Support Adaptive Testing: The effects of scaffolds in computer-based tests

Richard Watson Todd (irictodd@kmutt.ac.th)
King Mongkut's University of Technology, Thailand

Abstract
This paper presents an investigation of the viability of an innovative approach to computing-based multiple-choice adaptive testing, called Support Adaptive Testing, focusing on the effectiveness of scaffolds in testing. Support Adaptive Tests (SATs) are based on the principles of assessment for learning, scaffolding, and learner autonomy. The SATs under investigation initially appear as traditional computer-based multiple-choice tests of reading, but items answered incorrectly are repeated with scaffolds and explicit feedback is given at the end of the test. The SATs were taken by 276 Thai and 121 Vietnamese undergraduate students, and data concerning perceptions, responses and time collected. Results show that the students were positive about the SATs, Test Highlighting was the most preferred and most used scaffold, the format where a predetermined appropriate scaffold was provided proved most effective especially for Vietnamese students, and the scaffolds differed in effectiveness for different reading micro-skills.

INTRODUCTION

Assessment, especially in the form of tests, has a major impact on language learning, particularly in formal learning situations. Often, unfortunately, the impacts of assessment on learning are largely negative. With the rapid recent developments in technology, it might be hoped that computer applications in testing could ameliorate some of the negative impacts. Some attempts have been made in this direction, notably in the design and implementation of innovative computer-based item types (see e.g. Parshall, Davey, and Pashley, 2005), yet most computer-based testing is still multiple-choice (Boyle, Hutchinson, O'Hare, and Patterson, 2002; Conole and Warburton, 2005). This paper presents the results of the implementation of an innovative approach to multiple-choice computer-based tests called Support Adaptive Tests (SATs).

MULTIPLE-CHOICE TESTING AND COMPUTER-BASED TESTING

Multiple-choice testing (MCT) is the norm in many educational contexts. For English language testing, this is especially the case in countries where English is a foreign language (EFL). For instance, in the EFL countries of south-east Asia, MCT dominates testing (see Watson Todd, 2012; Watson Todd and Shih, 2013). In Thailand, over half of secondary school marks in English are derived from MCT (Piboonkanarax, 2007) and all high-stakes national-level exams are purely multiple-choice; in Vietnam, only one of the numerous national-level exams contains a component that does not rely on MCT. Unfortunately, this reliance on MCT, although enabling large numbers of students to be tested with few practical problems, has numerous negative washback effects on learning, including teaching being restricted to the receptive skills especially reading and to knowledge of grammar and vocabulary, the promotion of rote learning, and little focus on higher-order thinking (Watson Todd, 2008; Brown, 2005; Brown, Bull, and Pendlebury, 1997; Burke, 1999).

The move from paper-based to computer-based testing has perhaps increased the prevalence of MCT, since multiple-choice items are one of the easiest to present on computer (Roever, 2001) and the practicality benefits of using MCT are enhanced with automated marking. Even some computer-based testing innovations, such as computer-adaptive testing (CAT), are largely reliant on MCT. CAT is "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), with the aim of ensuring that the level of difficulty of the test items reflects the level of the test-taker (Ockey, 2009). Whatever the benefits of CAT, its reliance on structured-response formats, especially MCT, means that contexts where CAT is used are likely to suffer from the same pernicious washback effects as those where paper-based MCT dominates (Dunkel, 1999).

While CAT is unlikely to have positive washback effects, a different type of adaptive testing does promote learning. In CAT, an incorrect response to one item leads to the presentation of a new item (at an easier level) without the test-taker being aware that their response was incorrect. In what I will call Support Adaptive Testing, an incorrect response leads to the same item being repeated with a hint to help the test-taker. Previously applied primarily in testing mathematics and computer programming skills (Conejo, Guzmán, de-la Cruz, and Millán, 2006; Hu and Law, 2008; Wang, 2011), the hints provided can be viewed as scaffolds which increase the likelihood of the test-taker identifying the correct answer. If combined with further feedback on those items still answered incorrectly, Support Adaptive Testing enables MCT to provide opportunities for learning, both from the scaffolds and from the feedback.

PRINCIPLES UNDERLYING SPPORT ADAPTIVE TESTING

Since Support Adaptive Tests (SATs) provide opportunities for learning, they can fulfill the goals of both assessment of learning and assessment for learning. It has been argued that the traditional disjunction between assessment and learning is harmful for learning, and thus that there needs to be a move from this traditional model of assessment of learning to a new model of assessment for learning (Paran, 2010), which research has shown leads to substantial gains in student learning (Black and Wiliam, 1998). In EFL contexts, attempts to implement assessment for learning have produced mixed results, largely due to the effects of systemic inertia (Lee and Coniam, 2013). Implementing assessment for learning through an adaptation of a widely accepted testing procedure, such as MCT, is in line with existing practices in countries like Thailand and Vietnam and is therefore less likely to run into systemic resistance and so may be more successful.

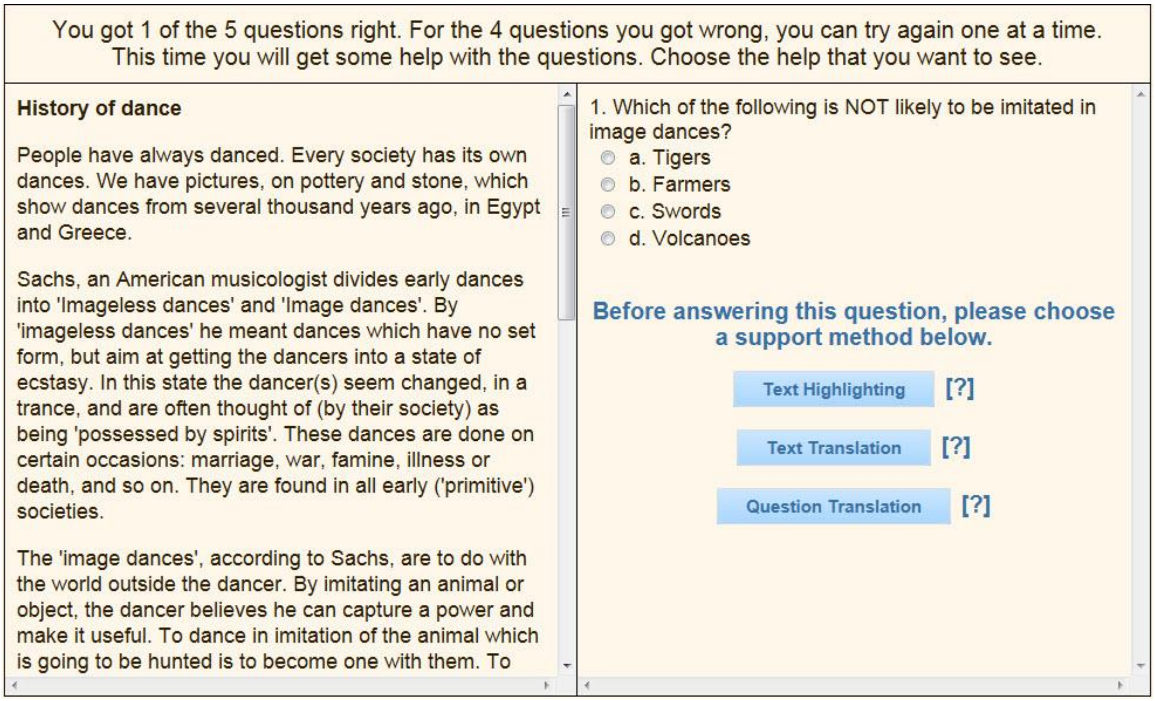
In SATs, the hints given in the repetition of an item answered incorrectly make the item easier to answer on the basis that the reason the item was answered incorrectly initially is that it was too difficult. A well-designed test consists of items around the test-takers' level of proficiency so that prior learning can influence scores (Khalifa and Weir, 2009), meaning that for a given test-taker some items are at or just below their level (and thus answered correctly) and some just above their level (or just too difficult and thus answered incorrectly). These latter items are in the test-taker's zone of proximal development, or the area just above independent problem solving ability where problems can be solved under guidance (Vygotsky, 1978). In SATs, this guidance comes in the form of hints or scaffolds.

Scaffolds are "the precise help that enables a learner to achieve a specific goal that would not be possible without some kind of support" (Sharpe, 2006: 212). In assessment contexts, the provision of these scaffolds may involve dialogue between teacher and learner (termed interactionist dynamic assessment) or may be signals directing a learner's attention to the key features of the environment necessary to complete the task (termed interventionist dynamic assessment; see Poehner, 2005). In most formal computer-based assessment, dialogue with a teacher is not possible; the scaffolds in SATs then are signals performing the function of mediation (Teo, 2012).

One problem with such automated interventionist dynamic assessment is that there is no ideal scaffold for a given item (although some may generally be better than others), since different test-takers may require different scaffolds for the same item (Lantolf, 2011). To overcome this problem, SATs allow test-takers to choose from a range of scaffolds allowing an element of learner autonomy, which involves offering choices and opportunities for decision making and supporting learners (Benson, 2003), to be incorporated on the basis that this is most likely to provide opportunities for learning.

THE DESIGN OF SATS

The SATs under investigation in this paper initially appear to be typical multiple-choice computer-based tests of reading with a text in the left half of the screen and five multiple-choice questions on the right. After test-takers have answered the five questions, they are informed how many of their answers were correct and incorrect and told that they will be given a second chance to answer the incorrect items with support. The text then re-appears together with the incorrect items presented individually. For each of these, test-takers can choose from three possible scaffolds as shown in Figure 1 (Text Highlighting: Figure 2, Text Translation: Figure 3 and Question Translation: Figure 4). Clicking on one of these either highlights the relevant part of the text or provides a translation of words likely to be problematic, after which the test-taker can answer the item again. On finishing the second round of answering, test-takers receive their combined score together with explanations of any items answered incorrectly twice (see Figure 5). The SATs used can be found at http://arts.kmutt.ac.th/crs/sat/ with the third test encountered being the full version of SATs.



You got 1 of the 5 questions right. For the 4 questions you got wrong, you can try again one at a time. This time you will get some help with the questions. Choose the help that you want to see.

History of dance

People have always danced. Every society has its own dances. We have pictures, on pottery and stone, which show dances from several thousand years ago, in Egypt and Greece.

Sachs, an American musicologist divides early dances into 'Imageless dances' and 'Image dances'. By 'imageless dances' he meant dances which have no set form, but aim at getting the dancers into a state of ecstasy. In this state the dancer(s) seem changed, in a trance, and are often thought of (by their society) as being 'possessed by spirits'. These dances are done on certain occasions: marriage, war, famine, illness or death, and so on. They are found in all early ('primitive') societies.

The 'image dances', according to Sachs, are to do with the world outside the dancer. By imitating an animal or object, the dancer believes he can capture a power and make it useful. To dance in imitation of the animal which is going to be hunted is to become one with them. To

1. Which of the following is NOT likely to be imitated in image dances?

○ a. Tigers

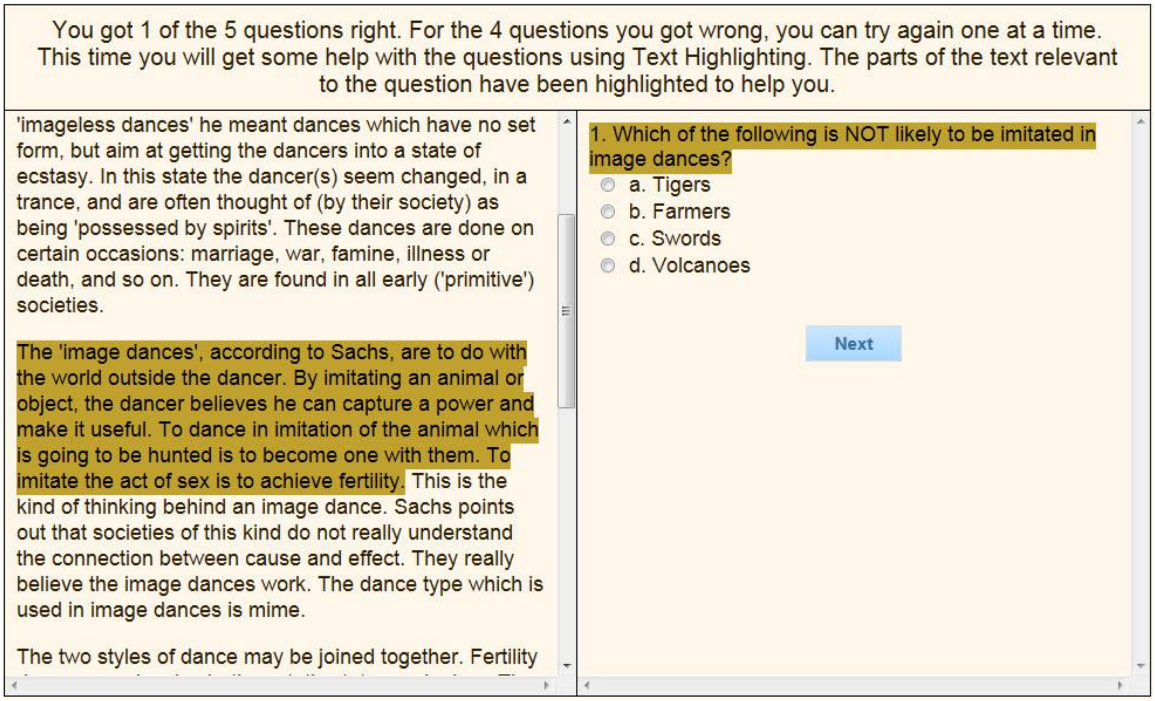
○ b. Farmers

○ c. Swords

○ d. Volcanoes

Before answering this question, please choose a support method below.

Figure 1. A repeated item in Format C (see below) of SATs



You got 1 of the 5 questions right. For the 4 questions you got wrong, you can try again one at a time. This time you will get some help with the questions using Text Highlighting. The parts of the text relevant to the question have been highlighted to help you.

'imageless dances' he meant dances which have no set form, but aim at getting the dancers into a state of ecstasy. In this state the dancer(s) seem changed, in a trance, and are often thought of (by their society) as being 'possessed by spirits'. These dances are done on certain occasions: marriage, war, famine, illness or death, and so on. They are found in all early ('primitive') societies.

The 'image dances', according to Sachs, are to do with the world outside the dancer. By imitating an animal or object, the dancer believes he can capture a power and make it useful. To dance in imitation of the animal which is going to be hunted is to become one with them. To imitate the act of sex is to achieve fertility. This is the kind of thinking behind an image dance. Sachs points out that societies of this kind do not really understand the connection between cause and effect. They really believe the image dances work. The dance type which is used in image dances is mime.

The two styles of dance may be joined together. Fertility

1. Which of the following is NOT likely to be imitated in image dances?

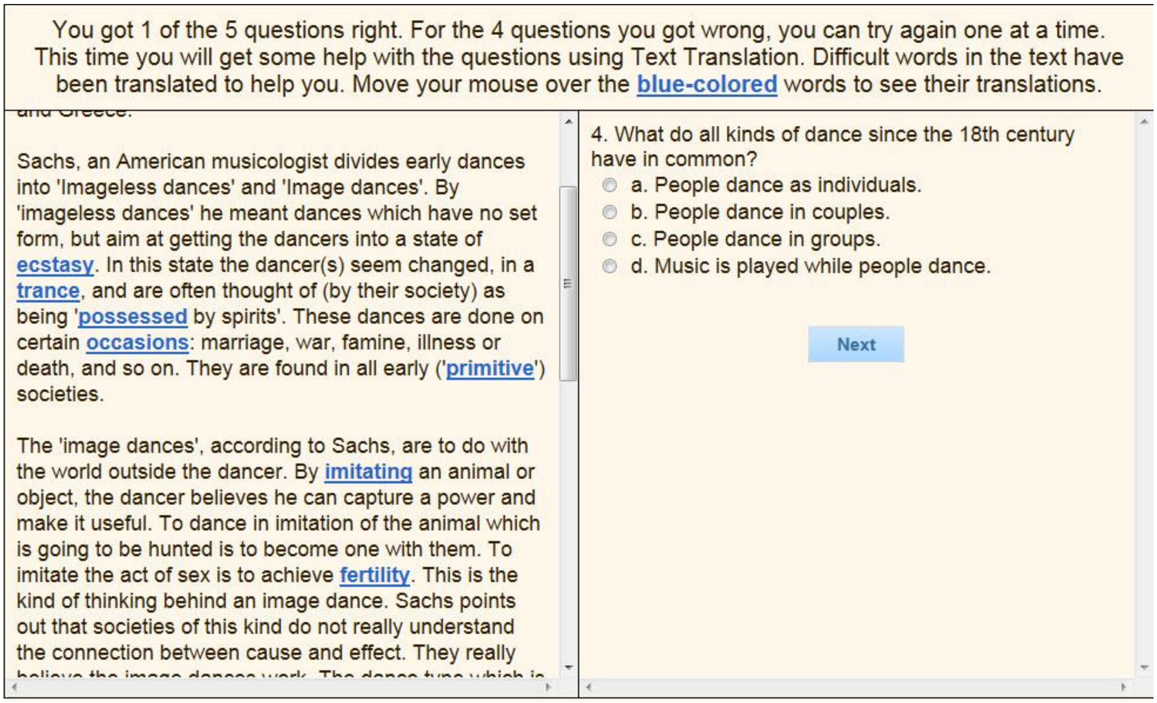
○ a. Tigers

○ b. Farmers

○ c. Swords

○ d. Volcanoes

Figure 2. Text Highlighting scaffold in SATs



You got 1 of the 5 questions right. For the 4 questions you got wrong, you can try again one at a time. This time you will get some help with the questions using Text Translation. Difficult words in the text have been translated to help you. Move your mouse over the blue-colored words to see their translations.

and Greece.

Sachs, an American musicologist divides early dances into 'Imageless dances' and 'Image dances'. By 'imageless dances' he meant dances which have no set form, but aim at getting the dancers into a state of ecstasy. In this state the dancer(s) seem changed, in a trance, and are often thought of (by their society) as being 'possessed by spirits'. These dances are done on certain occasions: marriage, war, famine, illness or death, and so on. They are found in all early ('primitive') societies.

The 'image dances', according to Sachs, are to do with the world outside the dancer. By imitating an animal or object, the dancer believes he can capture a power and make it useful. To dance in imitation of the animal which is going to be hunted is to become one with them. To imitate the act of sex is to achieve fertility. This is the kind of thinking behind an image dance. Sachs points out that societies of this kind do not really understand the connection between cause and effect. They really believe the image dances work. The dance type which is

4. What do all kinds of dance since the 18th century have in common?

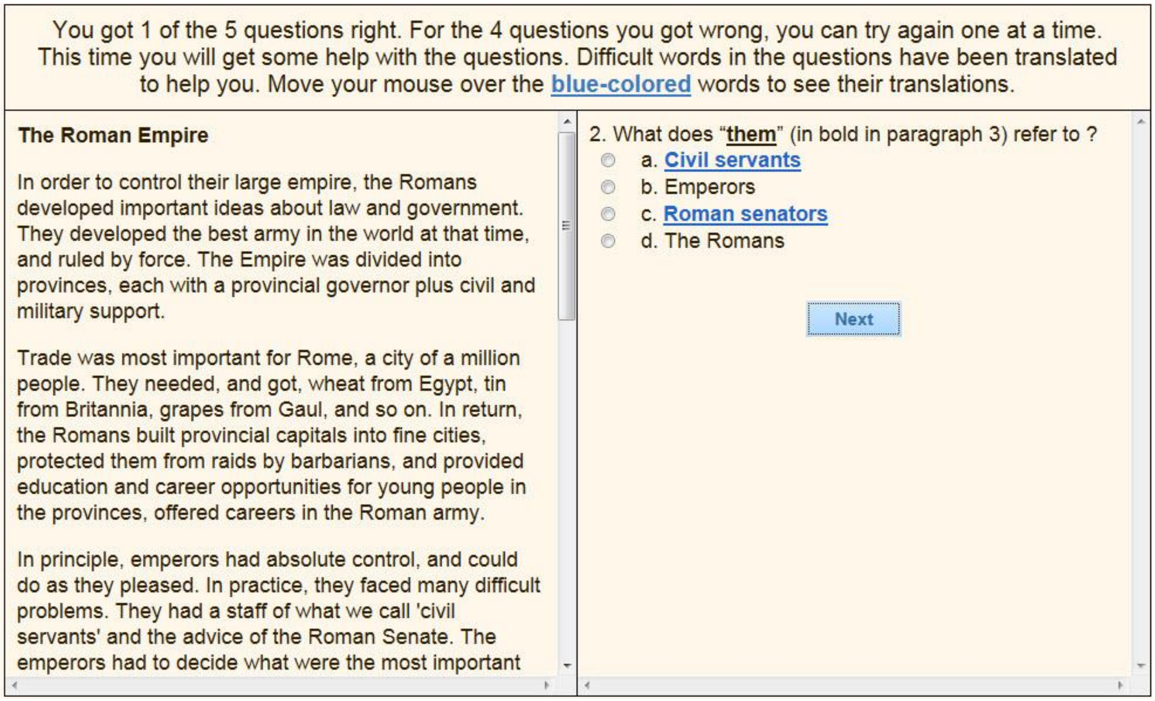
○ a. People dance as individuals.

○ b. People dance in couples.

○ c. People dance in groups.

○ d. Music is played while people dance.

Figure 3. Text Translation scaffold in SATs



You got 1 of the 5 questions right. For the 4 questions you got wrong, you can try again one at a time. This time you will get some help with the questions. Difficult words in the questions have been translated to help you. Move your mouse over the blue-colored words to see their translations.

The Roman Empire

In order to control their large empire, the Romans developed important ideas about law and government. They developed the best army in the world at that time, and ruled by force. The Empire was divided into provinces, each with a provincial governor plus civil and military support.

Trade was most important for Rome, a city of a million people. They needed, and got, wheat from Egypt, tin from Britannia, grapes from Gaul, and so on. In return, the Romans built provincial capitals into fine cities, protected them from raids by barbarians, and provided education and career opportunities for young people in the provinces, offered careers in the Roman army.

In principle, emperors had absolute control, and could do as they pleased. In practice, they faced many difficult problems. They had a staff of what we call 'civil servants' and the advice of the Roman Senate. The emperors had to decide what were the most important

2. What does "**them**" (in bold in paragraph 3) refer to ?

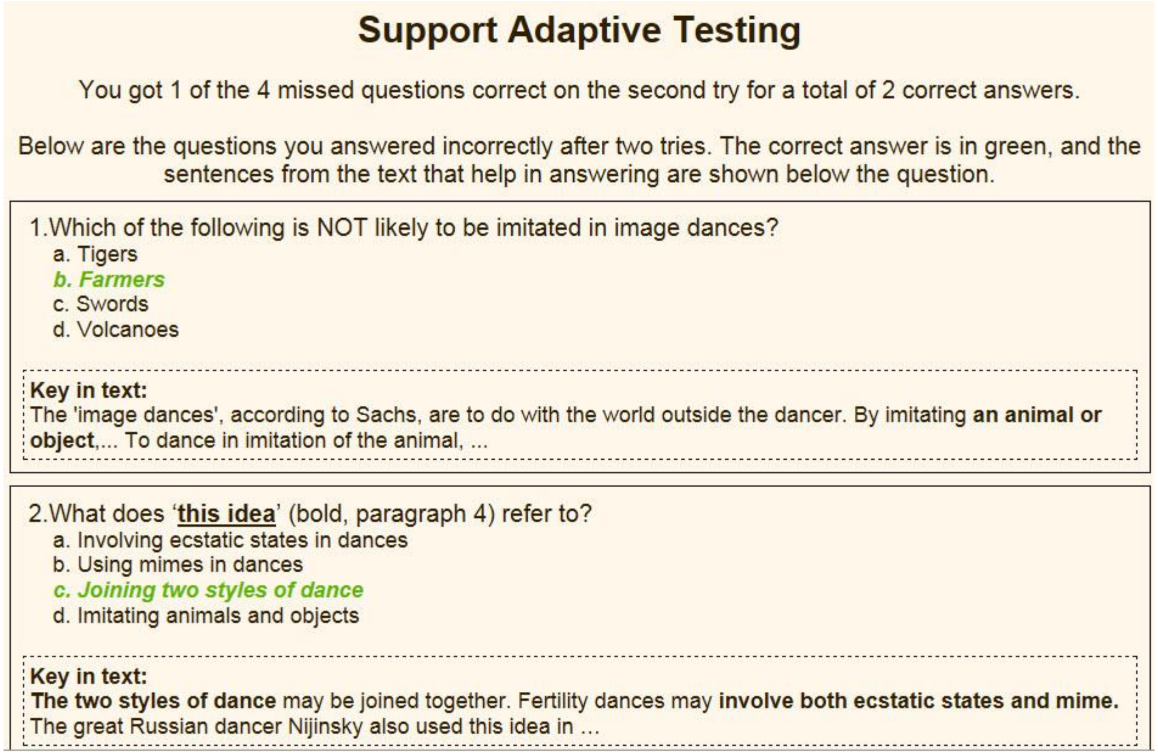
○ a. Civil servants

○ b. Emperors

○ c. Roman senators

○ d. The Romans

Figure 4. Question Translation scaffold in SATs



Support Adaptive Testing

You got 1 of the 4 missed questions correct on the second try for a total of 2 correct answers.

Below are the questions you answered incorrectly after two tries. The correct answer is in green, and the sentences from the text that help in answering are shown below the question.

1.Which of the following is NOT likely to be imitated in image dances?
a. Tigers
b. Farmers
c. Swords
d. Volcanoes

Key in text:
The 'image dances', according to Sachs, are to do with the world outside the dancer. By imitating **an animal or object**,... To dance in imitation of the animal, ...

2.What does '**this idea**' (bold, paragraph 4) refer to?
a. Involving ecstatic states in dances
b. Using mimes in dances
c. Joining two styles of dance
d. Imitating animals and objects

Key in text:
The two styles of dance may be joined together. Fertility dances may **involve both ecstatic states and mime.** The great Russian dancer Nijinsky also used this idea in ...

Figure 5. Feedback at the end of SATs

The SATs are designed based on the theoretical foundations discussed above. By starting from a typical multiple-choice test of the type prevalent in the contexts investigated, it is hoped that SATs are more likely to be accepted as they fit with current educational practices. The tests promote assessment for learning, hopefully reducing the negative washback of current testing practices, through the use of scaffolds and through the incorporation of test-taker autonomy by allowing choices. They also allow the extent to which test-takers are able to use feedback to be gauged. In addition, the design of SATs was also based on data collected through the process of test design.

THE PROCESS OF DESIGNING SATS

Test design needs to meet the needs of the stakeholders involved, but it is rare for the views of perhaps the most important group of stakeholders, the test-takers themselves, to be taken into account (Paran, 2010; Pino-Silva, 2008). The SATs test design therefore started with an open-ended survey focusing on perceptions of testing of 323 Thai university students (see Watson Todd and Jaturapitakkul, 2013 for details). Key points emerging from the survey that influenced SATs design were suggestions for improving current testing practices:

- Test rubrics and instructions should be clear and bilingual (in both Thai and English) with illustrations or examples to avoid misunderstandings.
- Test-takers should be informed of their scores and what they got incorrect immediately after the test.

- Reading passages in tests should not be excessively lengthy.
- Tests should cover a range of levels of difficulty commensurate with the various levels of proficiency of the students.
- Help, such as a glossary, should be provided in the test so that the test would assess their abilities, not their knowledge of a particular word.

These points were taken into account in the SATs design.

For reasons of practicality, it was decided that the pilot version of SATs should focus on reading. We therefore needed to identify the characteristics of appropriate texts, appropriate reading micro-skills to test, appropriate scaffolds for reading, and an appropriate test interface.

To identify appropriate reading tests, nine students took six multiple-choice reading tests that reflected current testing practices and were interviewed. The interview data showed that texts should have a length of 350-550 words and Flesch reading ease scores of 60-75, and should not be too technical. An open source fitting these criteria is Simple Wikipedia (http://simple.wikipedia.org/wiki/Main_Page), and texts on topics where the information in the texts was likely to be unfamiliar to test-takers including the history of dance, the atmosphere and paper were selected. In the interviews, the students also stated that tests could be more effective if two types of support were provided with the text: a glossary providing definitions or translations of problematic words especially low-frequency vocabulary on unfamiliar topics, and hints such as highlighting key information in texts or giving examples to clarify questions, providing further support for the design of SATs and indicating potential scaffolds.

Appropriate reading micro-skills were selected from a comprehensive list based on the literature (Alderson, 2000; Ediger, 2006; Grabe and Stoller, 2002; McDonough, 1995; Nunan, 1989; Urquhart and Weir, 1998). The testable micro-skills on the list were compared against existing tests and course objectives to identify five micro-skills to test:

- Identifying the referent of a referring expression (Reference)
- Identifying and interpreting specific information in the text (Specific information)
- Drawing an inference from specific information in the text (Specific inference)
- Synthesizing information from several parts of the text to provide a sequence, such as stages in a process or events in history (Synthesis sequence)
- Comparing or synthesizing information from several points in the text, often in the form of a NOT question (e.g. Which of the following is NOT mentioned ...) (Information synthesis)

For scaffolds, the students' suggestions in the survey and interviews were compared to the literature on text simplification and elaboration (e.g., Lorenzo, 2008; Nation, 2001; O'Donnell, 2009; Simard, 2009) to identify the three types of scaffolds used in SATs:

- Text Highlighting where the portions of the text relevant to a particular question would be highlighted with easy-to-see colour shading.

- Text Translation where words or phrases in the text likely to be unknown to students (usually because of their low frequency) would be glossed with an L1 translation. These words or phrases would be presented in a different colour font with the gloss becoming apparent on mouseover.
- Question Translation where words or phrases in the questions likely to be unknown to students would be glossed with an L1 translation in the same way as for Text Translation.

The web-based test interface was designed by analysing the top ten websites as returned by *Google* for reading tests and using http://web2.uvcs.uvic.ca/elc/studyzone/410/reading/deathcar.htm as a model.

PILOTING THE SATS

Two pilot studies into SATs were conducted. The first aimed to examine the face validity of SATs or the general acceptability as a form of testing (Perry, 2005). The SATs were exhibited at the Thailand National Research Exhibition, and 118 members of the general public took the test and completed a questionnaire. The responses showed an overwhelming (98.29% of respondents) perception that SATs were an appropriate way to test English reading ability, with Text Highlighting the preferred scaffold (64.56% of respondents). Most of the respondents viewed SATs as appropriate for all levels of education, with a few arguing that SATs were most appropriate as mock tests (2 respondents) or for self-study (4 respondents).

The second piloting was conducted for test validation and involved 68 Thai university students taking versions of SATs, each with a single scaffold type (to ensure sufficient data for each type of scaffold). Item facility and item discrimination were calculated and resulted in several items being rewritten. One unexpected result concerned the Question Translation scaffold. Where the translation only applied to words in one of the options, there was evidence that the students were choosing this option at a greater than expected rate even when it was incorrect, suggesting the use of testwise strategies (Cohen, 1994). To avoid this, additional translations were added to other options even though the words were believed to be familiar.

PURPOSE OF THE STUDY AND METHODOLOGY

This paper aims to investigate test-takers' perceptions of SATs and their effectiveness, specifically focusing on the usefulness of the scaffolds. To see the effectiveness of the scaffolds, the SATs need to be compared with tests which allow two rounds of answering but provide no scaffolds. In addition, to see if test-takers can choose the most appropriate scaffold, the SATs need to be compared with tests where an appropriate scaffold is provided for each item. In this study, therefore, three versions of tests were prepared:

Format A: Students complete the test. Any items answered incorrectly are repeated (with no support given).

Format B: Students complete the test. Any items answered incorrectly are repeated with a predetermined type of scaffold (one of TH, TT or QT). This scaffold is the one identified as most appropriate based on a combination of analysing the test and the results of piloting.

Format C: Students complete the test. Any questions answered incorrectly are repeated with a choice of three scaffolds (TH, TT and QT). Students choose the scaffold they want for each item and then answer after the scaffold has been provided.

After completing the second round of each test, the test-takers were given feedback with explanations for incorrect items in all three formats.

The tests in this study, therefore, consisted of six reading passages, each with five four-option multiple-choice items in three formats. Each subject took three tests, one in each format selected at random from a database. After completing the tests, a short questionnaire asking whether the scaffolds were helpful and which type of scaffold helped the test-takers understand the text and answer the questions was presented. The tests were taken by 276 Thai and 121 Vietnamese non-English major undergraduate students at respected universities recruited as intact classes for reasons of practicality. With the input into test design being based on Thai students only, the use of SATs needs to be compared between Thai students and a similar group from another country. The educational context in Vietnam is similar to that of Thailand, both English-as-a-foreign-language countries in southeast Asia where MCT dominates the educational system. The levels of proficiency of the two groups of subjects were similar.

The data collected consisted of:

- Test-takers' responses to the questionnaire
- Frequency of choice of the three types of scaffold in Format C
- Frequency of correct responses in round 1 and in the three formats
- Time taken to answer each item in round 1 and in the three formats

Using these data, the following questions can be answered:

Perceptions and use of scaffolds

1. How do test-takers perceive SATs?
2. How do test-takers perceive the three types of scaffold?
3. How frequently do test-takers choose each of the three types of scaffold?

Effectiveness of the scaffolds

4. How effective are the three formats in helping test-takers give correct responses? Are there differences in the level of effectiveness of the three formats?
5. How effective are the three types of scaffold in helping test-takers give correct responses?
6. Is there a relationship between test-takers' perceptions of the usefulness of the scaffold types and their effectiveness?

7. How effective are the three formats and the three types of scaffold in helping test-takers give correct responses to items testing each of the five micro-skills?

Comparing Thai and Vietnamese test-takers

8. Are there differences in the level of effectiveness of the three formats and the three types of scaffold between Thai and Vietnamese test-takers?

The majority of these questions can be answered by examining the percentages of responses. Questions 4 and 8 which look for differences are answered using the difference of proportions statistic with confidence intervals at 95% and using odds ratio to measure effect size (see Agresti, 2007). To answer Question 6 investigating the relationship between perceptions and effectiveness of the scaffolds, point biserial correlation between whether or not a particular scaffold was the type the test-taker stated a preference for and its effectiveness was calculated.

RESULTS

Test-taker's perceptions of SATs

In the questionnaire after taking the SAT, the test-takers were asked for their perceptions about SATs and the three types of scaffold. The findings are presented in Table 1.

Table 1

Test-takers' Perceptions of SATs

<table>
<tr><td rowspan="2">Q1: Was the support helpful?</td><td colspan="2">Yes (N)</td><td colspan="2">Yes (%)</td><td>No (N)</td><td>No (%)</td></tr>
<tr><td colspan="2">362</td><td colspan="2">96.02</td><td>15</td><td>3.98</td></tr>
<tr><td colspan="7"></td></tr>
<tr><td></td><td>TH (N)</td><td>TH (%)</td><td>TT (N)</td><td>TT (%)</td><td>QT (N)</td><td>QT (%)</td></tr>
<tr><td>Q2: Which kind of support helped you understand the essay better?</td><td>207</td><td>54.76</td><td>145</td><td>38.36</td><td>26</td><td>6.88</td></tr>
<tr><td>Q3: Which kind of support helped you answer the questions correctly?</td><td>241</td><td>63.93</td><td>95</td><td>25.20</td><td>41</td><td>10.87</td></tr>
</table>

From Table 1, we can see that the test-takers were overwhelmingly positive about the use of supports in SATs, with Text Highlighting being the preferred scaffold.

Test-takers' use of scaffolds

In Format A of data collection, test-takers were provided with no support; in Format B, the scaffold was predetermined; in Format C, test-takers could choose which type of scaffold they wanted for each question they had answered incorrectly. Examining the frequencies of choice of type of scaffold in Format C (shown in Table 2), then, allows us to see which type of scaffold the test-takers preferred to use.

Table 2
Frequency of Use of Types of Scaffold

	TH	TT	QT
Frequency of use	717	240	159
Percentage of use	64.25	21.50	14.25

Comparing Tables 1 and 2, we can see that the test-takers' choice of scaffold to use closely followed their preference for the type of scaffold that helped them answer the questions correctly with Text Highlighting being most frequent.

Effectiveness of scaffolds

To see whether the scaffolds in SATs were effective, we can compare the proportion of questions answered correctly in each format. Format A provides a benchmark for whether being given two opportunities to answer an item is helpful. If the test-takers remember their incorrect answer from the first round, they should have a one-in-three chance of answering correctly in the subsequent formats through randomly choosing a response. While Format A should give a greater than 33% correct response rate if answering twice is helpful, if the scaffolds are effective, we should expect correct response rates in Formats B and C to be higher than in Format A. For Format B, the most appropriate scaffold type was predetermined, but individual differences in test-taker preference may mean that the test-takers' own choice of scaffold type in Format C could be more effective. The correct response rates for the different formats, together with time taken to respond, are given in Table 3.

Table 3
Proportion of Correct Responses in Different Formats

	No. of items attempted	No. of correct responses	% of correct responses	Average time per question
Round 1 (no support)	5900	2512	42.58	01:49
Round 2: Format A (no support)	1107	393	35.50	00:36
Round 2: Format B (predetermined scaffold)	1136	471	41.46	00:47
Round 2: Format C (choice of scaffold)	1116	440	39.43	00:35[1]

[1] In addition, test-takers on average took 13 seconds to choose the type of scaffold

From Table 3, test-takers' responses in Format A were at about the level of chance, whereas there was a slight increase in correct responses for Formats B and C where support was provided with Format B being the most beneficial. Treating the number of correct and incorrect responses in Formats A, B, and C as independent binomial samples, we can calculate the difference of proportions, standard error and odds ratio between each pair of formats to see if the differences in proportions of correct responses are real.

Where the figures for the 95% confidence interval are either both positive or both negative, we can conclude that there is a significant difference in the proportions (see Agresti, 2007). These comparisons are shown in Table 4.

Table 4
Differences in Proportion of Correct Responses between Formats

	Format A v. Format B	Format A v. Format C	Format B v. Format C
Difference of proportions	-0.060	-0.039	0.020
Standard error	0.021	0.021	0.021
Odds ratio	1.287	1.183	0.919
95% confidence interval (max.)	-0.019	0.001	0.061
95% confidence interval (min.)	-0.100	-0.079	-0.020
Interpretation	Higher scores on Format B	No real difference	No real difference

Table 4 shows that the only real difference in the proportions of correct responses is between Formats A and B. We can conclude that Format B where the scaffold type provided is predetermined as the most appropriate is the method which best promotes higher test scores on a second round of answering.

In addition to comparing the different formats, we can also compare the effects of the three types of scaffold: Text Highlighting, Text translation and Question Translation. The effects of these in Formats B and C combined are shown in Table 5.

Table 5
Proportion of Correct Responses for Different Types of Scaffold

	No. of items attempted	No. of correct responses	% of correct responses	Average time per question
Text Highlighting	1114	425	38.15	00:38
Text Translation	629	265	42.13	00:35
Question Translation	509	221	43.42	00:29

From Table 5, Text Highlighting, the most preferred scaffold and the one most chosen in Format C, appears to be the least effective, although the differences between the three types of scaffold are minimal. Comparing the type of scaffold preferred in Question 3 of the questionnaire (see Table 1 above) with the effectiveness of this scaffold when used in Format C produces a point biserial correlation of -0.004 (not significant), implying that there is no relationship between test-takers' preferences for a type of scaffold and its effectiveness.

Effects of types of scaffold for different reading micro-skills tested

As described above, five reading micro-skills were tested in the SATs, namely, reference, finding specific information, making a specific inference, identifying a sequence through synthesizing points from different parts of the text, and synthesizing information. The percentages of correct responses for these five micro-skills in each format are given in Table 6.

Table 6
Percentage of Correct Responses for Different Reading Micro-skills Tested by Format

	Round 1 (no support)	Round 2: Format A (no support)	Round 2: Format B (predetermined scaffold)	Round 2: Format C (choice of scaffold)
Reference	48.64	42.35	40.95	40.61
Specific information	42.63	30.33	36.67	32.92
Specific inference	30.09	32.22	41.18	47.08
Synthesis sequence	55.89	41.12	52.94	44.55
Information synthesis	40.00	37.61	42.55	36.52

From Table 6, specific inference questions were the most difficult initially, but were also the type that most benefitted from the provision of scaffolds in Formats B and C. The scaffolds were also effective for synthesis sequence questions, the type that generally had the greatest proportion of correct responses, but the benefits of the scaffolds were less clear for the other three types of questions.

We can also investigate whether any particular type of scaffold was particularly effective for any particular reading micro-skill. The percentage of correct responses in Formats B and C combined for each type of scaffold and for each reading micro-skill are shown in Table 7. The figures in brackets give the total number of times this type of question with this type of scaffold was encountered.

Table 7
Percentage of Correct Responses for Different Reading Micro-skills Tested by Type of Scaffold

	Formats B and C combined		
	TH	TT	QT
Reference	42.15 (242)	40.41 (146)	26.32 (19)
Specific information	29.18 (353)	42.67 (150)	41.18 (119)
Specific inference	44.34 (221)	41.22 (131)	45.76 (177)
Synthesis sequence	57.89 (95)	53.03 (66)	32.35 (68)
Information synthesis	33.00 (203)	38.97 (136)	50.79 (126)

Surprisingly, the most frequent combination of reading micro-skill tested and scaffold type (Text Highlighting for specific information questions) was the least effective of the relatively frequent combinations, a point contrary to expectations since identifying the relevant part of the text should be especially helpful for questions focusing on specific information. Specific inference questions are the only reading micro-skill tested where the different types of scaffold provide consistent benefits; for the others, one type of scaffold is either particularly beneficial or notably unbeneficial, with Question Translation being the scaffold most likely to differ from the others.

Comparing Thai and Vietnamese test-takers

The test-takers who were subjects in this study came from two ASEAN countries where English is a foreign language: Thailand and Vietnam. This allows us to investigate whether there are any differences in the proportions of correct responses for the different factors in this study. Table 8 compares the proportion of correct responses for the different formats and the different types of scaffold between the two groups.

From Table 8, the Vietnamese test-takers generally produce slightly more correct responses than the Thai test-takers. To see if any of these differences in proportions of correct responses is noteworthy, the difference of proportions, standard error, odds ratio and difference at the 95% confidence interval were calculated for each of the formats and scaffold types. Table 9 summarizes these.

Table 8
Comparison of Thai and Vietnamese Test-takers for Different Formats and Different Types of Scaffold

	Thailand			Vietnam		
	No. of items attempted	No. of correct responses	% of correct responses	No. of items attempted	No. of correct responses	% of correct responses
Round 1 (no support)	4090	1716	41.96	1810	796	43.98
Round 2: Format A (no support)	770	268	34.81	337	125	37.09
Round 2: Format B (predetermined scaffold)	788	303	38.45	348	168	48.28
Round 2: Format C (choice of scaffold)	801	315	39.33	315	125	39.68
Formats B and C combined						
TH	801	298	37.20	313	127	40.58
TT	428	172	40.19	201	93	46.27
QT	360	148	41.11	149	73	48.99

Table 9
Differences between Thai and Vietnamese Test-takers for Different Formats and Different Types of Scaffold

	Round 1	Format A	Format B	Format C	TH	TT	QT
Difference of proportions	-0.020	-0.023	-0.098	-0.003	-0.034	-0.061	-0.079
Standard error	0.014	0.031	0.032	0.032	0.033	0.042	0.048
Odds ratio	1.086	1.104	1.494	1.015	1.153	1.282	1.376
95% confidence interval (max.)	0.007	0.039	-0.036	0.060	0.030	0.022	0.016
95% confidence interval (min.)	-0.048	-0.084	-0.161	-0.067	-0.098	-0.144	-0.173
Interpretation	No real difference	No real difference	Vietnamese higher than Thai	No real difference	No real difference	No real difference	No real difference

From Table 9, we can see that the only real difference between the Thai and Vietnamese test-takers was that the Vietnamese performed better on Format B of the SATs where the predetermined most appropriate scaffold was provided.

DISCUSSION

The key findings concerning SATs from this study are:

- Students find the SATs helpful.
- Students state a preference for Text Highlighting as a scaffold.
- Text Highlighting is the most frequently chosen scaffold.
- Format B (predetermined scaffold) is more effective than Format A (two rounds of answering but no support); Format C (choice of scaffold) is no more effective than the other formats.
- There are no clear differences in level of effectiveness between the different types of scaffold.
- There is no relationship between students' perceptions of the scaffold types and their effectiveness.
- The use of scaffolds appears to be most effective for items testing specific inference and synthesis sequence micro-skills; the most frequent combination of scaffold and micro-skill (Text Highlighting for specific information items) was not particularly effective.
- The effectiveness of Format B over other formats was particularly noticeable for the Vietnamese students.

The broad purpose of this paper is to investigate the viability of SATs as an alternative to traditional multiple-choice tests. The facts that the test-takers were overwhelmingly positive about SATs and that it is designed to counter students' concerns with current testing practices suggest that SATs should be a viable alternative. However, the findings about the effectiveness of the provision of scaffolds in tests are mixed. For some reading micro-skills, the scaffolds appear effective, but the full version of SATs (Format C) with test-taker choice of scaffold does not seem to be any more effective than simply providing two chances to answer items. Format B, on the other hand, where the most appropriate scaffold was predetermined removing the element of test-taker autonomy was more effective than either of the other versions, and this increased effectiveness was due to higher scores by the Vietnamese test-takers.

If we take Format B as the context where the highest possible scores are achievable since the most appropriate scaffold is provided, the differences between the Thai and Vietnamese test-takers on Formats B and C require explanation. The fact that the Thais scored about the same on the two formats suggests that they were able to choose appropriate scaffolds in Format C. For the Vietnamese, however, the difference in scores between Format B and Format C suggests that their choices of scaffold in Format C were not particularly appropriate. We might conclude then that the Thai test-takers are more effective in autonomous learning situations than the Vietnamese. This conclusion, however, conflicts with other sources. In a survey comparing students' attitudes towards learner autonomy in different countries, Littlewood (2000) found that Vietnamese students were consistently more positive about learner autonomy than Thais. Such attitudes may reflect a clear emphasis on learner development goals in English language curricula in Vietnam (Lap, 2005). It should be noted, however, that attitudes to autonomy do not necessarily lead to more effective autonomous decision making. The differences in performance on Formats B and C for Thai and Vietnamese test-takers could have implications for how SATs should be implemented in the two contexts.

Investigating the effectiveness of the scaffolds for helping test-takers to identify correct answers is not the only possible focus for research into SATs. Given the link between scaffolding and learning, and given the provision of feedback on incorrect responses, whether SATs lead to test-taker learning could be investigated. It should be noted, however, that SATs are designed as an alternative to formal testing, and are not designed to be used as a regular learning/assessment activity in the way that computerized dynamic assessment is (Teo, 2012), which means any learning from SATs is likely to be minimal given the short duration of the tests and to be limited to vocabulary items. SATs, however, are not designed primarily to be a learning instrument; rather, they are designed as a replacement for current testing practices which could lead to some changes in educational practices, most probably by reducing the negative washback effects of traditional multiple-choice testing. An alternative view of SATs is that they evaluate test-takers' ability to learn from feedback (in this study, most evident in the difference in number of correct responses between Formats A and B). A greater emphasis on the role of feedback in tests could have positive washback effects of encouraging a greater role for feedback throughout the teaching/learning process. The potential for SATs to reduce negative washback is an area requiring further research.

If SATs are to be implemented to reduce negative washback, it is somewhat unclear which version of SATs to use. For the Thai context (if we can generalize from the results of this study), given the theoretical arguments in its favor and the similarity of effectiveness of the different formats, it would appear that appear that Format C is most appropriate. For the Vietnamese context, however, the choice of most appropriate format is less clear. Format B is the most effective, and thus implementing Format B in Vietnamese contexts could be motivating. However, the problems of appropriacy of decision making in Format C for Vietnamese test-takers could mean that they would benefit more from greater exposure to such decision making. In other words, effectiveness in providing support to answer questions correctly might not be the most important criterion in choosing the format to use; rather, less tangible criteria which are difficult to measure, such as the effects of opportunities to make choices, could be more important in the long run.

CONCLUSION

Support Adaptive Tests, while not unequivocally effective in providing support for test-takers in all cases, appear to be a viable innovative alternative to traditional testing practices in English-as-a-foreign-language contexts where multiple-choice testing dominates. At worst, they are little different from current practices; at best, they take test-takers' subjective needs in testing (Brown, 1995) into account in test design, they provide some opportunities for learning through supported second chances at answering items and through explicit immediate feedback, and they hold the promise of reducing negative washback effects of testing. Given that innovations are more likely to be successful when they are congruent with current practices and student beliefs (Shamim, 1996), SATs are worthy of serious consideration as an addition to the existing range of computer-based 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. Natjiree Jaturapitakkul, Sonthida Keyuravong, Stuart Towns, Amin Dehgan, Nguyen Quang Tuyen and Worapong Kaewtong made substantial contributions essential to the success of this project.

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis*, 2nd edition. Hoboken, NJ: Wiley.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Benson, P. (2003). Learner autonomy in the classroom. In D. Nunan (Ed.), *Practical English language teaching* (pp. 289-308). Boston: McGraw-Hill.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice 5*(1), 7-74. doi: 10.1080/0969595980050102

Boyle, A., Hutchinson, D., O'Hare, D., & Patterson, A. (2002). Item selection and application in higher education. *Proceedings of the 6th International Computer Assessment Conference (CAA), 2002*. Loughborough, UK. pp. 269-284.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston: Heinle & Heinle.

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

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.

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 Press.

Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.

Conejo, R., Guzmán, E., de-la Cruz, J., & Millán, E. (2006). An empirical study about calibration of adaptive hints in web-based adaptive testing environments. In V. Wade, H. Ashman & B. Smyth (Eds.), *Adaptive hypermedia and adaptive web-based systems: Lecture notes in computer science 4018* (pp: 71-80). Berlin: Springer.

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *Association for Learning Technology Journal: Research in Learning Technology 13*(1), 17-31. doi: 10.1080/0968776042000339772

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology 2*(2), 77-93.

Ediger, A. M. (2006). Developing strategic L2 readers ... by reading for authentic purposes. In E. Uso-Juan & A. Martinez-Flor (Eds.), *Current trends in the development, and teaching of the four language skills* (pp. 303-328). Berlin: Mouton de Gruyter.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. London: Longman.

Hu, Q., & Law, N. (2008). Designing online peer assessment system in learning programming: An adaptive scaffolding framework and architecture. In J. Luca & E. Weippl (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2008* (pp. 1739-1744). Chesapeake, VA: AACE.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Studies in language testing 29*. Cambridge: Cambridge University Press.

Lantolf, J. P. (2011). The sociocultural approach to second language acquisition. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 24-47). London: Routledge.

Lap, T. Q. (2005). *Stimulating learner autonomy in English language education: A curriculum innovation study in a Vietnamese context*. PhD thesis, University of Amsterdam.

Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing 22*(1), 34-50. doi: 10.1016/j.jslw.2012.11.003

Littlewood, W. (2000). Do Asian students really want to listen and obey? *ELT Journal 54*(1), 31-36. doi: 10.1093/elt/54.1.31

Lorenzo, F. (2008). Instructional discourse in bilingual settings: an empirical study of linguistic adjustments in content and language integrated learning. *Language Learning Journal 36*(1), 21-34. doi: 10.1080/09571730801988470

McDonough, S.H. (1995). *Strategy and skill in learning a foreign language*. London: Edward Arnold.

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

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

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. doi: 10.1111/j.1540-4781.2009.00976.x

O'Donnell, M. E. (2009). Finding middle ground in second language reading: pedagogical modifications that increase comprehensibility and vocabulary acquisition while preserving authentic text features. *Modern Language Journal 93*(4), 512-531. doi: 10.1111/j.1540-4781.2009.00928.x

Paran, A. (2010). More than language: The additional faces of testing and assessment in language learning and teaching. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language education* (pp. 1-13). Bristol: Multilingual Matters.

Parshall, C. G., Davey, T., & Pashley, P. J. (2005). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129-148). Amsterdam: Kluwer.

Perry, F. L. (2005). *Research in applied linguistics*. Mahwah, NJ: Lawrence Erlbaum.

Piboonkanarax, K. (2007). *A survey of secondary school evaluation procedures focusing on continuous assessment*. MA thesis, King Mongkut's University of Technology Thonburi.

Pino-Silva, J. (2008). Student perceptions of computerized tests. *ELT Journal 62*(2), 148-156. doi: 10.1093/elt/ccl056

Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French*. PhD thesis, Pennsylvania State University.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology 5*(2), 84-94.

Shamim, F. (1996). Learner resistance to innovation in classroom methodology. In H. Coleman (Ed.) *Society and the Language Classroom* (pp. 105-121). Cambridge: Cambridge University Press.

Sharpe, T. (2006). 'Unpacking' scaffolding: Identifying discourse and multimodal strategies that support learning. *Language and Education 20*(3), 211-231. doi: 10.1080/09500780608668724

Simard, D. (2009). Differential effects of textual enhancement formats on intake. *System 37*(1), 124-135. doi: 10.1016/j.system.2008.06.005

Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology 16*(3), 10-20.

Urquhart, S., & Weir, C. (1998) *Reading in a second language*. London: Longman.

Vygotsky, L. S. (1978). Interaction between learning and development. In P. A. Richard-Amato (Ed.), *Making it happen: Interaction in the second language classroom* (pp. 342-353). New York: Longman.

Wang, T.-H. (2011). Implementation of web-based dynamic assessment in facilitating junior high school students to learn mathematics. *Computers & Education 56*(4), 1062-1071. doi: 10.1016/j.compedu.2010.09.014

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*. Bangkok, Thailand. pp. 118 – 127.

Watson Todd, R. (2012) English language assessment practices in ASEAN. *Proceedings of the 2012 International Conference "Cultural and Linguistic Diversity in ASEAN"*. Bangkok, Thailand. pp. 27-36.

Watson Todd, R., & Jaturapitakkul, N. (2013) Support Adaptive Testing: Towards a new future in language education. *Proceedings of the LIROD International Conference*. Bangkok, Thailand. pp. 54-59.

Watson Todd, R., & Shih, C.-M. (2013) Assessing English in Southeast Asia. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (chapter 102). New York: Wiley.