# WHY SHOULD THE WEB-BASED ACHIEVEMENT TESTS IN ENGLISH FOR TOURISM BE IMPLEMENTED?

Malinee Phaiboonnugulkij* and Kanchana Prapphal**

**บทคัดย่อ**

บทความนี้นำเสนอความต้องการการสนับสนุนกรอบทฤษฎีและขั้นตอนการตรวจสอบความตรงของแบบทดสอบการพูดบนเว็บไซต์สำหรับภาษาอังกฤษเพื่อการท่องเที่ยวซึ่งใช้เทคโนโลยีเวิลดไวด์เว็บในการสร้างแบบทดสอบ การทดสอบและการเก็บข้อมูล ในงานวิจัยฉบับนี้ การสร้างแบบทดสอบการพูดบนเว็บไซต์สำหรับภาษาอังกฤษเพื่อการท่องเที่ยวประกอบด้วย 4 ขั้นตอนหลัก คือ การวิเคราะห์สถานการณ์ที่ใช้ภาษาเฉพาะทาง การคัดเลือกและจัดหมวดหมู่กิจกรรมแบบทดสอบทางภาษาเฉพาะทาง การพัฒนาต้นแบบกิจกรรมแบบทดสอบโดยคำนึงถึงการออกแบบจอภาพ และขั้นตอนการตรวจสอบความตรงของแบบทดสอบ การสร้างแบบทดสอบออนไลน์ประกอบด้วยข้อพิจารณาหลักคือ ลักษณะเฉพาะของแบบทดสอบภาษาอังกฤษเฉพาะทางซึ่งเกี่ยวเนื่องกับความเสมือนจริงของแบบทดสอบ ความตรง และความเหมาะสมในการออกแบบจอภาพและการใช้สื่อผสมในแบบทดสอบ การศึกษานี้นำเสนอความเข้าใจอย่างชัดเจนในการผสมผสานเทคโนโลยีโดยการใช้แบบทดสอบการพูดบนเว็บไซต์สำหรับภาษาอังกฤษเพื่อการท่องเที่ยวในบริบทของประเทศไทย

**Abstract**

This article calls for a strong need to advocate the theoretical framework and validation procedures that underpin any web-based speaking tests in English for tourism which incorporates the World Wide Web technology in test construction, administration and data storage. In this research, an online web-based speaking test in English for tourism was developed and the test consisted of four main stages: Analysis of Target Language Use (TLU) situation, Selection and Categorization of the TLU tasks, Development of the Prototype Tasks with the consideration of the interface design, and Validation Procedures. There were a number of considerations when constructing this online test, particularly on the issue of salient features of LSP test that dealt with authenticity of test tasks and test

*Malinee Phaiboonnugulkij is a Ph.D. candidate in the English as an International Language program, Chulalongkorn University. Her research interests are SLA acquisition and testing.

**Dr.Kanchana Prapphal is a Professor Emeritus at Chulalongkorn University Language Institute. Most of her publications involve in the area of language teaching and testing.

validity, and the appropriate use of the interface design and multimedia. This study provides some insights in the integration of technology using web-based tests in English for tourism in the Thai context.

## INTRODUCTION

Due to the importance of tourism business that creates an estimated 6.7% of all the GDP in Thailand in 2007 (Thailand Tourism Review, 2008), a number of educational institutions offer English for tourism courses which aim to develop professional tourism staff including tour guides who are proficient in English language. In order to pass the course, students are required to pass a test that requires them to be at a certain level of oral proficiency. For this reason, a good test is needed to assess their skills precisely and accurately. The English for Tourism course is one of the Language for Specific Purposes (LSP) courses at Nakhon Ratchasima Ratchabhat University (NRRU) that requires a high quality LSP test that can assess students' speaking ability in tourism context. This tourism-oriented LSP speaking test is constructed for diagnostic purpose and expected to be used with a large number of students. Technology-integrated test, particularly the Web-based testing (WBT), is purposively selected in order to meet the requirements of oral assessment in the LSP course.

The sophistication of WBT has shaped the testing facets in test construction, administration and scoring methods (Roever, 2001, Hamilton, Klein &Lorie, 2000, Garcia Larboda, 2007a, Garcia Larboda, 2007b). WBT's advantages over Computer-based testing (CBT), particularly with regards to logistic flexibility, have resulted in WBT being used by the two most influential English language standardized tests, the TOEFL Internet-based test (iBT) and IELTS Computer-based test (CBT) (Garcia Larboda, 2007a, Garcia Larboda, 2007b and Alderson, 2009). Regarding pedagogical advantages, test takers also get immediate and specific feedback which is pertinent to some certain aspects of learner-centered second language assessment (Chapelle, Jamieson and Hegelheimer, 2003). This online test is mentioned by a number of scholars for its suitability as a low-stake test particularly for self-diagnostic purposes (Chapelle & Douglas, 2006, Roever, 2001). Hamilton, Klein & Lorie (2000) and Garcia Laborda (2007b) discuss the feasibility for using WBTs for large scale standardized tests due to the numerous technological advantages on inexpensive and rapid scoring, central storage of item banks, and less dependence on sophisticated software and hardware. All of these make WBT suitable for large scale testing. The authors also mention that the advance of technology made it possible to create new types of questions that can assess complex metacognitive skills. In addition, Garcia Laborda (2007b) projects that numerous standardized CBTs will eventually be available online and will include a speaking section. Garcia Larboda (2007a) proposes that the online speaking test should incorporate interactive audio input that is likely to effectively assess the complexity of language constructs. The reliability of the test can be fa-

cilitated by the grading system that allows numerous reevaluations. The use of video recording can make it possible for more effective assessment of non-linguistic features than traditional audio recording. Thus, the use of WBT has been claimed to enhance the reliability of scoring as raters could revise, reevaluate and adapt their scoring to optimally conform to common criteria. Consequently, both intra and inter rater discrepancy can be reduced. Although WBT has been praised for its practicality, this innovative test has some technical and practical limitations on the compatibility of technical devices and browser and cheating (Hamilton, Klein & Lorie, 2000, Roever, 2001 and Garcia Laborda, 2007b).

Consideration about practical and technological advantages and limitations aside, a number of scholars claim that WBT could be faster, more efficient and cost effective than the traditional test version; thus advocating that the benefits of this innovative test should outweigh its pitfall. (Garcia Larboda, 2007a, Garcia Larboda, 2007b and Hamilton, Klein & Lorie, 2000). From the point of practical and technological advantages, WBT is employed in a wide range of language testing skills with the exception on speaking ability assessment (Garcia Larboda, 2007a, Garcia Larboda, 2007b). The salient feature of the incorporation of technology into testing format requires a specific theoretical framework and validation procedures in order to create a good quality test particularly for the WBST-EFT. The test is expected to be used with a large number of test takers and as the first prototype online speaking test for the university. Thus, there is a strong need to understand the theory that underpins this innovative test, particularly on speaking skill assessment. The present study therefore aims to provide a theoretical framework and validation procedures for the online speaking skill assessment in the tourism context which is a subcategory of LSP testing. The emphasis will be on authenticity of test task and validation of test by taking into account the appropriate use of multimedia and interface design. The typical development procedures consist of five main stages: Analysis of specific purpose language use situation, Selection and classification of tasks and situations, Development of the WBST-EFT and the rating scale, Validation procedures and Stage five: Empirical evidence.

## THEORETICAL FRAMEWORK FOR THE WBST-EFT AND THE RATING SCALE CONSTRUCTION

In test construction, the theoretical framework is required and the WBST-EFT underpins the contextualized communicative language ability which is LSP testing theory. The WBST-EFT was developed under the theoretical framework of an LSP test development proposed by Douglas (2000) which is modified from the framework of Bachman and Palmer (1996) in line with the Interface Design framework for technology integrated test from Fulcher (2003a).

According to Douglas (2000), the prominent scholar in LSP testing, LSP testing poses salient aspects that distinguish it from more general purposes language testing including authenticity of task and the interaction between language knowledge and specific purpose content knowledge. Test contents and methods arise from the analy-

sis of language use in the Target Language Use (TLU) situations. Thus, the test tasks and contents correspond closely to the tasks specifically performed in specific context. In the LSP test constructs referred to by Douglas (2000), LSP ability incorporates language knowledge and most importantly, specific tourism background knowledge which is a prominent feature of an LSP test. Therefore, the inference of the LSP test performances can be made to the actual and specific TLU domain.

Similar to other frameworks, the LSP ability model is questioned by a number of scholars on the issues of the inclusion of the field specific content knowledge on language constructs and its effect on language performances (Wu & Stansfield, 2001, Clapham, 1996, Tan, 1990 and Krekeler, 2006). However, the findings of these studies remain inconclusive. Another issue is on the specificity of LSP test that has been asserted by several researchers on the lack of precise theoretical basis in this language testing skill (Davies, 2001, Cumming, 2001, Wu & Stansfield, 2001 and Elder, 2001). On the other hand, Douglas (2000) mentions that LSP testing is a sub-category of language testing, and that the content and test method result from the Target Language Use analysis (TLU) (Douglas, 2000:19). LSP therefore poses certain and precise characteristics used by people in the field that people who are not in the field do not have a thorough understanding of.

Since WBST-EFT is a technology-based test, a framework on this innovative test is required. In the current study, the Interface design framework by Fulcher (2003a) is used. This framework has taken into account the appropriate use of technological devices in test design, particularly the components that will appear on the computer screen. The inclusion of multimedia prompts could facilitate more realistic test tasks for the test takers and the technology makes it possible to measure metacognitive skill and effectively measure complex language constructs (Garcia Larboda, 2007a, Garcia Larboda, 2007b and Hamilton, Klein & Lorie, 2000). As part of the WBST-EFT, the rating scale employs similar constructs of the LSP ability proposed by Douglas (2000) and speaking ability framework proposed by Fulcher (2003b). These frameworks were integrated into the test development and the following figure 1 illustrates the framework of the WBST-EFT.

## ANALYSIS OF SPECIFIC PURPOSE LANGUAGE USE SITUATIONS

According to Douglas (2000), the first step in the LSP test design procedure is to describe and to analyze the target language use situations in order to establish the characteristics of context and task. This will help to ensure that the test tasks incorporate the important features of the target language use domain.

The content from the Tourism Authority of Thailand Tour Guide Training Curriculum (Foreign Language) (1996) indicates that one of the prominent roles of tour guides is to be the cultural representative of the nation, and to be familiar with the following language use task in tourism context: presenting tourism information, informing the tourists about Do's and Don'ts of Thai culture and laws. Additionally, the professional tour guides also act as problem solvers and

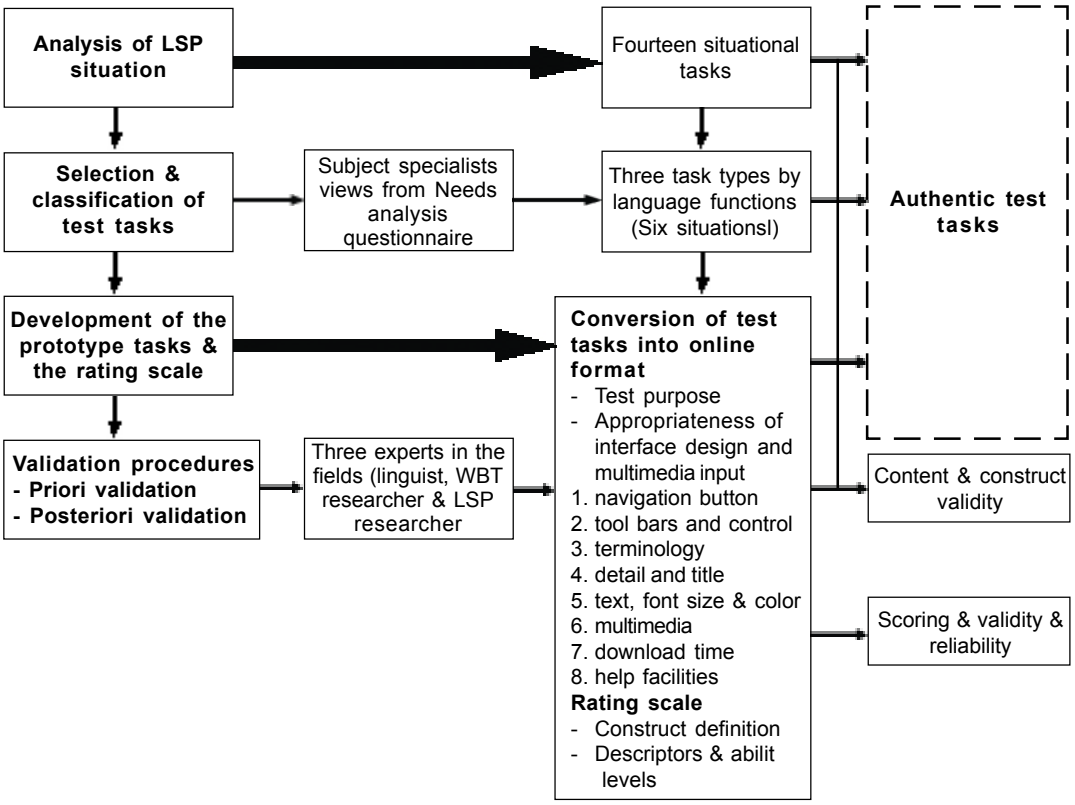


**Figure 1: The WBST-EFT Framework**

negotiators for the tourists in situations which could arise at places of tourist attractions, on the bus and at the hotel.

In order to achieve the purpose of the test as the final achievement test, the language use situations stated in the English for Tourism II (EF II) course description and course syllabus were analyzed in terms of context, and the tasks provided guidance on task construction and content coverage in the WBST-EFT. The situations and tasks that were pertinent to EF II course syllabus were reviewed and included in the Needs Analysis Questionnaire and the WBST-EFT.

In order to obtain the authentic language use tasks, three subject specialists in the tourism field were consulted regarding the characteristics of language use tasks that were frequently performed by the tour guides during tour operation. The Needs Analysis Questionnaire was distributed to obtain their opinions. The consensus of the subject experts was used by some LSP direct speaking tests to derive the test tasks and contents (Brown, 1995).

- **Needs Analysis Questionnaire**

A needs analysis questionnaire was developed by the researcher to investigate:

1) the target language use tasks and situations

2) language knowledge required in professional tour guides

3) criteria for assessing the language knowledge from the subject specialists in

the fields. The information from the needs analysis questionnaire was gathered from the related literature and the contents from EF II course syllabus analysis. The information obtained from this instrument was used in the development of the WBST-EFT and the rating scale to create the tasks that closely corresponded to the real world tasks and actual language knowledge used by professional tour guides.

There were four parts in the questionnaire. Part one related to the demographic information of the content area specialists. Part Two and Three provided a 4-point Likert scale asking the degree of importance about the tasks and situations most likely to be used by professional tour guides and their language knowledge. In addition, Part Three touched on additional language knowledge and testing. Part Four was also a 4-point Likert scale with open-ended questions that asked about the appropriateness of criteria for assessing the language knowledge of the tour guides.

The questionnaire was validated by three experts in the field with the Index of Item-Objective Congruence value at 1.00 for each item. Some parts of the questionnaire were revised according to the experts' suggestions and tried out with three subject specialists. Then, the instrument was revised again particularly on the clarity of language in some parts. The final questionnaire was administered with fifteen subject specialists who each had a minimum of seven years of experience in tourism and English for tourism instruction. This group of people included five travel agency owners who used tour guides, five English-speaking professional tour guides who hold bronze type tour guide license, and five English for Tourism II course lecturers. To ensure the validity of the information, they were required to provide demographic information which was directly related to their expertise in Tourism or in English. Almost all of them had master's degrees in tourism or in English and one held a bachelor's degree in tourism. Their experience in their profession ranged from seven to nine years, except for two of the specialists who had 35 to 37 years in the field.

Results from the needs analysis questionnaire were based on the degree of importance from very important (4) to important (1) and are discussed in the next step.

## SELECTION AND CLASSIFICATION OF TASKS AND SITUATIONS

The next step is the selection and classification of tasks and situations. The tasks that were rated by most of the specialists only in "very important" and "important" category were included in the WBST-EFT. After that they were classified into three language functions based on Douglas (2000) LSP ability framework. The classification was as follows.

- Presenting tourism related information (heuristic function)
- Informing tourists about what they should do and should not do in Thailand (manipulative function)
- Dealing with tourists' enquiries and complaints (ideational function)

In addition, the subject specialists were asked to specify the degree of importance for the components of language knowledge used by the tour guides when organizing trips. These components were also proposed in Douglas (2000)'s LSP ability

framework in line with Fulcher (2003b)'s speaking ability model. The components that were pertinent to English for Tourism II course syllabus were included. The information obtained in this part was used in the rating scale construction.

The results of the questionnaire revealed that the knowledge of vocabulary was rated as a very important feature by almost all of the subject specialists, and fluency and content knowledge also posed the same degree of importance. There was also general consensus by the specialists that knowledge of grammar was another important component in the tourism domain. Knowledge of pronunciation, language function and cohesion were also considered important features.

With regards to the appropriateness of the criteria for assessing language knowledge, all of the specialists agreed that the range of speech was the most appropriate criterion. Appropriate use of grammatical structures and language functions with the consideration of sociolinguistic domains was also rated as an important criterion, whereas both accuracy and fluency were rated as very appropriate. Therefore, all these criteria were included in the rating scale.

It is clear at this stage that the analysis of the specific characteristics of the target language use in context and tasks is a vital procedure in the WBST-EFT development. Without this procedure, the WBST-EFT will not cover the important elements of the actual tasks in tourism context and it will directly affect the authenticity of the test tasks.

Regarding the specificity of LSP tests, Douglas (2001) mentions that all tests are developed for some purposes and they will fit in the particular point of the continuum of specificity. LSP tests therefore must include certain precise characteristics used by people in the profession such as specific pronunciation, vocabulary, word meaning, and sentence structures. People who are not in the fields will not have a thorough understanding of these. Thus, the subject specialists' view is purposively obtained in this study in order to provide the specific features that must be included in the WBST-EFT tasks and rating scale.

## DEVELOPMENT OF THE WBST-EFT AND THE RATING SCALE

### - WBST-EFT Prototype Tasks Development Procedures

Another step deals with the test specifications and the actual test tasks development. The test specifications or test blueprint was written to be used as a plan for the WBST-EFT construction. This is an important step that cannot be excluded from any test developments. This planning guides the WBST-EFT test developer about the test purposes and language ability to be measured. The blueprint will also be used as the guideline for the WBST-EFT item writing and task construction. In addition, it also gives details on the scoring criteria, procedures and interpretation for the raters. It provides information on test objectives, test constructs and interpretation of test performances for the test users. Finally, the details of the test specification can be used as part of the validation procedure to provide empirical evidence on test validity.

Concerning the test constructs,

Douglas's (2000) LSP ability framework and Fulcher's (2003b) Speaking Ability were incorporated in the WBST-EFT. Some components of LSP language ability were purposively selected on their relatedness to English for Tourism II course. As a part of LSP constructs, background knowledge was investigated by some studies with varied results (Clapham, 1996, Tan, 1990 and Krekeler, 2006). However, some studies revealed the supportive effects of this knowledge on test performance (Clapham, 1996); therefore, it was included in the WBST-EFT. Fluency of speaking ability was used as a criterion in several LSP speaking tests (Brown, 1995, ILEC Handbook, 2008, BULATS Handbook, 2009); thus, it was included in the WBST-EFT constructs.

After the specifications have been drafted, the actual test tasks were constructed and converted into the online format. In the WBST-EFT, test takers would act as the tour guides organizing the trip in the central part of Thailand. Drawing on previous studies, English for Tourism II course syllabus analysis, and the data derived from the Needs analysis questionnaire, the researcher created three target language use task types and six situational tasks as these types of language and situations are most likely to be used by professional tour guides. There were three sections in the WBST-EFT, which were categorized by task types, and each task type had two subtasks. Each test task purposively incorporated multi-media in order to simulate a real world task and make it live. The test takers had preparation time in sections one and two. The information about the preparation and response time and marking criteria was available in the instructions part of the test. The whole test lasted approximately twenty four minutes. The test takers could take the sample test so as to be familiar with the test. They were required to respond to each test task by clicking on the record button and start speaking through the microphone when they heard the sound "beep" and saw the "start speaking" prompt on the screen. After they finished speaking, they had to click on the same button again to stop. They were allowed to record their response only once. Then, they clicked the next button to move onto the next section.

In Sections One and Two, motion pictures with audio input were presented to elicit the heuristic and manipulative functions of the test takers. In Section Three, short video clips which simulated real world scenarios were presented. The last part was reciprocal in nature requiring the test takers to interact with the scenarios by using the ideational language function.

In Section Three, after the scenarios that are likely to happen in the organized tour were selected, short video clips were created. The TLU characteristics were used in creating the dialogues between the tour guide and the tourists. There were six scenarios in this section and three of them were classified into responding with tourists' enquiries, and the other three dealt with complaints. All of the topics and situations were selected from previous studies and at the suggestion of the subject specialists. The dialogues were checked by a linguist and a professional tour guide. Some parts were revised to improve the clarity and accuracy of language particularly at the discourse level.

The test tasks were then converted into short films. Eight young ambassadors at

Nakhon Ratchasima Rajabhat University voluntarily participated in this research project. This group of students was trained as professional tour guides. There was a rehearsal of all six scenarios before the actual filming took place. To increase the authenticity of test tasks, all of the scenes were recorded at actual sites such as at the hotel, on the bus, and at the tourist attractions. The films were edited by the researcher using free downloadable software programs: Window Movie Maker version 2.6 and Sound Forge Trial version 9.0.

All of the six prototype tasks were posted on Moodle version 1.9.5, a free online platform that is currently being used at NRRU. The test takers' speech productions were recorded with Sound Forge software program version 9.0. Their responses were stored in this platform and can be retrieved online by the raters. The details and objectives of the test tasks are presented below.

Section 1 (Task type 1): Presenting tourism related information

This section aims to elicit the test takers' ability in presenting national tourist attractions and explaining the tour program.

Task 1: Presenting tourist attractions

The first task aimed at eliciting the test takers' ability to present two of the most famous national attractions in Thailand: the Emerald Buddha Temple and the Grand Palace. The test takers were provided with seven pictures about the two sites (four about the Emerald Buddha Temple and three about the Grand Palace) and they were asked to explain these pictures in detail. They had seven minutes to work on this task. For each picture, they had twenty seconds to prepare their responses and the remaining 40 seconds were for their speech production.

Task 2: Describing one day tour program in the central region of Thailand

In Task 2, the test takers first read the one-day tour itinerary, and then they were required to present the information to tourists. They were asked to provide additional details of the underlined attractions.

Section 2 (Task type 2): Giving polite suggestions to tourists

The objective of Task type 2 is to assess the test takers' ability in giving polite suggestions to tourists in two different situations.

Task 3: At the Summer Palace

The test takers first watched the video clip containing a monologue of the tour guide at the Summer Palace. Then, there were six pop up pictures on the clip which required the test takers to give polite suggestions on what the tourists should do and should not do in each situation based on Thai cultural and religious beliefs.

Task 4: At Jatujak Market

The test takers attempted this task in a similar fashion to Task Three. They first watched a video clip and were asked to respond to the six pop up pictures containing different scenes by giving polite suggestions regarding what the tourists should do at the crowded shopping center.

Section 3 (Task type 3): Dealing with enquiries and complaints

Task type 3 emphasizes the test takers' ability to deal with tourists' enquiries and complaints on a variety of topics.

Task 5: Dealing with enquiries

The test takers first watched the video clip containing the dialogue of three different enquiries: asking for help in recovering

a stolen wallet, requesting a guide to explore the night life, and requesting medical assistance. At the end of each dialogue, the test takers were asked to respond to the enquiry politely and appropriately.

Task 6: Dealing with complaints

Task 6 incorporated three complaints: an incomplete tour program, an unrequested hotel room, and prolonged wait for a bus. The test takers first watched video clips containing different complaints, and they were required to respond each complaint politely and appropriately.

- **The Rating Scale Construction**

The WBST-EFT rating scale is an essential instrument in scoring the speaking performance. It is specifically used with the target population and test purpose. The WBST-EFT rating scale provides operational definitions of LSP constructs in tourism and levels of mastery of these features in completing the test tasks. The description in the scale must be explicit, precise and able to differentiate the test takers' levels of mastery of the constructs. Raters are needed to be trained to use the scoring scale in order to obtain the reliability of their rating. As with the test, the rating scale must be designed under the theoretical frameworks and has to undergo the validation process. The WBST-EFT rating scale was developed under the rating scale development proposed by Fulcher (2003b). The details of the development procedure were presented to justify the use of this instrument in scoring procedure and interpretation.

First, the language ability in the rating scale was defined. As part of the WBST-EFT, the rating scale employs similar constructs of the LSP ability proposed by Douglas (2000) and speaking ability framework proposed by Fulcher (2003b).

Then, the purpose and type of rating scale were decided. It is assumed that both the purpose and type of rating scale guide the rating procedures and scoring interpretation. Thus, they are purposively decided with the consideration of their usefulness and suitability with the WBST-EFT. The purpose of the WBST-EFT rating scale is to be used for guiding the rating process emphasizing the quality of the performance. Therefore, it is considered as the "Assessor-Oriented Scale" (Fulcher, 2003b). The scale contains operational construct definitions that are easy to comprehend within a short time.

Regarding the type of scale, the analytic rating scale was used in this study due to its appropriateness with the test purpose as the classroom final achievement test and for the diagnostic purpose. Hence, the analytical scale is appropriate for these purposes. The analytical rating scale allows for assessing specific components of language ability defined by the constructs definitions. Additionally, each of the scale descriptors contains a specific level of mastery of language ability. Therefore, either the mastery or failure of the specific language components can be indicated. This analytic scale can provide information related to the strengths and weaknesses of the test takers, and the information can be used for remedial courses and instructional approach designs.

In terms of criteria for correctness, accuracy was used in rating the speaking responses. The notions of accuracy in linguistic elements, range and appropriateness of

speech production were used as the criteria by a number of LSP speaking tests (Brown, 1995, ILEC Handbook, 2008, BULATS Handbook, 2009). Therefore, these elements will be included in the WBST-EFT rating scale.

After that, the number of levels of ability on the scale was decided and the band descriptors were written. Like any test, the rating scale requires an appropriate design to derive the number of levels of ability and descriptors that are clear and precise in differentiating the mastery of the test takers' language ability in completing the test tasks. Approaches to a rating scale design proposed by Fulcher (2003b) were used to develop the descriptors of the WBST-EFT rating scale. The design of the scale was based on the expert-judgement method. The main researcher who taught the English for Tourism II course for four years and worked as a professional tour guide for seven years wrote the band descriptors. The expert-judgement method in rating scale design requires a number of years in field experience. Prior to band descriptor development, the number of levels of ability was decided. The WBST-EFT consisted of five language abilities starting from level 0 (a very poor user) to 4 (a very good user). The number of ability levels was pertinent to the course grades which started at F and went up to A. In this way, the band levels were conveniently converted into course grades. The sequence of band descriptors meaningfully and clearly reflected a progression in LSP language ability. The band descriptors were relevant to the language requirements stated in the course syllabus, and were based on the experts' recommendations from the needs analysis questionnaire. The LSP ability and speaking theoretical framework were included in the rating scale. The criteria for correctness were modified from Fulcher's (2003b) Speaking ability.

Before the next stage, rater selection and training were conducted to ensure that all raters are qualified and consistent in rating procedures. There were two raters in this session. The criteria for rater selection were from the years of experience in their profession and their language proficiency scores. One of the two raters is an English speaking professional tour guide who holds a Bronze type tour guide license with seven years of experience in the field. This rater achieved a band score of 6.5 in the IELTS. The other rater was English for Tourism II course lecturer who had 37 years of teaching experience and was also a trainer of the TAT tour guide training course.

The final step is the pre-rating session. All the raters were provided with the rating form, descriptors of the criteria and description of rating procedures. They were trained to understand the descriptors and criteria, to follow the rating procedures and to appropriately apply the rating scale. This session was carefully arranged because it could affect the scoring validity of the test.

## VALIDATION PROCEDURES

In validating the WBST-EFT, both priori and posteriori validation procedures were adopted. To establish the content and construct validity evidence, the two instruments were validated by three experts in the field by using the index of Item-Objective Congruence (IOC>.75). The description of the test specifications was used to establish

its validity evidence for both contents and constructs (Weir, 2005, Bachman & Palmer, 1996). The test and rating scale were revised according to the experts' suggestions.

After the two instruments were validated, they were piloted with 30 samples who were classified into two groups based on their English for Tourism I course grades. There were fifteen samples in each proficiency group, and they were further divided into three equivalent subgroups. Each subgroup consisted of five subjects. In each proficiency level, the students were randomly assigned into three task type groups.

As for the posteriori validity evidence regarding scoring validity and reliability, item analysis was carried out. IF values ranging from .20 to .80 and the minimum ID value of .30 for each test task were set. Pearson Correlation was applied to assess the inter-rater reliability of the rating scale and the reliability coefficient was established at .70. Since the WBST-EFT is a subjective test that requires judgmental scoring, the raters should be consistent in scoring the performance of the subjects. As mentioned earlier, scoring consistency can be derived through rater training.

## EMPIRICAL EVIDENCE

From the analysis, it was found that the index of Item-Objective Congruence by the three experts in the field for each test task and descriptor in the rating scale posed the value of 1.00. This reflected a high validity of the contents and constructs of the instruments. In addition, the result from the item analysis of the WBST-EFT yielded high values of item discrimination index which ranged from .58 to .63 for the six tasks. This means all the test tasks could effectively differentiate the mastery levels of all the test takers. For the difficulty level of the test tasks, the values ranged from .28 to .35, which could be interpreted that the test was quite difficult. Regarding the reliability of the test, Cronbach's alpha value was .98, which reflected a high reliability of rating and could be claimed that the raters were highly consistent in their rating.

## CONCLUSION AND DISCUSSION

Due to the need on specific theoretical framework and validation procedures on the innovative online speaking test, the WBST-EFT was constructed. This web-integrated test is an LSP test in which test contents and methods result from the analysis of language use in tourism situations. Thus, the test tasks and contents closely corresponded to the tasks specifically performed in tourism context. On the aspect of the test constructs, the WBST-EFT meets the LSP ability framework proposed by Douglas (2000) which incorporates language knowledge and most importantly, specific tourism background knowledge which is a prominent feature of an LSP test. Therefore, the inference of the WBST-EFT performances can be made to the actual tourism settings.

The analysis of TLU characteristics and situations is a significant and initial procedure in an LSP test development that leads to authenticity of the test tasks and contents. Prior to the actual test construction, the selection and categorization of TLU tasks and situations is an essential stage that also affects the authenticity of test tasks. With this

stage, subject specialists should be consulted about the importance of selecting TLU tasks performed by the professional tour guides because LSP poses certain features that only people who are in the field can understand.

The next stage is the development of the actual test tasks and the rating scale which includes the use of technology in test construction. The WBST-EFT, as its name suggests, integrates the World Wide Web technology in test construction and administration. Considerations on the appropriateness and effectiveness of the multimedia types are vital. Not only can the multimedia input e.g. audio, visual and video clips make the test live but also effectively create interaction between the test takers and the test tasks. However, this type of technology can be construct-irrelevant that affects the validity of the test (Fulcher, 2003a, Gracia Larboda, 2007a) when it is not carefully used. For this reason, TLU tasks and objectives of the test play a crucial role in the design of the test. In this stage, all of the six dialogues from the last two tasks were validated from two subject specialists before the actual prototype tasks construction. They were revised according to the specialists' suggestions. The computer-based framework (the previous version of the web-based test), was used as a guideline in the design of the WBST-EFT.

To ensure the quality of the test, the validation procedure by the experts in the field related to technology-based testing, linguistics and tourism was carried out. The empirical evidences from both priori and posteriori validation stages were then used to claim the validity of the contents and constructs including the reliability of the WBST-EFT and the rating scale. Finally, the test was revised according to the experts' suggestion before the main study.

The Web-based testing poses a number of advantages such as logistic flexibility with a large number of test takers, no expertise requirement in programming, sophisticated software and hardware, integration of interactive input and, most importantly, cost-effectiveness in test construction by using free online software (Hamilton, Klein & Lorie, 2000 and Roever, 2001 and Garcia Laborda, 2007b). However, the theoretical issue of test validity has been raised by several scholars (Chapelle, Jamieson and Hegelheimer, 2003). Following Chapelle, Jamieson and Hegelheimer (2003), the influential scholars in WBT, the objective of the WBST-EFT and the washback as the final achievement test have been primarily considered in the test design and validation procedure. The empirical evidence from both priori and posteriori validation procedures was used to claim both content and construct validity and reliability of the test.

Apart from the theoretical issue of the WBT, practical issues on technical limitations, item-confidentiality and testing environment are significant, and may prove to create threats to the present study. The first issue deals with the data storage of Moodle 1.9.5 that allows a maximum of 10 Megabytes for each video file. It may sometimes affect the quality of pictures on the screen, and the test developers should keep in mind about this limitation when using this online platform. Another technological aspect is the breakdown of technology, and the test developer should have a backup plan for this unexpected situation. Item-security is an-

other limitation and is considered as the most influential drawback of the WBT. This limitation can be overcome by creating a large pool of item banking. As for the testing environment (particularly the computer laboratory in the present study), noise distraction from other test takers may possibly affect the test takers during test taking process. Test takers should therefore be provided with seatings that are sufficiently far apart without creating disruptions to others.

WBST-EFT is one example of an in-house online speaking tests that is expected to be used with a large number of the test takers. Thus it must be constructed with a particular theoretical framework and put through the validation procedures to ensure an acceptable standard. WBST-EFT underpins both the concepts of LSP test and WBT construction which are illustrated in figure 1. Assessing speaking ability is a challenging task for test developers. However, making use of the technological and practical advantages then specific framework, offers web-based tests as an achievement test may be feasibly employed in the Thai context as explained in this article.

## REFERENCES


Bachman, L. and Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, A. (1995). "The effect of rater variables in the development of an occupation-specific language performance test". *Language Testing*, 1-15.

Chapelle, C. A., Jamieson, J. and Hegelheimer, V. (2003). "Validation of a web-based ESL test". *Language Testing*, 20, 4, 409-439.

Clapham, C. (1996). *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Cumming, A. (2001). "ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes?" *Language Testing*, 18, 2, 207-224.

Davies, A. (2001). "The logic of testing Languages for Specific Purposes". *Language Testing*, 18, 2, 133-147.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Douglas, D. (2001). "Three problems in testing language for specific purposes: Authenticity, specificity and inseparability". In Elder, C., Brown, A., Grove, W., Hill, K., Iwashita, N., Lumley, T, McNamara, T. and O'Loughlin, K. (eds.), *Studies in Language Testing: Experiencing with uncertainty*. Essays in honor of Alan Davies. Cambridge: Cambridge University Press, pp.45-52.

Elder, C. (2001). "Assessing the language proficiency of teachers: are there any teacher controls?" *Language Testing*, 18, 2, 149-170.

Fulcher, G. (2003a). "Interface design in computer-based language testing". *Language Testing*, 20, 4, 384-408.

Fulcher, G. (2003b). *Testing second language speaking*. London: Longman/ Pearson Education.

Garcia Laborda, J. (2007a). "From Fulcher to PLEVALEX: issues in interface design, Validity and reliability in internet based language testing". CALL-EJ

Online, 9. Retrieved on September 10, 2008 from: http://www.tell.is.ritsumei.ac.jp/callejonline/journal/9-1/laborda.html.

Garcia Laborda, J. (2007b). "On the net: introducing standardized EFL/ESL exams". *Language Learning and Technology*, 11, 2, 3-9.

Hamilton, L.S., Klein, S.P., & Lorie?, W. (2000). *Using web-based testing for large-scale assessment*. Santa Monica, CA: RAND Corporation.

International Legal English Certificate Handbook. (2008). Retrieved on November 20, 2009, from: http://www.legalenglishtest.org/downloads/ilec_handbook.pdf.

Krekeler, C. (2006). "Language for academic purposes (LSAP) testing: the effect of background knowledge revisited". *Language Testing*, 23, 1, 99-130.

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

Tan, S. (1990). "The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates". In J.de Jong and D. Stevenson (eds.), *Individualizing the assessment of language abilities*. Clevedon, UK: Multilingual Matters, pp. 214-224.

Thailand Tourism Review. (2008). "Diethelm Travel's Thailand Tourism Review 2007". Retrieved October 20, 2009, from http://www. bangkokpost.com/tourismreview2007/

The Business Language Testing Service (BULATS) Test Specification: A Guide for Clients. (2009). Retrieved on October 25, 2009, from: http://www.bulats.org/docs/BULATS-Test-Specifications-a-Guide-for-Client-June09.pdf.

Tour Guide Training Curriculum (Foreign Language) Handbook. (1996). Bureau of Tourism and Guide Registration, Department of Tourism, Tourism Authority of Thailand.

Weir, C. (2005). *Language Test Validation: an evidence-based approach*. Oxford: Palgrave.

Wu, W.M. and Stansfield, C.W. (2001). "Towards authenticity of task in test development". *Language Testing*, 18, 2, 187-206.

## Appendix

## Web-Based Speaking Test in English for Tourism (WBST-EFT)

Task type 1(Task 1)


Task type 2 (Task 3)

Task type 3 (Task 5)