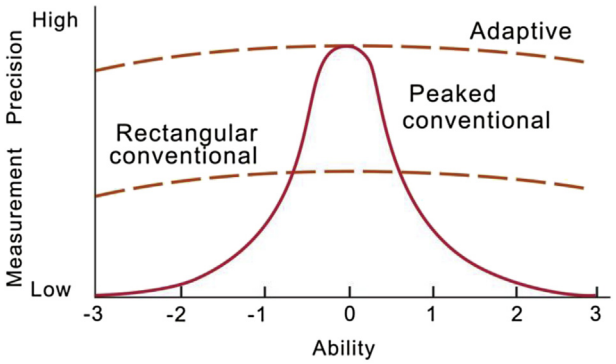


Fig. 1. Measurement precision as a function of trait level for an adaptive test and for peaked and rectangular conventional tests.
Source: Weiss (1985, p. 775).

optimally targeted to their ability estimates. As illustrated in Fig. 1, an adaptive test can ensure a full realization of high measurement precision and a wide range of bandwidth provisions across all ability levels. Clearly, the adaptive item selection strategy taken by CAT enables a more fine-grained distinction between test-takers abilities which are of approximately the same level than fixed-item tests. This high level of discriminating power is constant across all levels of ability. Therefore, regarding vocabulary assessment, if a fixed-item test is used, adopting either a peaked or a rectangular test format is less likely to secure both fidelity and bandwidth in light of test quality. Specifically, the fixed length test format may result in unstable and imprecise estimates of vocabulary size for very high and very low ability groups of vocabulary learners (Schultz et al., 2014). Early research has shown that adaptive tests are more predictive of achievement test scores than conventional tests (Thompson & Weiss, 1980), and the concurrent validity coefficients for conventional tests have been consistently lower than those of adaptive tests (Martin, McBride, & Weiss, 1983). Similarly, adaptive tests have been shown to be more efficient by saving at least 50% more of test items than those that could be used in a typical P&P test (Vispoel, 1993; Ward, 1984). A more recent study in personality assessment further indicates that adaptive tests are more precise than P&P tests (Kantrowitz, Dawson, & Fetzer, 2011). These empirical findings have led Schultz et al. to conclude that "CAT, as compared to traditional paper-and-pencil tests, can be both more effective and more efficient" (p. 326).

Still another important test quality related to CAT is test fairness. It is critical that a test should be fair, in a way that all the test takers, regardless of their ability level, should have an equal chance of receiving a sufficient amount of test items that are targeted to their ability level (Kunnan, 2014; McNamara & Ryan, 2011). Bridgeman, Bejar, and Friedman (1999), applying computer-based tests to an architectural licensure examination, also suggested that in their pass/fail decision study, sufficient tasks of comparable difficulty needed to be designed and constructed in order to achieve test fairness. Although these researchers did not adopt CAT in the study, their argument that test takers need to receive sufficient tasks to elicit their optimal performance echoes the sentiments expressed in Kunnan's and McNamara and Ryan's works, and CAT offers tremendous potential for providing this called-for fairness in test administration. In a strict sense, the 'fairness' argument suggested by Kunnan, McNamara and Ryan, and Bridgeman et al. is isomorphic with the 'bandwidth' thesis proposed by Weiss (1985).

Because of its psychometric advantages over fixed-item tests as well as its potentials in realizing test fairness, CAT has been enacted and implemented in a variety of domains for different test decisions over the past two decades. These applications of CAT are inclusive of not only large-scale standardized testing but also classroom-based diagnostic checking. For example, CAT has been used for learning diagnoses in classroom settings (Tatsuoka & Tatsuoka, 1997; Wang, Chang, & Huebner, 2011), a mastery (pass-fail) examination for licensure in nursing (National Council of State Boards of Nursing, 2015), and high stakes international language proficiency tests (Economides & Roupas, 2007). Typically, CAT licensing exams require fewer items than high stakes tests, in that the goal of the latter is to estimate and differentiate examinees' abilities as precisely as possible (Schultz et al., 2014).

Comparing the score equivalence between P&P and computer-based tests (CBT) is essential if CAT is intended to be taken as a more efficient and precise test alternative (Wang, Jiao, Young, Brooks, & Olson, 2008). Because of the high-quality and flexibility of interface design, test administrations through computers have become more and more popular and welcome by both practitioners and stakeholders. Over the past three decades, numerous comparative studies regarding the two modes of test delivery have been undertaken in other disciplines such as education, psychology, mathematics, and ergonomics (Sawaki, 2001). Synthesizing the findings from diverse disciplines, Paek (2005) concluded that "evidence has accumulated to the point that it appears that the computer may be used to administer tests in many traditional multiple-choice test settings without any significant effect on student performance" (p. 1). Empirical studies tend to support the notion that most CBTs are equivalent to the test scores of P&P tests (Bridgeman et al., 1999; Poggio, Glasnapp, Yang, & Poggio, 2005; Pommerich, 2004; Pomplun, Frey, & Becker, 2002), leading Wang et al. (2008) in their meta-analysis study to argue that "the administration mode had no statistically significant effect on K-12 student reading achievement scores" (p. 1).

Research on score comparability between CAT and P&P tests in language testing, however, remains scant (Sawaki, 2012). Although there have been a number of CAT-application studies in language testing, early comparative studies were mostly undertaken in an L1 (English as native language) vocabulary context (e.g., Vispoel, 1993, 1998; Vispoel et al., 1994), and focused on the comparability between CAT and fixed-item (i.e., P&P) tests and the effects of individual differences on CAT administration. Vispoel (1993) and Vispoel et al. (1994) found that CAT required only 13 items (or 67.5% fewer) to yield better reliability and validity than those acquired by the P&P test with 40 items. However, Vispoel's studies were predominantly carried out in the L1 context, and the P&P vocabulary tests in the studies were short-length in format and did not provide adequate evidence for construct validity. More importantly, as mentioned above, they were not particularly designed for estimating non-native speakers' English vocabulary size, the role of which is pivotal to language development. To elaborate further, a typical P&P-based English vocabulary size test developed for non-native speakers may go up to 180 items (e.g., The Eurocentres Vocabulary Size Test: Meara & Buxton, 1987), and the vocabulary items in such vocabulary size instruments have a much wider range of item difficulties. Vispoel's 40-item P&P test now appears dated and immaterial to EFL context. Therefore, it has yet to be demonstrated whether the same level of efficiency can be achieved if the vocabulary items in P&P tests are validated for measuring English vocabulary size, as opposed to short-length tests. For this reason, we argue that there is a pressing need for a score comparability study of CAT and P&P measurements of English vocabulary size, and it is hoped that the findings of the present study will provide a new approach for the measurement of English vocabulary.

A typical P&P-based vocabulary size test focuses on measuring learners' written receptive knowledge about the form-meaning connection of words. Exemplar vocabulary size tests such as the Vocabulary Levels Test also adopt this format, in which learners are required to select target words from the options to match their corresponding English definitions (Schmitt, Schmitt, & Clapham, 2001). Although knowing a word involves more than just establishing the link between word form and word meaning, Laufer and Goldstein (2004) remark that "The most important component of word knowledge is the ability to establish *the link between word form and word meaning* " (p. 409, *italics* original). Furthermore, the acquisition of more advanced types of word knowledge such as collocations and associations also inevitably entails an understanding of stimulus words' meanings (Schmitt, 2010). Last but not least, Schmitt and Schmitt (2014) critically review a wide range of empirical studies (e.g., Cobb, 2007; Nation, 2006) and argue for the significance of teaching mid-frequency vocabulary, the range of which varies between high-frequency (3000) and low-frequency (9000+) vocabulary boundaries. They suggest that mid-frequency vocabulary levels should be included in learning materials and programs, but in actuality such a wide range of levels can pose a great challenge for instruction. Aside from the pedagogical challenge, equally challenging is how to efficiently and reliably measure a word band that covers up to 6000 word families. In sum, based on the foregoing review of previous studies, we set up a research agenda to conduct a series of empirical studies to examine the potentials of applying CAT to assess EFL learners' English vocabulary size.

3. Purpose of the study

To bridge the existing research gap, the aim of this study was to examine the feasibility and practicality of CAT as a mechanism to measure English vocabulary *size*. Three research questions thus guided the current study:

1. Can a unidimensional item bank for a vocabulary size test be validated?
2. To what extent are the CAT-based vocabulary size estimates comparable to the P&P-based ones?
3. Can the CAT-based vocabulary size estimates be used to determine mastery/non-mastery of a vocabulary size threshold as set by a wordlist from a national curriculum guideline?

3.1. Phase I: development and validation of the item bank

3.1.1. Word selection from College Entrance Examination Center wordlist

The basis for this study's CAT item bank was a wordlist that was compiled under the sponsorship of the College Entrance Examination Center (CEEC) in Taiwan, and that has been used to define national curriculum goals since its publication in 2002. The wordlist was designed for two levels of English tests — the Scholastic Aptitude English Test (SAET) and the Department Required English Test (DRET) — in college entrance examinations in Taiwan and referred to by textbook publishers to develop English textbooks for high school students. The wordlist was originally composed from a variety of other lists, including from a number of different kinds of textbooks and wordlists used in Britain, Canada, China, Japan, Taiwan, and the United States. Both English-as-first-language and English-as-foreign-language materials were selected, compared, and balanced in a way that the best word candidates and optimal levels of words could be determined for learning. The wordlist was divided into six frequency levels and has served as a baseline for EFL learning and teaching in Taiwan since its publication. Each level contains 1080 words, totaling 6480 words for the whole list. Typically, the six levels were ordered on the basis of word frequency, with words of lower levels being more frequent. Research has shown that Level 1 and Level 2 can cover on average 85.18% of the number of words used in the EFL General English (GE) textbooks published by international publishers, ranging from low-intermediate to the advanced level (Hsu, 2009). Therefore, learners' efforts dedicated to the learning of the wordlist will not be in vain and can be rewarding in their future pursuit of a higher education degree.

The items adopted in CAT came from each level of the CEEC wordlist, such that for every sequence of twenty-seven words, a word was selected as a test item, resulting in each level having an equal proportion of selected words (40 words). Firstly, a multiple-choice format with four options was adopted to design the vocabulary size test because this test format was most familiar to ordinary test-takers regardless of their ethnic backgrounds, and secondly, it was adopted because this format provided the best likelihood of local independence, a requisite construct assumption for Rasch modeling (exemplified in the following sample item).

Although other format options are possible, such as the matching format appearing on the Vocabulary Levels Test (Nation, 2013; Schmitt, 2010) and the Yes/No format used by Meara and Buxton's (1987) The Eurocentres Vocabulary Size Test, these measures have been criticized as "relatively superficial" (Meara, 2009, p. 74) and suffer from evident psychometric problems. For instance, the Eurocentres Vocabulary Size Test simply requires test-takers to check whether they know the meanings of words, and this greatly increases the likelihood of inflated estimates of English vocabulary size (Schmitt, 2010). Similarly, the issue with the Vocabulary Levels Test is its matching format, and the items within each block are unavoidably interdependent (Milton, 2009; Ronald & Kamimoto, 2014). If a test-taker marks an item wrong and that wrong answer is the correct answer for another item in the same block, this will lead to local dependence within the test block, and the test-taker's vocabulary size will be biasedly estimated. This psychometric problem of the test has been acknowledged by its original test developers (Nation, 2013; Nation & Webb, 2011; Schmitt, 2010), leading Schmitt (2010) to remark that "the test is not really designed to provide an estimate of a person's overall vocabulary size" (p. 198). However, if the problem of local dependence can be solved, the Vocabulary Levels Test can continue to be used to shed new light on the field. In brief, in light of the psychometric weaknesses of the other two format alternatives, using a multiple-choice test format appears to be the safest way to elicit test-takers' test responses because each array of options is uniquely valenced with a single item stem and contains only one correct response.

3.1.2. Rasch analysis for the item bank

An initial item pool of 240 vocabulary items based on the CEEC wordlist was developed and validated using the Rasch unidimensional measurement model, for which item difficulty and person ability were simultaneously computed and calibrated. The 240 vocabulary test items were evenly divided into five parallel tests. Each test involved a set of 45 independent items and a set of 15 common (linking) items, for an overall number of 240 items. In total, 1536 senior high school students from Taiwan were recruited to take the five tests. It should be noted that the subjects did not need to take all the five tests, as a concurrent equating method was used to calibrate all 240 item parameters through the linking items. The 1536 test-takers were divided into five groups, with each group containing around 300 test-takers and each accounting for a different vocabulary test.

The Rasch model was adopted to calibrate the item difficulty parameter in the test bank and to check the quality of each vocabulary item. The computer software RUMM2030 (Andrich, Sheridan, & Luo, 2010) was used to model the test response matrix (1536 × 240). The results showed that 53 items failed to fit the Rasch model, and were subsequently removed from the item pool. To ensure a balanced distribution of the number of words sampled from each word level, it was decided that 30 words should be retained for each level, and this totaled 180 items to form the item bank for the computerized adaptive test. With the inclusion of 30 items for each frequency level, the reliability indices (Cronbach's Alpha) of the six levels all surpassed 0.80. The item bank was then crossed-validated with the Vocabulary Size Test (Beglar, 2010), and the correlation between the two tests was quite high ($r = 0.85$).

The mean fit statistics of the item bank and the associated mean person measures are presented in Table 1. The fit residuals, with mean and standard deviation approximating to 0 and 1 respectively, provided robust evidence that in the current study the empirical test data fit the Rasch unidimensional measurement model. Furthermore, Table 2 indicates that the Person Separation Index (PSI) calculated by RUMM2030 was 0.90, suggesting that the power of the test of fit was excellent, and the whole item bank could reach a high level of reliability (see Discussion).

Table 2 also reports the results of the item-trait interaction of the item bank. As can be seen from this table, the item–trait interaction probability was 0.167 ($\chi^2 = 1674.823$, $df = 1620$), meaning that there was no significant interaction between item difficulty and person measure in the latent trait space. This finding provides solid support for the assumption that unidimensionality underlying the Rasch model was met, as there was a very high agreement among the responses for the different difficulties of all the items on the latent scale (Andrich & Van-Schoubroeck, 1989).

Since the property of unidimensionality underlying the item bank is paramount to the implementation of CAT (Flaugher, 2000; Steinberg, Thissen, & Wainer, 2000), it was necessary to provide an additional, complementary check to examine whether the item bank was truly unidimensional. To triangulate the findings derived from the Rasch model, the DIMTEST statistical test (Stout, Froelich, & Gao, 2001) was further employed. Unlike parametric IRT models such as the Rasch model, the DIMTEST solution is essentially a non-parametric method and hence can provide extra information for detecting unidimensionality. DIMTEST 2.1 was used to test whether the null hypothesis – i.e., essential unidimensionality – could be

Having doubt (A) Skeptical	(B) Fabulous	(C) Controversial	(D) Alcoholic

Table 1
Means and standard deviations of fit residuals for item difficulty and person measure.

	Item	Person
	Fit residual	Fit residual
Mean	−0.013	−0.336
Standard Dev.	1.063	1.071

Table 2
The person separation index and item-trait interaction.

Person separation index (PSI)	0.90
Item-trait interaction	
Total item chi-square	1674.823
Total degree of freedom	1620
Total chi-square probability	0.167

Note. The PSI refers to the proportion of the observed variance that is considered true, ranging between 0 and 1.

maintained. The results show that the value of the *T* statistic was 0.372 ($p = 0.3586$), indicating that the null hypothesis could not be rejected. In sum, the DIMTEST provided complementary support for the unidimensionality underlying the vocabulary item bank.

With the confirmation of unidimensionality, the test information function (TIF) and the corresponding conditional standard error measurement (CSEM) of the item bank were further plotted (Fig. 2). The TIF, denoted by the blue solid curve, and the CSEM, indicated by the red solid-circle dash curve, are mutually reciprocal: the higher the TIF, the smaller the CSEM. In other words, the higher the TIF, the more informative the test; and, the smaller the CSEM, the more precise the whole test. Hence, the two curves in Fig. 1 reveal that the whole item bank was both informative and precise in estimating the targeted English vocabulary size level. For example, the standard error of the theta estimate approximated zero in the theta scale between −2 and +2 logits and only increased approximately 0.15 toward either −3 or +3 logits on the theta scale.

The score conversion from theta into vocabulary size was also calculated by RUMM2030. As indicated in Fig. 3, the score conversion between theta and vocabulary size was not straightforwardly linear, but rather non-linear.

3.2. Phase II: comparability study

3.2.1. CAT system, item selection, and termination criteria

CATPRO 1.0.0.4, developed by *Assessment Systems*, was used as the platform to implement the CAT vocabulary size test. The difficulty indices of the vocabulary items in the item bank (180 items) were keyed into the CAT system. A Bayesian estimation approach—*Bayes Modal Estimation (MAP)* – was employed to estimate the test-takers' CAT scores because it could estimate the all-correct and all-incorrect CAT scores before the termination criterion was reached (van der Linden & Pashley, 2010). The

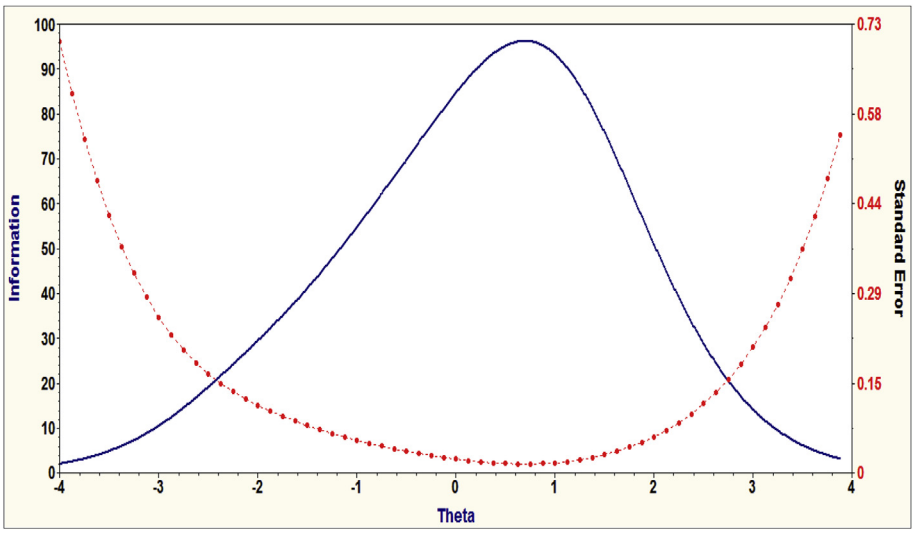


Fig. 2. Test information function and the corresponding conditional standard error measurement of the item bank.

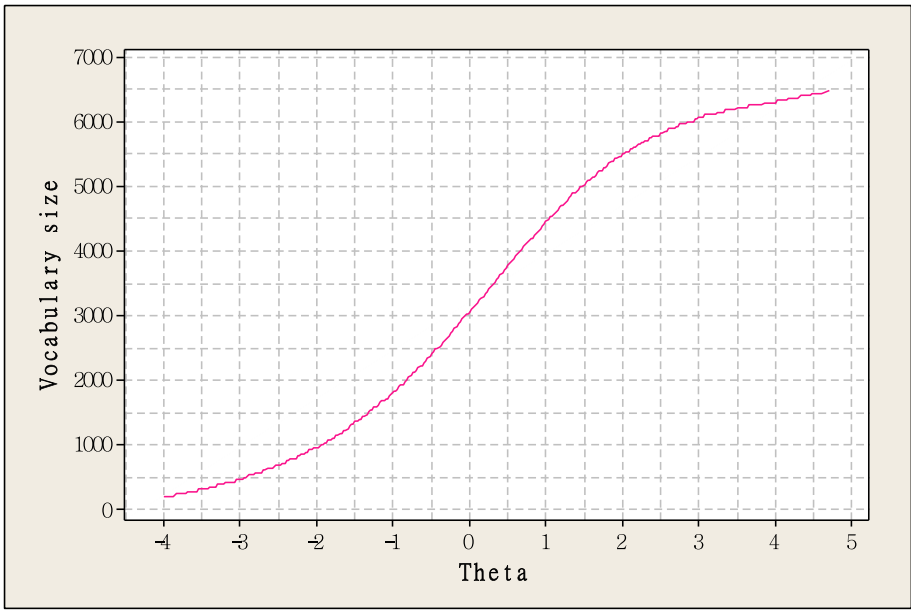


Fig. 3. The non-linear relationship between theta and vocabulary size.

scores of the P&P test were also estimated using the *MAP* method to ensure the subsequent comparability between the CAT scores and the P&P scores for the same test-takers. Regarding the item selection algorithm, the items with difficulty indices falling between −0.5 and +0.5 were randomly selected as the 1st items. Moreover, regarding the termination criterion, two types of context were created: the fixed-length had three different specified conditions (30-Item, 60-Item, and 90-Item), and the variable-length contexts used three other conditions (*SE* at 0.40, *SE* at 0.30, and *SE* at 0.20). The sample item screen of the CAT implementation and the item bank of the CAT vocabulary size test are respectively illustrated in Figs. 4 and 5.

3.2.2. Procedure

In total, 211 senior high school students participated in the second phase of the study. Seventy of the participants were recruited from the tenth grade, 70 from the eleventh grade, and another 71 from the twelfth grade. Participants' term grades were also collected. Then, the participants were further classified into six groups according to the six different termination conditions: 37 test-takers in the 30-Item condition, 33 in the 60-Item, 36 in the 90-Item, 34 in the 0.40 *SE*, 34 in the 0.30 *SE*, and 37 in the 0.20 *SE*.

The subjects in all the six conditions were first required to complete the CAT test. As the CAT test required that each test-taker receive different initial items and undergo a diverse item selection process, this procedure prevented the memory effect of systematically remembering the same vocabulary items that would otherwise appear in a P&P test. The subjects in the fixed-length context could complete the CAT test within 30 min, whereas most test-takers in the variable-length only spent 5–10 min to finish the test, with the exception of only a few cases. After the CAT test was completed, the full-bank test was then administered, with test-takers having a 20-min break before taking the full-bank test. On average, the full-bank test could be finished in an hour, which was much longer than what would be required in the CAT context.

3.2.3. Results

Table 3 reports the correlations between CAT-theta and full-bank theta. In the fixed-length termination context, the computed CAT-theta based on the three fixed-length conditions all had very strong correlations with the full bank theta. It was found that all the correlation coefficients in the fixed-length termination context exceeded 0.95. However, in the variable-length termination context, the correlations between the CAT-theta and the full bank theta were on average lower than those in the fixed-length context.

Table 4 further shows the average, minimum, and maximum test items needed in different termination conditions. The average numbers of test items required in the variable-length conditions were 10 for *SE* fixed at 0.40, 18 for *SE* at 0.30, and 43 for *SE* at 0.20. The *SE* fixed at 0.40 utilized the fewest items (Min. $N = 9$), followed by *SE* at 0.30 (Min. $N = 16$) and *SE* at 0.20 (Min. $N = 35$). The maximum number of test items observed for the SE at 0.30 (Max. $N = 90$) and the *SE* at 0.20 (Max. $N = 90$) entailed the use of 50% of all the test items in the item bank, whereas for the *SE* at 0.40, only 30% of the test items in the full bank had to be administered.

Paired *t*-tests were adopted to examine whether significant differences existed between the CAT-theta and full-bank theta for the six different termination criteria. The results of the paired *t*-tests are presented in Table 5. It was found that with the exception of *SE* at 0.20, the theta comparisons between the CAT and full-bank test format reached statistical significance in the

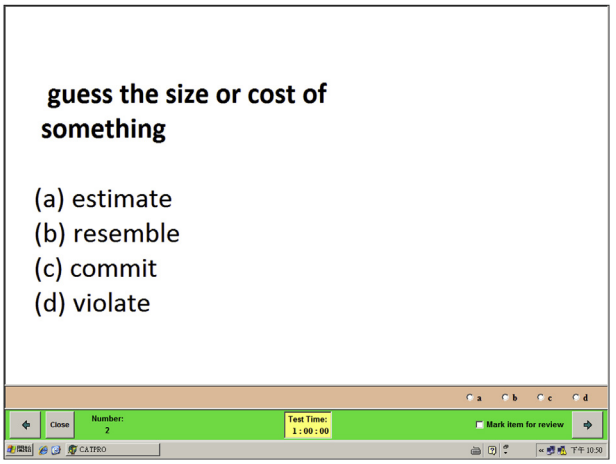


Fig. 4. Sample item of CAT vocabulary size test.

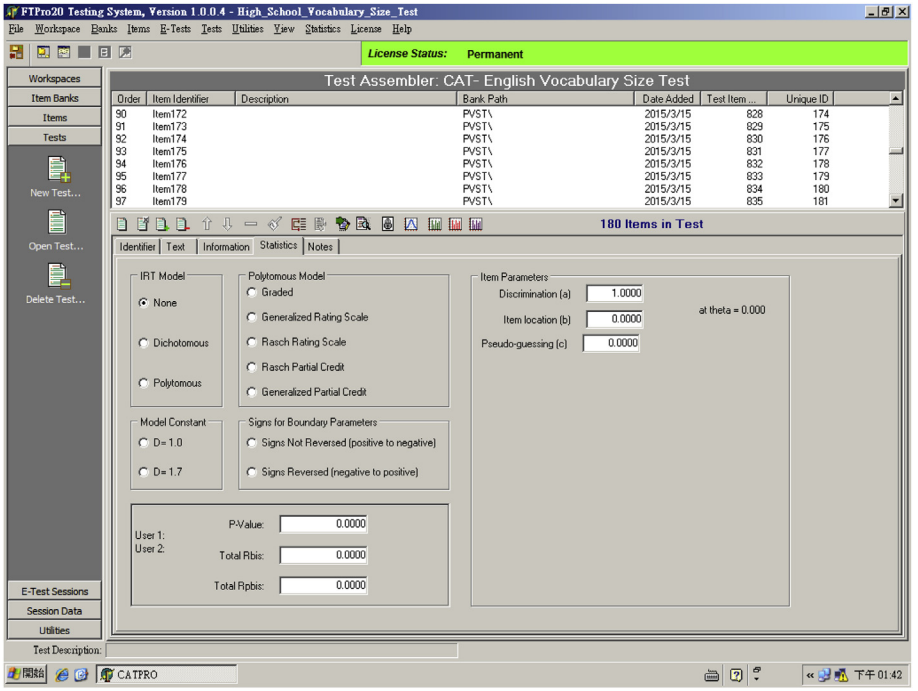


Fig. 5. The item bank of CAT vocabulary size test.

Table 3
Correlations between CAT-Theta and full-bank theta.

	Fixed-length context			Variable-length context		
	30-Item	60-Item	90-Item	SE at 0.40	SE at 0.30	SE at 0.20
Corr.	0.96	0.98	0.99	0.91	0.94	0.97

Note. All the correlation coefficients reached statistical significance with $p < 0.001$.

other two variable-length termination conditions. In the fixed-length conditions, statistical significance was only achieved in the 30-Item condition but not in the 60-Item and 90-Item conditions.

The conversions from theta to the expected number of correct responses and expected vocabulary size estimates are further listed in Table 6. In the 90-Item condition, the expected number of correct responses and the expected vocabulary size

Table 4
Maximum, minimum, and average test items in the fixed-length and variable-length context.

	Fixed-length context			Variable-length context		
	30-Item	60-Item	90-Item	SE at 0.40	SE at 0.30	SE at 0.20
Max.	Null	Null	Null	54	90	90
Min.	30	60	90	9	16	35
Average	30	60	90	10	18	43
% of saving	83.3%	66.7%	50%	94.4%	90%	76.1%

Table 5
Paired *t*-test between CAT-Theta and Full-Bank Theta.

	No. of subjects	CAT-theta		Full-bank theta			
Termination criterion	*n*	*M*	*SD*	*M*	*SD*	$t_{(df)}$	*p*
30-Item	37	0.18	0.76	0.05	0.73	4.44	<0.001
60-Item	33	0.31	0.53	0.35	0.61	−1.78	0.084
90-Item	36	−0.36	0.44	−0.38	0.48	1.32	0.197
SE at 0.40	34	0.11	0.53	−0.17	0.53	5.79	<0.001
SE at 0.30	34	0.20	0.66	0.08	0.74	2.21	<0.05
SE at 0.20	37	0.56	0.38	0.59	0.52	−1.04	0.306

Table 6
Theta conversion into expected number correct and vocabulary size estimate.

Termination criterion	CAT-theta	Expected number correct	Expected vocabulary size	Full-bank theta	Expected number correct	Expected vocabulary size
30-Item	0.18	92	3312	0.05	87	3168
60-Item	0.31	98	3528	0.35	99	3564
90-Item	−0.36	71	2556	−0.38	71	2556
SE at 0.40	0.11	90	3240	−0.17	79	2844
SE at 0.30	0.20	93	3348	0.08	89	3204
SE at 0.20	0.56	107	3852	0.59	109	3924

estimates were found to be exactly identical. In the 60-Item condition, the CAT-theta and the full-bank theta only differed by one item in terms of the expected number of correct responses (99−98 = 1) and 36 words regarding the expected vocabulary size estimate (3564−3528 = 36); two items and 72 words in the 0.20 *SE* condition; four items and 144 words in both the 0.30 *SE* condition and the 30-Item condition. The largest mismatch was observed in the 0.40 *SE* condition, in which the CAT-theta and the full-bank theta differed by eleven items, or the equivalent of 396 words.

Fig. 6 illustrates the scatter plot between the CAT-theta and the corresponding standard error in the six termination conditions. It can be seen that the fixed-length termination conditions in general had smaller standard errors than the variable-length conditions. Likewise, all the standard errors in the fixed-length termination conditions were below 0.30, but the 30-Item condition appeared to have larger standard errors than the 0.20 *SE* condition.

3.3. Phase III: mastery/non-mastery study

3.3.1. Deciding the cut-off value of English vocabulary size

The decision on the cut-off value of English vocabulary size was based on the sample in the first-phase study. A K-Means cluster analysis was used to determine the pass/fail vocabulary size threshold, with the algorithm for conducting a K-Means cluster analysis being adopted from the R package "trimcluster." To run a K-Means cluster analysis, it is necessary to decide the number of clusters before running the formal analysis. As only two groups – pass vs. fail – were required, the number of clusters was set at two. Further, the contrasting group method under the K-Means cluster analysis was further taken to objectively determine the cut-off score. The results based on the contrasting group method indicated that 100 was the cut-off point with 1096 test takers passing the test and a pass rate of 63.8%. The estimated vocabulary size corresponding to 100 items was 3360 words [(100/180) × 6048 = 3360]. In Taiwan, 1200 words are stipulated as the vocabulary requirements to be learned at the junior high school level, and 2440 words are specified for regular senior high English textbooks as *words for production* that must be acquired, which means that learners must be able to spell out the words. The total number of words to be mastered from junior and senior high school education are thus about 3640 words as included in the CEEC wordlist, which is slightly higher than the vocabulary size cut-off. It can be argued that 3360 words is a reasonable threshold score, given that the cut-off point is very close to the minimum requirement of 3640 words prescribed by the curriculum guidelines.

To transform the vocabulary size into the Rasch logit score, it was necessary to first convert backward, from the vocabulary size to the expected number of correct responses, and then convert the expected number of correct responses to the Rasch

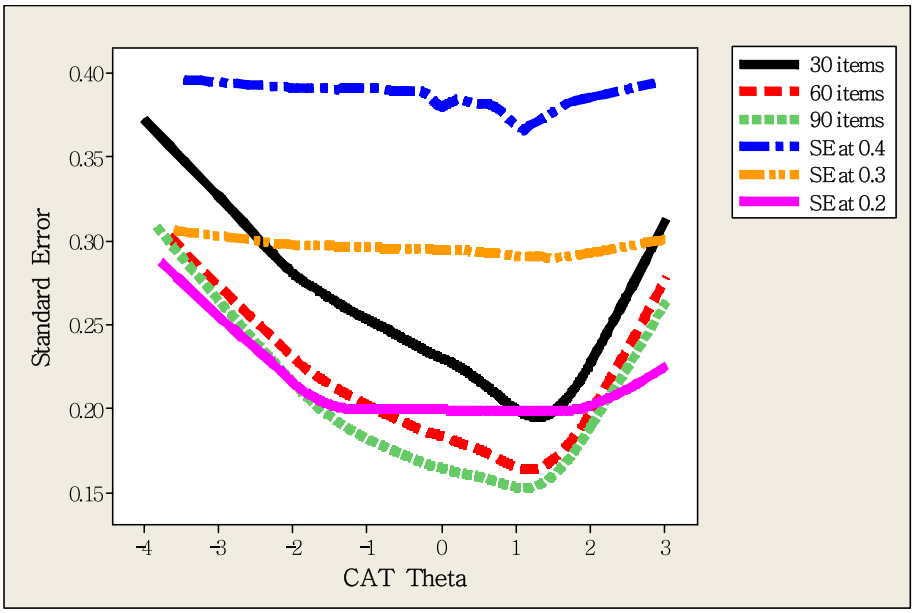


Fig. 6. The scatter plot between the CAT-Theta and the corresponding standard error in the six termination conditions.

logit score. Through this conversion procedure, the cut-off theta estimate corresponding to the expected number of correct responses positioned at 100 was 0.21 logit.

3.3.2. Procedure

Participants in the third-phase study were additionally recruited, and in total 563 senior high school students partook in the third-phase classification study. As in the second phase, the subjects in all the six conditions were first required to complete the CAT test. After a 10-min break, the P&P test was administered. The Bayesian estimation method (*MAP*) was again taken to estimate the logit value in both test modes. The termination criterion in this mastery/non-mastery phase of the study was specified as follows: To successfully pass the cut-off score of 0.21, the 95% confidence interval of the error band around the final theta estimates had to be above and could not overlap the cut-off score of 0.21. Conversely, to determine whether one failed to pass the cut-off value, the 95% confidence interval of the error band around the final theta estimates had to be below and could not overlap the cut-off score of 0.21. An example of the four-item adaptive theta estimation process in the mastery/non-mastery study is given in Fig. 7, and the corresponding dynamic item selection sequence by increasing difficulty steps is further shown in Fig. 8.

3.3.3. Results

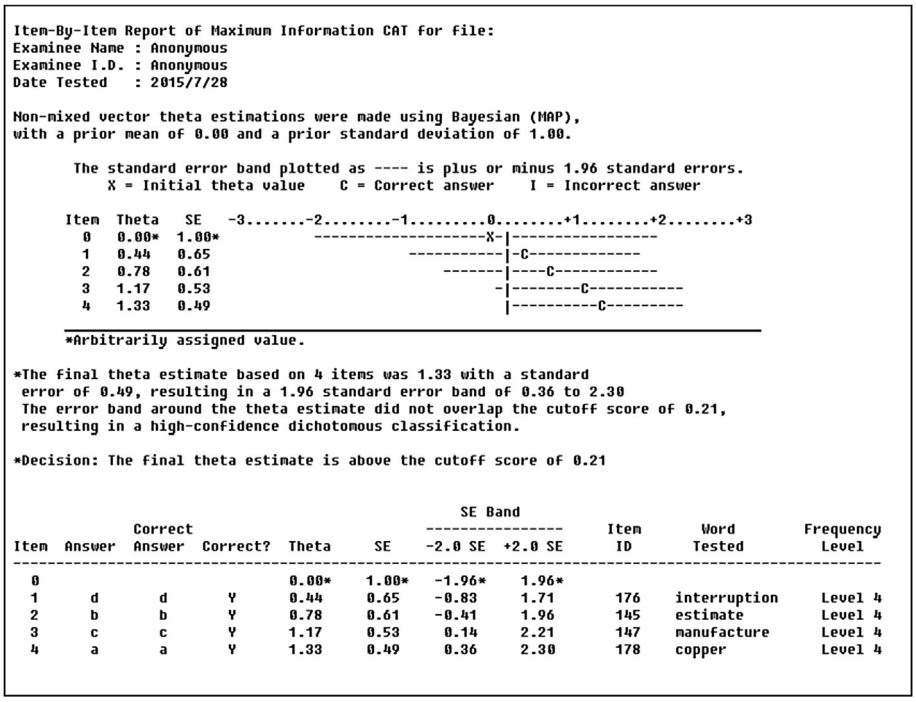
Regarding the agreement of classification on pass/fail of the cut-off theta, it could be seen from Table 7 that CAT classification exhibited 100 percent agreement with the full-bank classification in judging whether test-takers had successfully passed the cut-off logit (0.21), whereas a minor disagreement (3.82%) was detected on deciding whether test-takers had unambiguously failed to pass the cut-off theta. A chi-square goodness-of-fit test showed there was no significant difference in the classification above or below the vocabulary size threshold between the two test modes, $\chi^2(1) = 0.10$, $p = 0.75$. Compared to the P&P test condition in which all the 180 items needed to be used, the CAT test mode only had to administer a handful of items ($M = 21$, $SD = 20$, $Min. = 12$, $Max. = 37$) to complete the classification task.

As shown in Fig. 9, in the area where the two standard error curves intersected, it was found that the standard error of CAT theta estimates revealed a more consistent and stable performance, while that of the full-bank theta estimates showed an erratic, unexpected high-peaked anomaly. This classification disagreement could be possibly attributed to the larger, higher-peaked standard error observed on the full-bank standard error curve at the theta value of around 0.21 logit.

4. Discussion

4.1. RQ1: can a unidimensional item bank for a vocabulary size test be validated?

In the current study, computerized adaptive testing was employed to resolve the trade-off of conventional vocabulary size tests that require a “daunting” number of items to achieve precise ability estimates. A vocabulary item bank of 180 items was developed and validated with the Rasch model, and the selection of vocabulary items was based on the CEEC wordlist.



```
Item-By-Item Report of Maximum Information CAT for file:
Examinee Name : Anonymous
Examinee I.D. : Anonymous
Date Tested   : 2015/7/28

Non-mixed vector theta estimations were made using Bayesian (MAP),
with a prior mean of 0.00 and a prior standard deviation of 1.00.

      The standard error band plotted as ---- is plus or minus 1.96 standard errors.
          X = Initial theta value    C = Correct answer    I = Incorrect answer

     Item  Theta   SE   -3.......-2.........-1.........0.........+1.........+2.........+3
       0   0.00*  1.00*           --------------------X-|------------------
       1   0.44   0.65                         ----------|-C----------
       2   0.78   0.61                             -------|----C------------
       3   1.17   0.53                                   -|--------C----------
       4   1.33   0.49                                    |----------C---------

     *Arbitrarily assigned value.

*The final theta estimate based on 4 items was 1.33 with a standard
 error of 0.49, resulting in a 1.96 standard error band of 0.36 to 2.30
 The error band around the theta estimate did not overlap the cutoff score of 0.21,
 resulting in a high-confidence dichotomous classification.

*Decision: The final theta estimate is above the cutoff score of 0.21


                                                    SE Band
              Correct                               ----------      Item       Word        Frequency
Item  Answer  Answer  Correct?  Theta     SE      -2.0 SE  +2.0 SE   ID       Tested       Level
-------------------------------------------------------------------------------------------------
  0                             0.00*   1.00*    -1.96*    1.96*
  1     d       d       Y       0.44    0.65     -0.83     1.71      176    interruption    Level 4
  2     b       b       Y       0.78    0.61     -0.41     1.96      145    estimate        Level 4
  3     c       c       Y       1.17    0.53      0.14     2.21      147    manufacture     Level 4
  4     a       a       Y       1.33    0.49      0.36     2.30      178    copper          Level 4
```

Fig. 7. The adaptive theta estimation process in the cut-off termination condition.

The results of the Phase I study (constructing an item bank) showed strong support for the high reliability and strong unidimensionality of the item bank. Few items worked against the prediction made by the Rasch model. The high value of PSI suggests that the true variance being measured in the observed score accounted for a large proportion, which is to say that the high PSI value of the item bank indicates that the calibrated person measures had very small measurement errors (<0.015; see Fig. 1) and could be well-separated along the theta scale. Furthermore, in the current research context, the fulfillment of unidimensionality means that only the *size* dimension – and not other forms of vocabulary knowledge – was being targeted and measured. Robust evidence for the support of unidimensionality of the whole vocabulary size bank can be seen in the absence of significant item-trait interaction, which suggests that every item parameter (i.e., item difficulty) was applicable to all the test-takers and that each person measure (i.e., vocabulary size) was attainable by all the items in the item bank. This Rasch-guided check of unidimensionality was also further corroborated by DIMTEST, a non-parametric method that can complement the Rasch-based findings. In sum, both the parametric and non-parametric statistical methods demonstrated the fulfillment of the assumption of unidimensionality underlying the item pool in the study. Given that "adaptive testing assumes unidimensional pools" (Flaugher, 2000, p. 56) and also that "The assessment of unidimensionality is central for CAT" (Steinberg et al., 2000, p. 209), the confirmation that the item bank can be unidimensional lends support for the comparability of CAT thetas that are derived from adaptively administered vocabulary items since each test-taker appears to receive items sequenced differently according to his/her interim test performance in the CAT test context. In brief, ample evidence has been procured for the construct validity of the item bank, suggesting the vocabulary items in the item bank measure the single construct of vocabulary size. This is a very important property to lend support for the subsequent deployment of CAT implementation. Without a valid item bank, the results of the CAT study simply would not be reliable and trustworthy.

This study makes an important step forward in estimating English vocabulary size by proposing a creative, dynamic, and efficient way of measuring EFL learners' English vocabulary size, which has long been a focal concern for vocabulary researchers. To cover a wide range of vocabulary size, common practice has demonstrated that more than a hundred vocabulary test items are required: the Vocabulary Levels Test (VLT) contains 150 test items (Nation, 1990, 2013; Schmitt, 2010); the Eurocentres Vocabulary Size Test (EVST) includes 180 items (Meara & Buxton, 1987); and (3) the Vocabulary Size Test (VST) consists of 100 items (Beglar, 2010). Furthermore, specifically regarding the number of items needed in the five word frequency levels of the EVST, Meara (2010) admitted that "This would be daunting even for an extremely keen test-taker" (p.3). Clearly, there is a trade-off between reliability and parsimony regarding measuring EFL learners' English vocabulary size. On the one hand, to achieve high reliability, numerous vocabulary items are necessary from each word level; on the other hand, if parsimony becomes the concern in the test design, then lower reliability is the cost. Currently no vocabulary size test can simultaneously accomplish both reliability and parsimony at fewer than 100 items. Completing this type of P&P vocabulary size test is easily influenced by undesirable mental and physiological consequences such as testing fatigue, boredom, anxiety, or loss of attention. Arguably, when these confounding test variables intervene in the test process, the measurement errors of

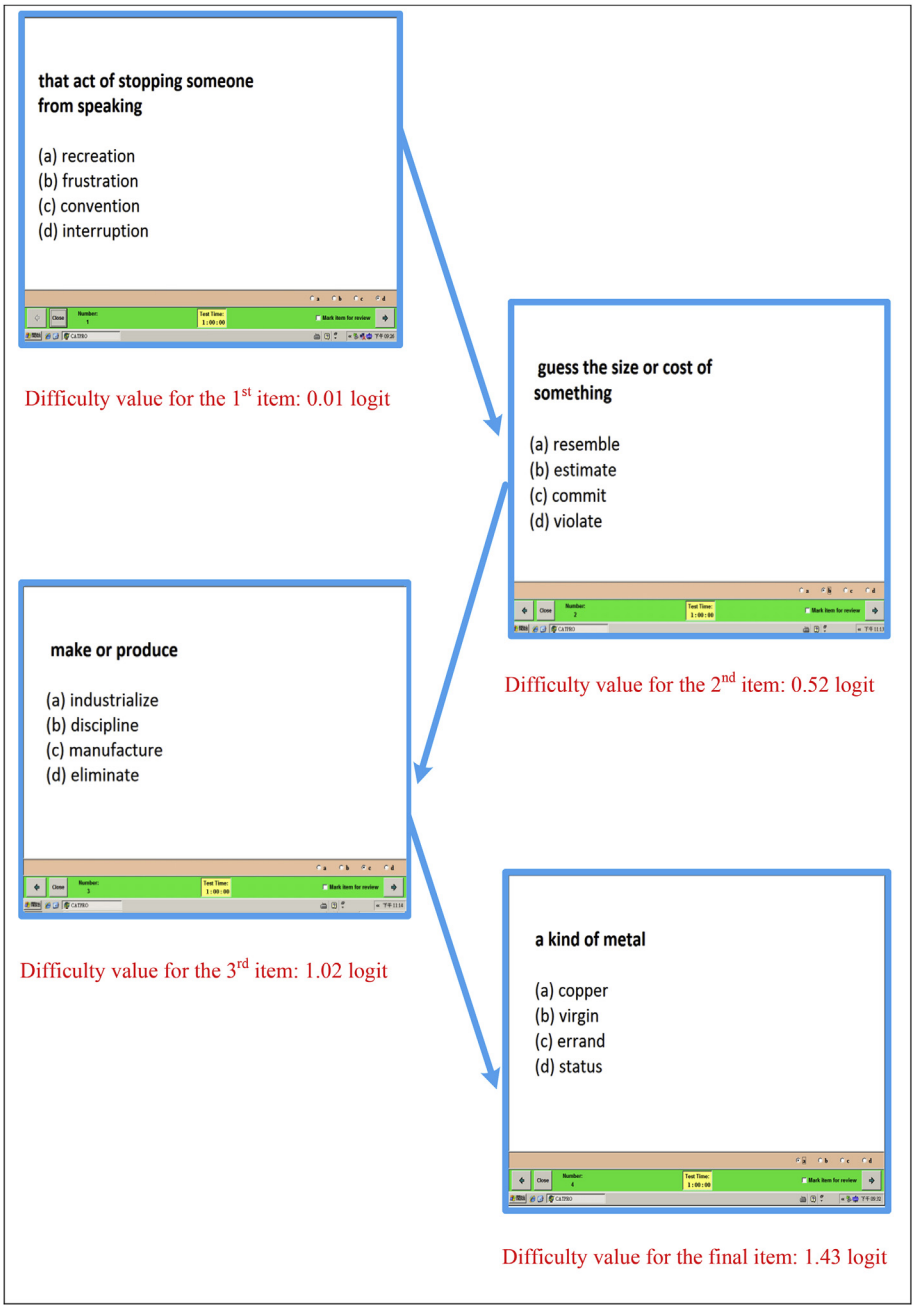


Fig. 8. The dynamic item selection sequence of four contiguous correct responses.

Table 7
Classification agreement summary.

	N	Percent	Percent of total
Full bank classification: ABOVE	327		58.08
CAT classification: Above	327	100.00	58.08
CAT classification: Below	0	0.00	0.00
Full bank classification: BELOW	236		41.92
CAT classification: Above	9	3.82	1.60
CAT classification: Below	227	96.18	40.32

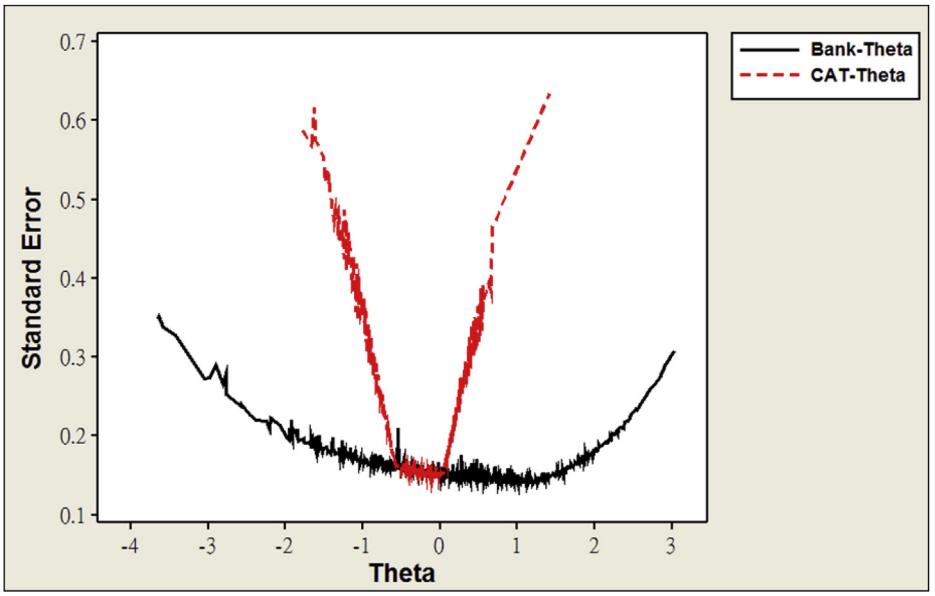


Fig. 9. The Scatter Plot between the Theta and the Standard Error in the classification of above or below the cut-off value.

the test scores may become large, as evidenced in phase three of the study, rending the theta estimates questionable and even misleading. Ultimately, conventional P&P tests in reality suffer from a fundamental psychometric problem: the item selection process is always fixed and static. That is, the test length is fixed, and the capacity to determine the interaction between test items and test takers is also absent. In a CAT testing context, however, these psychometric and test-taking performance issues during the testing process can be effectively resolved and efficiently amended.

4.2. RQ2: to what extent are the CAT-based vocabulary size estimates comparable to the P&P-based ones?

Although the total number of vocabulary test items in the item bank of this CAT study is the same as that of the EVST (Meara, 2010), the results indicate that the number of the test items required for the CAT condition is at least half that of the full-bank. It is important to note that in the CAT testing scenario all the items except for the very 1st item arbitrarily assigned were chosen so that the difficulties of the items could be tailored to locally fit the test-takers' interim theta estimates during the test process. This dynamically adaptive and constantly informative item selection process contrasts starkly with the P&P test approach and could never be achieved by a P&P test. Because the test items required in the CAT context can on average be saved up to 94.4% (Table 4), the test-takers do not as easily experience test anxiety or boredom and hence can keep fully focused on facing vocabulary test items that most suit their ability without being influenced by test fatigue. More critically, the results of the study provide supportive evidence for implementing CAT in estimating English vocabulary size, as the correlations between CAT estimates and the full-bank estimates in all six termination conditions were all fairly high ($r > 0.90$). This finding provides robust criterion validity for our proposed CAT option, and this finding is consistent with earlier studies' (Klinkenberg et al., 2011; Nirmalakhandan, 2007), which also corroborate the close link between the two test modes. Furthermore, in addition to the high correlations, a number of *insignificant* theta estimate contrasts between the variable-length and fixed-length termination context further support the feasibility of replacing the P&P test with the CAT test in estimating EFL learners' English vocabulary size. The CAT-based theta estimates were not only strongly correlated with but also highly comparable to those derived from the P&P test. Among the series of theta comparisons using the paired *t*-tests, three termination conditions showed no statistical difference: 60-Item, 90-Item, and *SE* at 0.20. This finding echoes the study by Huang, Lin, and Cheng (2009), who also found no statistical difference between CAT-based and P&P-based ability estimates.

In consideration of which termination criterion should be used in estimating English vocabulary size via CAT, the 60-Item condition appears to be the most appropriate choice. In the *SE* at 0.20 condition, although the average number of items administered was the smallest (saving 76.1% of item bank) among the three, the maximum number administered nonetheless reached 90, which was identical to the 90-Item condition. Hence, where parsimony is concerned, *SE* at 0.20 can only be deemed at best equally efficient as the 90-Item condition. Similarly, although the average number of items required in the 60-Item condition was approximately 1.4 times ($60/43 = 1.40$) more than that needed in the *SE* at 0.20 condition, the test-takers in the 60-Item condition did not have to take up to 90 items as would be likely observed in the *SE* at 0.20 condition. In actuality, the correlation between the CAT-theta and the full-bank theta was higher in the 60-Item condition than in the *SE* at

0.20 condition. Hence, executing CAT with the fixed-length stopping rule at 60 items can maintain both parsimony in administering vocabulary items and reliability in producing comparable theta estimates with the full-bank ones. More importantly, the CAT test mode can ensure that every test taker will receive a sufficient amount of vocabulary items that are adaptively targeted to his or her ability level, so that the test itself will be deemed fair across all the ability levels (Kunnan, 2014; McNamara & Ryan, 2011). In connection to early comparability studies (Vispoel, 1993; Vispoel et al., 1994), both Vispoel's studies and the current study confirm that CAT-based tests are not only much more efficient in estimating test-takers' vocabulary ability but are also more economical in saving item administration, with the scores derived from the two test modes being still highly comparable. Therefore, the results of the study further demonstrate that using computerized adaptive tests to replace traditional P&P tests to estimate EFL learners' English vocabulary size is indeed viable.

4.3. RQ3: Can the CAT-based vocabulary size estimates be used to determine the mastery/non-mastery of the vocabulary size threshold as set by a wordlist from a national curriculum guideline?

The results of the study provide supportive evidence for implementing CAT in estimating English vocabulary size regarding the written receptive dimension of word knowledge, as the correlation between CAT estimates and the full-bank estimates was fairly high ($r = 0.90$). In reality, the CAT estimates can be even more convincing and trustworthy in classifying test-takers in a context where a predetermined, curriculum-based vocabulary size threshold is used as the termination criterion. As shown in Fig. 6, the standard error curves of both CAT and full-bank estimates overlapped in the theta interval between −0.5 and +0.5. The two standard curves should be expected to closely fit each other if the CAT and P&P theta estimates are equally precise. However, this was not the case in this study. We found that the standard error of the full-bank theta estimates experienced an abrupt high-peaked, unstable fluctuation, while that of the CAT theta estimates remained in a constantly unwavering pattern. This misfit phenomenon can be possibly explained by the odds that in taking the full-bank P&P vocabulary size test, lower-proficiency learners are in practice more likely to experience a high level of physical fatigue, boredom and test anxiety than higher-proficiency learners. The indirect evidence for this claim comes from the health field in which CAT has been widely employed (Polit & Yang, 2015). As Polit and Yang remark, “CAT administration can decrease patients' fatigue, boredom, and frustration, especially when multiple traits are being assessed” (p. 91). Hence, based on our research findings, we argue that the adaptive, dynamic items selection approach taken by CAT is more precise in classifying test-takers than the fixed, static item selection method adopted by the P&P-based test.

Test fairness is an essential factor to be considered in test implementation. Test fairness refers to an operationalized sense of justice underlying the procedures that are taken to “avoid the effects of any *construct-irrelevant factors* on the entire testing process” (Walters, 2012, p. 470; *italics* in the original). As shown in the mastery/non-mastery study, the P&P-based test incurred more unexpected randomness than the CAT-based test. Insofar as the impact and far-reaching consequences that pass/fail decisions can have on test takers, our findings suggest that although a long vocabulary size test can surely guarantee a high reliability, it may at the same time undermine the construct validity of the test (Walters, 2012). The construct-irrelevant factors in this classification study can be described as boredom, fatigue, or systematic guessing behaviors that are likely to arise in taking all the 180 items at once. Both the current study and Bridgeman et al. (1999) have investigated the effects of implementing computer-based tests on pass/fail decision-making of ability mastery in a specific domain. The results of the current study are not only in line with those of Bridgeman et al. but also further suggest that test fairness in the context of licensure or vocabulary size tests is more likely to be realized and maintained in CAT-based tests compared to P&P-based tests.

That CAT can be more efficient and precise in classifying test-takers into mastery and non-mastery groups has significant practicality and utility in diagnosing whether EFL learners have mastered the English vocabulary size either prescribed by the curriculum or guided by theory. For instance, based on an investigation of the relationship between vocabulary size and text coverage across nine written and spoken corpora, Nation (2006, 2013) theorizes that 4000 word families are required for understanding 95% of text coverage of novels and newspaper articles, whereas to cover an additional 3% (i.e., up to 98%), at least 8000–9000 word families are needed. Laufer and Ravenhorst-Kalovski (2010) also share similar views and suggest 4000 word families as the minimal threshold and 8000 word families as the optimal threshold for reading comprehension. Therefore, in relation to the utility of mastery/non-mastery decision making, the test can be further employed to determine whether EFL learners have passed the two fundamental vocabulary size thresholds for reading comprehension. More importantly, research has argued for the usefulness and benefits of acquiring mid-frequency vocabulary for proficient and authentic language use (Schmitt & Schmitt, 2014). Because a mid-frequency vocabulary level covers another 6000 word families on top of high-frequency vocabulary, to reliably and efficiently measure such a wide frequency band of words *independent* of both high-frequency and low-frequency vocabulary poses a great challenge to the designers of the conventional P&P test format. Because high-frequency words are learned before mid-frequency words, a vocabulary test focusing on mid-frequency vocabulary should exclude high-frequency words for the purpose of test efficiency. If a P&P vocabulary test is used for this purpose, it is unlikely to check whether learners have the preparatory knowledge of the 3000 high-frequency vocabulary terms in taking the test. However, this problem can be easily handled in a CAT-delivery test. For instance, in a large calibrated item bank that includes all the high-, mid-, and low-frequency words, the starting rule of theta estimate in the CAT test can be fixed at the boundary value between high- and mid-frequency words, and then the test can progress adaptively according to learners' vocabulary knowledge. Likewise, to check whether learners have mastered the mid-frequency vocabulary, the termination criterion can be set at the cut-off theta value corresponding to 9000 words in the calibrated item

bank. This way, the CAT-delivery vocabulary test makes it possible to diagnose each learner to a high level of precision, as well as to measure all ability groups with high and equal levels of precision. Both fidelity and bandwidth psychometric principles can be solidly upheld in the CAT test format (Weiss, 1985, 2004). Pedagogically, school faculty can help implement CAT software in the classroom to help English teachers not only efficiently measure but also precisely diagnose their students' English vocabulary size.

Regarding the diagnostic potential of CAT, because the items selected by the CAT algorithm are the most informative and targeted to the test-takers' ability level, teachers can track which items have been taken and determine the factors that cause students to answer incorrectly. With this knowledge, effective, individualized vocabulary learning strategies can be implemented. Likewise, during the period of high school education, EFL high school students in Taiwan are normally required to learn up to several thousand new words. In this situation, EFL high school students in Taiwan could effectively receive CAT diagnoses at different time points to regularly track their growth in vocabulary size. Upon receiving the CAT diagnostic feedback, English teachers can quickly realize how many words students have learned and systematically check the strengths and weaknesses of students' language learning progress from students' advances in word levels. Therefore, because of the diagnostic potential that CAT holds, measuring English vocabulary size with CAT may help create an autonomous classroom environment in light of 'assessment *for* learning' rather than simply an 'assessment of learning' (Kalyuga, 2013), and this can provide the great opportunity to promote and facilitate 'self-directed learning' in the task of acquiring English vocabulary (Hsu, Zhao, & Wang, 2013).

In his discussion of the strengths of CAT on measuring lexical competence, Schmitt (2010) argues that there are two advantages of CAT-based test delivery modes: One is its swift adaptiveness and dynamism of moving between word frequencies in response to examinees' lexical competence, and the other one is its better and more valid sampling of words that are appropriately tailored to and targeted on examinees' true lexical profiles. Schmitt concludes that "This adaptiveness has a great advantage over static tests, where either the test administrators must guess the frequency levels to give to the examinees, or the examinees must work their way through the whole test in a lockup fashion" (p. 202). It can be said that the measurement of vocabulary knowledge has entered a new era, where CAT measurements of English vocabulary knowledge are opening a new window through which L2 mental lexicons can be more fully and precisely understood, diagnosed and assessed.

In sum, the research implications of this CAT study not only support its use in replacing traditional P&P tests, but also reveal its potential in tracking the trajectory and development of vocabulary size from the fundamental 2000-word level to the essential 5000-word level. The results of this CAT study serve as a heuristic point of departure in guiding future work in this line of research inquiry.

5. Conclusion

In a conventional mode of P&P test delivery, being assessed for 180 vocabulary items to gauge English vocabulary size can be daunting and de-motivating for most test-takers. The current study shows that testing EFL learners' English vocabulary size using CAT can reduce the needed test items by up to two-thirds while still producing fairly close and comparable vocabulary size estimates to the ones using all of the 180 items. In the dynamic mode of CAT test administration, the process of item selection by CAT is tailored to test-takers' provisional vocabulary size estimates. This study suggests that CAT has incredible potential in not only efficiently measuring EFL learners' English vocabulary size but also precisely diagnosing whether they have mastered a targeted vocabulary size threshold. The results of the study constitute a *prima facie* rationale for implementing CAT in both assessing and diagnosing EFL learners' English vocabulary size.

This study, however, is not without limitations. First, it is limited in its scope of item selection from a prescribed curriculum. Researchers need to move one step further to refer to other standardized vocabulary size tests and run more replication studies. Second, the study is also limited in its coverage of word knowledge. Word knowledge is multi-faceted in nature. Although this pioneering study shows the potentials of using CAT to measure the receptive dimension of form-meaning links of words, its utility in measuring other facets of word knowledge such as collocation or register has yet to be determined. Future research efforts should therefore be dedicated to these two lines of research inquiries.

Acknowledgements

This study is supported in part by the National Science Council of the Republic of China under contract numbers NSC-97-2410-H-003-054 and NSC-99-2410-H-003-081-MY2. The author would like to thank Prof. David Weiss from University of Minnesota and Dr. Nate Thompson from Assessment Systems for their assistance in updating the computerized adaptive testing system used in the study.

References

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

Andrich, D., Sheridan, B., & Luo, G. (2010). *RUMM2030: A windows-based item analysis program employing Rasch unidimensional measurement models*. Perth: Western Australia: RUMM Laboratory.

Andrich, D., & Van-Schoubroeck, L. (1989). The General Health Questionnaire: a psychometric analysis using latent trait theory. *Psychological Medicine, 19*, 469–485. http://dx.doi.org/10.1017/S0033291700012502.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*, 101–118. http://dx.doi.org/10.1177/0265532209340194.
Bridgeman, B., Bejar, I. I., & Friedman, D. (1999). Fairness issues in a computer- based architectural licensure examination. *Computers in Human Behavior, 5*, 419–440.
Chalhoub-Deville, M. (1999). *Issues in computer-adaptive testing of reading proficiency*. Cambridge: University of Cambridge Local Examinations Syndicate.
Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*, 1–20.
Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology, 11*, 38–63.
Dunkel, P. (1999). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency*. Cambridge: University of Cambridge Local Examinations Syndicate.
Economides, A. A., & Roupas, C. (2007). Evaluation of computer adaptive testing systems. *International Journal of Web-Based Learning and Teaching Technologies, 2*, 70–87.
Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed.). Mahawah, New Jersey: Lawrence Erlbaum Associates, Inc.
Gusev, M., & Armenski, G. (2014). E-assessment systems and online learning with adaptive testing. In M. Ivanović, & L. C. Jain (Eds.), *E-Learning paradigms and applications*. Heidelberg: Springer.
Hsu, W. H. (2009). College English textbooks for general purposes: a corpus-based analysis of lexical coverage. *Electronic Journal of Foreign Language Teaching, 6*, 42–62.
Hsu, C. L., Zhao, Y., & Wang, W. C. (2013). Exploiting computerized adaptive testing for self-directed learning. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-Pacific*. New York: Springer.
Huang, Y. M., Lin, Y. T., & Cheng, S. C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education, 52*, 53–67.
Kalyuga, S. (2013). Rapid dynamic assessment learning. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-pacific*. New York: Springer.
Kantrowitz, T., Dawson, C., & Fetzer, M. (2011). Computer adaptive testing (CAT): a faster, smarter, and more secure approach to pre-employment testing. *Journal of Business and Psychology, 26*, 227–232.
Kaya-Carton, E., Carton, A. S., & Dandonoli, P. (1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.
Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*, 1813–1824.
Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1098–1114). Malden, MA: Wiley.
Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing, 21*, 202–226. http://dx.doi.org/10.1191/0265532204lt277oa.
Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: size, strength, and computer adaptiveness. *Language Learning, 54*, 399–436. http://dx.doi.org/10.1111/j.0023-8333.2004.00260.x.
Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*, 15–30.
Lilley, M., Barker, T., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education, 43*, 109–123.
van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden, & C. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer.
Madsen, H. S. (1991). Computer-adaptive testing of listening and reading comprehension: the Brigham Young University approach. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.
Malabonga, V. (2000). *Trends in foreign language assessment: The computerized oral proficiency instrument*. NCLRC Newsletter. http://www.cal,org/nclrc.
Malabonga, V., & Kenyon, D. (1999). Multimedia computer technology and performance-based language testing: a demonstration of the computerized oral proficiency instrument. In M. B. Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processing*. New Brunswick, NJ: Association for Computational Linguistics.
Martin, J. T., McBride, J. R., & Weiss, D. J. (1983). *Reliability and validity of adaptive tests in a military recruit population (ONR TR 83–1)*. Arlington, VA: Personnel and Training Programs, Office of Naval Research (NTIS No. AD-A129 324).
Mayrath, M. C., Clarke-Midura, J., & Robinson, D. H. (2012). *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.
McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: the place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*, 161–178.
Meara, P. (2009). *Connected words: Word associates and second language vocabulary acquisition*. Amsterdam: John Benjamins.
Meara, P. (2010). *EFL vocabulary tests* (2nd ed.). Swansea: Lognostics.
Meara, P., & Buxton, B. (1987). An alternative to multiple choice will vocabulary testing. *Language Testing, 4*, 142–154. http://dx.doi.org/10.1177/026553228700400202.
Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Maters.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Mass: Newbury House.
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59–82.
Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.
National Council of State Boards of Nursing. (2015). *Computerized adaptive testing*. CAT. Retrieved from https://www.ncsbn.org/1216.htm.
Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle.
Nirmalakhandan, N. (2007). Computerized adaptive tutorials to improve and assess problem-solving skills. *Computers & Education, 49*, 1321–1329.
Paek, P. (2005). *Recent trends in comparability studies* [Research Report]. Pearson Education Measurement. Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TrendsCompStudies.pdf?WT.mc_id=TMRS_Recent_Trends_in_Comparability_Studies.
Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3*, 1–30.
Polit, F. F., & Yang, F. M. (2015). *Measurement and the measurement of change*. New York: Wolters Kluwer.
Pommerich, M. (2004). Developing computerized versions of paper tests: mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2*, 1–44.
Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*, 337–354.
Ronald, J., & Kamimoto, T. (2014). Confidence in word knowledge. In J. Milton, & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge*. Basingstoke: Palgrave Macmillan.
Rudner, L. M. (2010). Implementing the graduate management admission test computerized adaptive test. In W. J. van der Linden, & C. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer.
Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning and Technology, 5*, 38–59.
Sawaki, Y. (2012). Technology in language testing. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing*. Oxon: Routledge.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. London: Palgrave MacMillan.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*, 484–503. http://dx.doi.org/10.1017/S0261444812000018.
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing, 18*, 55–88. http://dx.doi.org/10.1177/026553220101800103.
Schultz, K. S., Whitney, D. J., & Zickar, M. J. (2014). *Measurement Theory in Action* (2nd ed.). Hove, Sussex: Taylor & Francis.
Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed.). Mahawah, New Jersey: Lawrence Erlbaum Associates, Inc.
Stevenson, J., & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.
Stout, W., Froelich, A., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijin, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–375). New York: Springer-Verlag.
Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*, 139–152. http://dx.doi.org/10.1080/09571730802389975.
Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*, 577–607. http://dx.doi.org/10.1017/S0272263109990039.
Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: effect on remedial instruction as empirical validation. *Journal of Educational Measurement, 34*, 3–20.
Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*. Mahawah, NJ: Lawrence Erlbaum Associates, Inc.
Thompson, J. G., & Weiss, D. J. (1980). *Criterion-related validity of adaptive testing strategies (RR80–3)*. Minneapolis: Department of Psychology, University of Minnesota.
Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*, 1–9.
Verschoor, A. J., & Straetmans, G. (2010). MATHCAT: a flexible testing system in Mathematics education for adults. In W. J. van der Linden, & C. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer.
Vispoel, W. P. (1993). Computerized adaptive and fixed-item versions of the ITED vocabulary subtest. *Education and Psychological Measurement, 53*, 779–788. http://dx.doi.org/10.1177/0013164493053003022.
Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: the role of answer feedback and test anxiety. *Journal of Educational Measurement, 35*, 155–167. http://dx.doi.org/10.1111/j.1745-3984.1998.tb00532.x.
Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: a comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*, 53–79. http://dx.doi.org/10.1207/s15324818ame0701_5.
Walters, F. S. (2012). Fairness. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing*. Oxon: Routledge.
Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Education Measurement, 48*, 255–273. http://dx.doi.org/10.1111/j.1745-3984.2011.00145.x.
Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: a meta-analysis of testing mode effects. *Education and Psychological Measurement, 68*, 5–24.
Ward, W. C. (1984). Using microcomputers to administer tests. *Educational Measurement, 3*, 16–20. http://dx.doi.org/10.1111/j.1745-3992.1984.tb00744.x.
Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774–789. http://dx.doi.org/10.1037/0022-006X.53.6.774.
Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*, 70–84.
Wu, H. K., Kuo, C. Y., Jen, T. H., & Hsu, Y. S. (2015). What makes an item more difficult? effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. *Computers & Education, 85*, 35–48.