Support Adaptive Testing: Towards a New Future in Language Education
Jaturapitakkul, N. and Watson Todd, R.
CD Proceedings of the 4th Language in the Realm of Social Dynamics International Conference "The Multi-Dimensions in an Era of Language and Teaching". pp. 54 - 59. University of the Thai Chamber of Commerce, Bangkok.

The definitive version of this article was published as Jaturapitakkul, N. and Watson Todd, R. (2012) Support Adaptive Testing: Towards a New Future in Language Education. CD Proceedings of the 4th Language in the Realm of Social Dynamics International Conference "The Multi-Dimensions in an Era of Language and Teaching". pp. 54 - 59. University of the Thai Chamber of Commerce, Bangkok.

Support Adaptive Testing: Towards a New Future in Language Education

NATJIREE JATURAPITAKKUL, RICHARD WATSON TODD
King Mongkut's University of Technology Thonburi
Bangkok
natjiree.jat@kmutt.ac.th, irictodd@kmutt.ac.th

Abstract

Tests play a central role in language education, often with detrimental effects. In Thailand, multiple-choice testing dominates with very little evidence of learning, unfortunate washback effects, and no clear meaning to test results. Given that the role of tests in Thai education is unlikely to change, new testing approaches that fit the practicalities of test use while having more positive effects are needed. This paper presents an innovative approach to computer-based testing that incorporates scaffolds, namely, Support Adaptive Tests (SATs), and compares these against other forms of language tests. In SATs, an incorrect response results in the same item being presented again but with support provided, either through text adaptation (e.g. highlighting the relevant part of the text) or item adaptation (e.g. reducing the number of options in multiple-choice). SATs should provide a better

measure of test-taker potential, facilitate learning while testing, and encourage scaffolded instruction through washback.

1. The role of tests in education and directions for development

Assessment (or the collection of any types of measurement of learning) is crucial in education, and in many contexts the most common form of assessment is testing (usually formal, one-off measurements). While educational assessment reflects the values and expectations of the educational systems and societies in which it is conducted (McNamara & Roever, 2006), assessment practices often have additional wide-ranging and unintended consequences. In Thailand, for instance, multiple-choice testing dominates English language assessment with many pernicious effects. Currently, all locally-produced high-stakes tests, such as the university entrance exam, rely solely on multiple-choice items, and multiple-choice accounts on average for over half of the scores for English courses at secondary schools (Piboonkanarax, 2007). The dominance of multiple-choice testing means that it acts as an implicit policy instrument (Darasawang & Watson Todd, 2012) and is the greatest cause for concern among teachers, even more worrying than large class sizes, low student proficiency and excessive workload (Thongsri, Charumanee & Chatupote, 2006), perhaps because of its negative washback effects. These include a focus on grades rather than learning, a focus on receptive skills, the promotion of rote learning of simplistic knowledge, and a reduced emphasis on higher-level thinking (Brown, 2005; Burke, 1999; Forsyth, Joliffe & Stevens, 1999; Watson Todd, 2008).

The somewhat depressing assessment situation in Thailand is common in many other similar EFL countries (Watson Todd & Shih, forthcoming), and, despite numerous calls for change, is likely to remain an issue for the foreseeable future. Making large-scale changes in educational systems, such as shifting from multiple-choice testing to other forms of assessment, is a time-consuming process with a low chance of success (Fullan, 1987). A more realistic approach is to investigate how multiple-choice testing can be developed to, at least, minimise its negative effects and, hopefully, to enable multiple-choice testing to have positive impacts on education.

The goal of this paper is to investigate ways to develop multiple-choice testing and to propose an innovative method of conducting multiple-choice tests based on scaffolding, namely, support adaptive tests (SATs). In doing this, we take a two-pronged approach. First, from a theoretical perspective, we will examine how current multiple-choice testing practices measure up to the ideals of testing. Second, we will investigate the perceptions of students, a largely overlooked group of important stakeholders in assessment (Pino-Silva, 2008, being a notable exception), and their suggestions for developing tests. From the findings of these two investigations, we will examine how the proposed SATs have the potential to benefit English language education.

2. The ideals of testing and existing multiple-choice tests

While it is impossible to create a definitive list of ideal test characteristics since each test serves its own unique purpose, for tests to be useful, they need to manifest reliability, validity, authenticity, interactiveness, appropriate impact and practicality (Bachman, 2004). Multiple-choice tests, when well-designed, generally have high reliability and practicality. The key issues we will look at in this section, then, are validity and impact, especially in terms of washback.

Numerous sub-categories of validity have been proposed in the literature (see e.g. Fulcher, 2003; Khalifa & Weir, 2007; Weir, 2005). Several of these, such as context validity, construct validity and criterion-related validity, are dependent on the design and implementation of individual tests in specific contexts and thus are not of relevance in a discussion of generally applicable testing methods. We will therefore focus on four types of validity in this section, namely, content validity (concerning whether the appropriate objectives are covered), method validity (concerning whether the method of testing is appropriate), learning validity (concerning

whether learning is facilitated during testing, see Tomlinson, 2005), and consequential validity (or washback).

Traditional multiple-choice testing, as used in many contexts in Thailand, has, as we have seen, substantial negative washback. In part, this is because it limits language objectives assessed to reading and linguistic knowledge. If not supplemented by other forms of assessment, there are therefore likely to be content validity problems. These tests are also inherently inauthentic meaning that their method validity suffers since it is unclear how applicable results from such inauthentic tests are to authentic situations. Furthermore, traditional multiple-choice tests provide no opportunities for learning since scores are often delayed, no information concerning which items were answered incorrectly is provided, and the tests promote ephemeral memorisation of surface knowledge (Watson Todd, 2007). It would therefore seem that traditional multiple-choice tests fail on all four aspects of validity under consideration.

An alternative form of multiple-choice testing that became fashionable in the 1990s is computer-adaptive testing (CAT). "Computer-adaptive refers to the procedure where an item(s) is selected on-line for each test taker based on his/her performance on previous items" (Chalhoub-Deville, 1999: ix). Claimed advantages of CAT include the level of difficulty being challenging but not discouraging, and fewer questions being needed to measure a test taker's level (Chapelle & Douglas, 2006; Ockey, 2009). Even though CAT can be more efficient than traditional testing, it suffers the same validity problems and may even create additional problems. For instance, since different test takers answer different items depending on their level, judgments about the extent to which the test covers the stated objectives, or content validity, are more difficult to make than for traditional tests. CAT would therefore not appear to be the answer in our search for more effective multiple-choice testing.

Before we examine how the various types of validity are manifested in SATs, we will first look at two investigations of students' views on language testing to see if they can help in identifying directions for developing multiple-choice tests.

3. Students' views on language testing

3.1 Study 1: Students' perceptions of traditional English testing in Thailand

In many formal language learning situations, including those in Thailand, traditional testing dominates. Paper-and-pencil tests are typically used for the assessment of separate components of language (grammar, vocabulary) and for receptive understanding (listening and reading comprehension). Test items in such tests are often in convergent fixed-response formats, such as multiple-choice, and students are rated in relation to how many right answers they give. There are, however, some problems with using such tests, such as the lack of assessment of productive skills and ignoring the assessment of higher-order thinking.

Most previous research studies and published articles in relation to traditional assessment have examined the views of researchers and educators. However, students' views as a large important group of stakeholders have usually been overlooked (Pino-Silva, 2008). Therefore, an investigation of students' views on problems and preferences in traditional testing, particularly in Thailand, was conducted. 323 undergraduate students from multidisciplinary programmes at King Mongkut's University of Technology Thonburi were randomly selected to participate in the study. To give a chance for students to voice what they perceived about traditional English tests freely, an open-ended questionnaire was used to collect data. Four main questions were included in the questionnaire concerning fairness of the tests, format of the tests, content validity of the tests, and learning opportunities from the tests. The students' responses were transcibed and analysed as a corpus for frequency, allowing the data to be categorised into salient themes.

In regard to the fairness of the tests, the issue of unclear marking was prominent since minimal feedback on performance was given and no information concerning which items were answered incorrectly is provided. The issue of fairness in testing was also related by the students to unreal assessment of all English skills, but especially speaking and listening. Moreover, most of the tests the students had taken focused on grammar and vocabulary that the test takers had never been taught or never used in daily life. Tests were also perceived as being unfair if the test takers could comprehend the test content or meaning of the test task, but misinterpreted instructions and test questions.

Regarding the format of the tests, students viewed the format of the tests as being an issue of little relevance for those who are proficient in English. Indeed, some formats of tests could help facilitate test taking in some skills, such as multiple-choice items in reading tests, but these were not applicable in speaking, listening and writing. In addition, although the present format of the tests does not seem to be a good way to assess students' English ability for all skills, it helps facilitate understanding for weak students.

Concerning content validity in the tests (i.e. whether the test measures what they have learned in class), to some extent the tests can measure important content in English courses they learned in class particularly linguistic knowledge (i.e. grammar and vocabulary). Most of the time, however, they had been assessed on general knowledge of English with unseen topics and difficult vocabulary. If there is difficult content that needs application (analyzing or reasoning), it is likely that they cannot perform to their best on the tests.

The last concern is if the tests promote learning. Most respondents stated that the tests do not help them learn but put them under pressure, a fact especially applicable for weaker students. The tests are not administered for learning purposes but only for the sake of assessing how well the test-takers interpret and understand the test questions. Furthermore, the tests do not promote learning since they have never been told what the right answers in the test are. Last but not least, some parts of the test which contain previously unseen words and complex structures cause difficulty and then block their learning from the tests.

On a different note, the following are students' suggestions in developing their preferred tests. These might help in identifying directions for developing multiple-choice tests.

1. Clear bilingual rubrics with illustrations or examples should be provided.
2. Test takers should be informed of their scores and mistakes right after the tests.
3. There should be four or five options in multiple-choice questions.
4. Lengthy texts should be reduced and the tests should concentrate more on quality of test items and appropriate format/ method.
5. Tests should be used to assess cognitive ability, not ephemeral memorization of knowledge.
6. There should be various levels of difficulty in each test.
7. The difficulty levels of the tests should be in accordance with test-takers' ability or proficiency.
8. Some kinds of supports (such as a glossary) should be provided while taking the tests.

3.2 Study 2: Students' views on a pilot study of reading test

As a preliminary study before developing the SATs focusing on reading, 6 reading passages with 5-multiple-choice questions for each were piloted with 9 students. Introspective interviews were undertaken right after the test to examine how attempted to answer the questions and how the test was perceived. In this part, students' views towards the reading test are reported.

Regarding the difficulty level of the texts, this set of reading texts consisted of varied levels of text difficulty which depended on topic familiarity, vocabulary used and types of questions (i.e. identifying the referent, drawing interpretations and extending beyond the text).

In terms of the difficulties they experienced while reading, tackling unknown vocabulary was highlighted especially where the topic itself was unfamiliar (i.e. friction, and Murphy's Law),

with their lack of background knowledge affecting their comprehension of the passages. In addition, lengthy texts (especially those of more than 500 words) were viewed as being necessarily more difficult. Last but not least, the ease with which they could interpret the test, both the reading passages and the test questions, was considered a key cause of their reading difficulty.

With regard to kinds of support needed while reading, two main kinds of supports were mentioned. The first one is a glossary; providing definitions or explanations of problematic vocabulary items either in Thai or in English could assist students' reading comprehension, particularly for low-frequency vocabulary in texts on unfamiliar topics. The second support needed is to provide hints, for example by highlighting key information in the text or giving examples to clarify the questions where possible.

These two investigations of students' views on language tests can help identify directions for developing multiple-choice tests which would both address the concerns of students with present-day testing practices and provide opportunities for learning from test, and thus the responses from stduents in these two studies were taken into consideration in the new approach to language testing of SATs.

4. A possible solution: Support Adaptive Tests (SATs)

As a solution to challenging multiple-choice testing, an innovative method of testing by implementing the principles of scaffolding in a computer adaptive test, namely, support adaptive tests (SATs), is proposed. In SATs, an incorrect response results in the same item being presented again but with support provided, either through *text adaptation* (e.g. highlighting the relevant part of the text) or *item adaptation* (e.g. reducing the number of options in multiple-choice). For instance, if a 4-option multiple-choice item for reading is answered incorrectly, it is repeated either with only 3 options (item adaptation) or with the relevant parts of the test highlighted (text adaptation).

Vygotskyan or sociocultural approaches which emphasis the role of scaffolds in learning have had a large impact on the teaching of English as a second or foreign language in the last twenty years (Gibbons, 2002). However, the scaffolding principles have not influenced testing or assessment to the same extent. Therefore, it is interesting to attempt to apply the principles of scaffolding to testing. Applying Vygotskyan scaffolding into SATs could provide two main advantages over current testing approaches. First, a test which includes scaffolds can serve learning as well as assessment purposes. Second, SATs could overcome the problems of Computer Adaptive Tests (CATs) while retaining their benefits. CATs are individualised in that they vary the level of difficulty of succeeding items based on previous responses with the benefit that test-takers do not waste time on items which are too difficult but with the problem that objectives covered may not be the same across test-takers (Chapelle and Douglas, 2006). The scaffolds in SATs also ensure that test-takers do not encounter excessively difficult items but that all test-takers are assessed on the same objectives.

In addition to these two major benefits, various types of validity could be manifested in SATS. First, learning validity could be promoted since learning is faciliated during testing. SATs could provide clear opportunities for learning since scores are informed immediately after the tests, various kinds of support are provided promptly if items are answered incorrectly, and the tests promote cognitive ability, not ephemeral memorization. Consequently, positive washback or the extent to which the tests drive learning (Bailey, 1999) would also be exhibited. Furthermore, SATs are more authentic, meaning that their method validity improves since it is clear how applicable results from the test are to authentic situations. SATs may also reflect teaching better than other forms of tests, since in most teaching contexts teachers provide support for students to answer questions. It would therefore seem that SAT multiple-choice reading tests promote several aspects of validity.

The students' perspectives on testing from the two studies reported above were used as guidelines for SATs design and provided a justification for the development of SATs. Three main characteristics which were salient in the students' responses and which are applicable for SATs design are promoting learning as well as assessing English ability at the same time, providing some kinds of supports or scaffolds to facilitate answering the test items (i.e. text and item adaptation), and informing immediate scores or feedbacks to test-takers.

5. Directions for future development of SATs

Since the SATs design is in its preliminary stages and because there has been little previous work on support adaptive tests (SATs) especially as applied to language teaching, a series of pilot studies that each aim to inform the appropriate design of SATs are needed. The first two pilot studies to inform the SATs design were reported above. Two more studies on reading will be conducted with a focus on text and item adaptation as scaffolds in SATs. It is hoped that the results from those investigations can provide principles which can be used in the design of a full-scale SAT.

5.1 Piloting plan for Text and Item Adaptation

For Studies 3 and 4, each piloting (one for Text Adaptation, one for Item Adaptation) consists of 4 texts each with 5 items (i.e. 4 sets). Each set focuses on a different type of adaptation. All items are set at a level where we might expect test-takers to score 40-50%. Piloting will be conducted with around 50 test-takers in Thailand. To ensure that the piloting results are not specific to Thai students, parallel mini-piloting will be conducted in Vietnam. The following is the provisional plan of piloting.

Piloting for Text Adaptation

Qs 1-5	5 multiple-choice questions, repeat incorrect items with a simplified text
Qs 6-10	5 multiple-choice items, repeat incorrect items with relevant parts of text highlighted
Qs 11-15	5 multiple-choice items, repeat incorrect items with added headings, pictures and diagrams in text
Qs 16-20	5 multiple-choice items, repeat incorrect items with added glosses and paraphrases of difficult points in text

Piloting for Item Adaptation

Qs 1-5	5 short-answer questions, repeat incorrect items as multiple-choice items
Qs 6-10	5 short-answer questions with one-word answers, repeat incorrect items with first letter of answer inserted
Qs 11-15	5 multiple-choice questions, repeat incorrect items with number of options reduced from 4 to 3
Qs 16-20	5 multiple-choice questions, repeat incorrect items after a video demonstration in Thai (Vietnamese) of how to answer a parallel question

5.2 Data analysis

Item analysis will be conducted for each set to see the increase in scores after scaffolds are provided. Ideally, the proportion of items answered correctly would improve from 40-50% to 70-80%. Also, individual item analysis will be conducted to see how scores on each item are affected by the scaffold (especially to see if an individual item has an undue effect on the set score). In addition, 4 test-takers per piloting will engage in think-aloud introspection while taking the test.

6. Conclusion

This paper presents a preliminary investigation into an innovative approach to testing: Support Adaptive Tests (SATs). Integrating scaffolding and Computer Adaptive Testing, SATs involve varying the amount of support or scaffolds (through text or item adaptation) available depending on whether an item is answered correctly or incorrectly. SATs could provide an opportunity for learning as well as testing language at the same time and should promote positive washback. If SATs prove to be feasible and of potential benefit, they can become parts of the repertoire of test approaches for test designers who wish to incorporate their advantages into their tests.

Acknowledgements

This paper is part of a larger project to develop Support Adaptive Tests funded by the School of Liberal Arts, King Mongkut's University of Technology Thonburi (Project no. 2555102) and by the British Council Access English Research Partnership.

References

Bachman, L. F. (2004) *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bailey, K. M. (1999) *Washback in Language Testing.* (TOEFL Monograph Series, MS 15). Princeton, NJ: Educational Testing Service.

Brown, J. D. (2005) *Testing in Language Programs*. Singapore: McGraw-Hill.

Burke, K. (1999) *How to Assess Authentic Learning*. Arlington Heights, IL: Skylight Professional Development.

Chalhoub-Deville, M. (1999) *Studies in Language Testing 10: Issues in Computer-Adaptive Testing of Reading Proficiency*. Cambridge: Cambridge University Pres.

Chapelle, C. A. & Douglas, D. (2006) *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press.

Darasawang, P. & Watson Todd, R. (2012) The effect of policy on English language teaching at secondary schools in Thailand. In Low, E.-L. and Hashim, A. (eds.) *English in Southeast Asia: Features, Policy and Language in Use* (pp. 207-220). Amsterdam: John Benjamins.

Forsyth, I., Jolliffe, A. & Stevens, D. (1999) *Planning a Course*, 2nd edition. London: Kogan Page.

Fulcher, G. (2003) *Testing Second Language Speaking*. Harlow: Pearson.

Fullan, M. (1987) Managing curriculum change. In Preedy, M. (ed.) *Approaches to Curriculum Management* (pp. 144-149). Milton Keynes: Open University Press.

Gibbons, P. (2002) *Scaffolding Langugae Scaffolding Learning.* Richmond: Heinemann.

Khalifa, H. & Weir, C. J. (2007) *Studies in Language Testing 29: Examining Reading*. Cambridge: Cambridge University Press.

McNamara, T. & Roever, C. (2006) *Language Testing: The Social Dimension*. Oxford: Blackwell.

Ockey, G. J. (2009) Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal* 93, focus issue, 836-847.

Piboonkanarax, K. (2007) *A Survey of Secondary School Evaluation Procedures Focusing on Continuous Assessment*. Unpublished MA thesis, King Mongkut's University of Technology Thonburi, Bangkok.

Pino-Silva, J. (2008) Student perceptions of computerized tests. *ELT Journal* 62(2), 148-156.

Thongsri, M., Charumanee, N. and Chatupote, M. (2006) The implementation of 2001 English language curriculum in government secondary schools in Songkhla. *ThaiTESOL Bulletin* 19(1), 60-94.

Tomlinson, B. (2005) Testing to learn: a personal view of language testing. *ELT Journal* 59(1), 39-46.

Watson Todd, R. & Shih, C.-M. (forthcoming) Assessing English in South East Asia. In Kunnan, A. (ed.) *Companion to Language Assessment*. Wiley-Blackwell.

Watson Todd, R. (2007) Encourage real scholarship. *The Bangkok Post* 23rd September 2007.

Watson Todd, R. (2008) The impact of evaluation on Thai ELT. *Selected Proceedings of the 12th English in South – East Asia International Conference: Trends and Directions* (pp. 118 – 127). Bangkok.

Weir, C. J. (2005) *Language Testing and Validation*. Basingstoke: Palgrave Macmillan.