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