PASAA
Volume 44
July 2012

LANGUAGE FOR SPECIFIC PURPOSES (LSP) WEB-BASED ASSESSMENT AND THE SPEAKING PERFORMANCES OF TWO ABILITY GROUPS [1]

Malinee Phaiboonnugulkij
Kanchana Prapphal
English as an International Language Program
Chulalongkorn University

Abstract

This study examines the validity and reliability of an Internet-integrated test to assess language for specific purposes (LSP) speaking abilities of the students in the English for Tourism course at a Thai university, and explores whether this test can identify differences in the students' abilities in three task types in the context of tourism in Thailand. The sample group was comprised of 120 third-year university students. A two-way ANOVA was conducted to explore the differences in the performances of the two ability groups in attempting the three test-task types. Content analysis was employed to investigate the similarities and differences in each LSP component of the two ability groups. Statistical analysis indicated that the web-based speaking test in English for Tourism (WBST-EFT) was an effective assessment tool for a large number of students, and that it posed high content and construct validity, reliability, and practicality. The results showed that the test effectively identified differences in LSP speaking performances between the two ability groups across the three task types in range, accuracy, complexity, and appropriateness of the LSP production, particularly in the content knowledge component. This insightful information should be used in future LSP curriculum development and assessment.

[1] This study is part of the first author's doctoral dissertation.

Introduction

Tourism is one of the most important industries in Thailand, contributing approximately 6.7% of the country's gross domestic product. This is the result of approximately 18.82 million tourists arriving in 2011 (Thailand Tourist Arrivals, 2011). For this reason, a large number of educational institutions offer English for Tourism courses to produce proficient English-speaking staff, particularly tour guides. Since tour guides are some of the key individuals in various tourism enterprises who directly communicate with linguistically diverse tourists, English speaking skills are essential for their chosen career.

According to Douglas (2000), English for Tourism is considered to be one area of language for specific purposes (LSP). This classification is related to the utilization of English for a particular purpose in a targeted setting. The English for Tourism course is taught at Nakhon Ratchasima Rajabhat University to a large number of students. As part of the objectives of the course, a diagnostic test is required to evaluate students' LSP abilities, particularly their speaking skills.

Due to the increasing number of English for Tourism courses and the need to be able to evaluate students' LSP skills and abilities in the University, there is a strong need for an assessment instrument that can precisely and accurately measure the speaking ability of a large number of these LSP majors. For these reasons, a technology-integrated test was purposively selected for this study primarily due to the advantages in the administration of the test, and more importantly due to the logistic flexibility in time and place that the test offers. This advanced technology made it possible to administer the test to a large number of students with more interactive input, which is capable of eliciting more complex speech performance than a more traditional format (Garcia Larboda, 2007a; Hamilton, Klein & Lorie, 2000).

Apart from the practical need for a good-quality assessment tool, the similarities and differences in LSP test performances construct the insightful information for LSP speaking ability improvement and curriculum development. Bachman and Palmer (1996) stated that the variations in test performances were affected by the test-takers' target language proficiency levels. Another potential factor linked to variation in language test performances was task type, which has been widely investigated in the traditional testing format with inconclusive results (Kim, 2009; Teng, 2008; Turner & Upsher, 1995).

Considering practical and pedagogical needs, this article aims to examine the validity and reliability of the LSP online speaking test used in one of the English for Tourism courses at the University, and to investigate whether this test can discriminate between differences in the LSP speaking abilities of the students in language knowledge and content knowledge in three task types. These three task types are as follows: presenting tourism information regarding Thai attractions, giving polite suggestions, and dealing with complaints and inquiries.

Literature Review

Language for specific purposes (LSP) speaking ability

According to Douglas (2000, p. 40), language for specific purposes (LSP) ability "results from the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics." Douglas's definition of LSP ability is based on Bachman and Palmer's (1996) model of communicative language ability, with some modification to the strategic competence component, and he added the notion of background knowledge to the model. In a specific purposes context, the relationship between language ability and specific background knowledge is one of the key features of LSP. The LSP

ability model comprises three factors: language knowledge, strategic competence, and background knowledge.

Language knowledge incorporates grammatical knowledge, textual knowledge, functional knowledge, and sociolinguistic knowledge. Language knowledge deals with the process of language production and its appropriate use in context situations. Strategic competence refers to the metacognitive strategies or higher order thinking and communication strategies which are hierarchically employed by language users. Background knowledge is the central issue that marks distinctive characteristics of LSP ability. Individuals relate this long-term memory knowledge, which is based on previous experience, with the present input to predict upcoming events and to make a decision. In the testing context, test-takers retrieve pertinent background knowledge and relate it with language knowledge to interpret the communicative situation and to respond to the test tasks that resemble the target language use situation. Communication strategies serve as the mediator to facilitate the interaction between these language ability components.

Only a limited number of studies have investigated the inclusion of strategic competence in the LSP construct as indicated by Elder (2001). The author reported the mismatch of the test-takers' LSP communicative ability between linguistic competence and non-language ability, such as strategic competence and teaching skills. Elder proposed that linguistic ability should be separated from non-language competence. This approach of separating these factors could be employed in situations when assessing the sensitivity of contextual situational English usage and can be substituted for inadequate linguistic competence. From the problematic findings, strategic competence was not included in the test construct in this study.

In addition, the inclusion of field-specific content knowledge was also questioned by Wu and Stansfield (2001) in

the Listening Summary Translation Exam in Taiwanese (LSTE/T), which viewed information in a law-oriented context. Instead, they proposed that language in test tasks is the key feature that creates specificity of the test along with the authenticity of the test tasks. Apart from the controversial issue concerning the inclusion of background knowledge in test construct, a number of studies have investigated the effects of this field-specific knowledge on language test performance. However, the findings of these studies have been inconclusive.

Clapham (1996, cited in Douglas, 2000) studied ten reading-sub tests performance of the International English Language Testing System (IELTS) with three proficiency groups. She found that the test-takers achieved higher scores on the reading test in their own subject areas than in general topics. The finding also revealed that there was a highly significant effect of subject area knowledge of the test takers when they scored more than 60% on the grammar test. In contrast, test takers with scores lower than 60% did not benefit from their background knowledge. Therefore, the level of language knowledge, particularly concerning grammar, influences the effect of background knowledge on test performance.

Krekeler (2006), however, reported contradictory results to the findings of Clapham (1996). More than 500 subjects participated in this study. Two discipline-related business and technical texts were selected and C-test scores were used as a measure of L2 proficiency. In general, there was a strong effect of background knowledge on reading performance. The findings revealed that the test takers performed better on topics related to their own discipline regardless of their L2 proficiency levels but that the interaction effect between background knowledge and L2 proficiency levels was limited. The majority of test takers were able to take advantage of their background knowledge. Krekeler's findings concerning technical-related texts contradicted those of

Clapham (1996) in that the medium-level test-takers profited least from their background knowledge. With this being a prominent feature of Douglas's (2000) LSP test, background knowledge was included in the web-based speaking test in English for Tourism (WBST-EFT) construct as the content knowledge component.

Technology-integrated speaking test

Due to the number of technological advantages concerning test construction and administration, web-based language tests (WBT) have been increasingly used in testing contexts (Roever, 2001). Web-based language tests share similar features to computer-based language tests (CBT). Two advantages of the WBT testing approach over a CBT testing approach are its flexibility and convenience in test administration (Roever, 2001). According to Hamilton, Klein and Lorie (2000) and Roever (2001), the WBT has gained status in the assessment context based on its very user-friendly approach. This is particularly true regarding a low-tech approach that does not require expertise in programming and sophisticated hardware and software. Internet technology also allows test developers to create interactive, semi-direct speaking tests due to the availability of free software and the capability to post tests online for free; hence, the test is considered to be cost-effective. The low-tech WBT is an integral part of this study due to its practicality in being a user-friendly program, which is less dependent on technological expertise, and the reduced financial concerns in test development.

A few studies have claimed that the WBT is suitable for low-stakes assessment (Chapelle & Douglas, 2006; Roever, 2001), particularly for self-assessment. One example of a WBT is the self-diagnostic Dialang standardized test, but it has limitations due to limited item security and the potential for cheating. However, Garcia Laborda (2007a) wrote on the use of

the World Wide Web platform on standardized high-stakes tests. He projected that numerous standardized technology-based tests will eventually be available online and will include speaking skill evaluation. Likewise, Hamilton, Klein and Lorie (2000) discussed the feasibility of using the WBT for large-scale standardized tests due to the numerous related technological advantages, such as being inexpensive, its rapid scoring capability, the central storage of item banks, and less dependence on sophisticated software and hardware. All these attributes of the WBT makes it suitable for large-scale testing projects. One of the most prominent standardized tests in the field is the Test of English as a Foreign Language (TOEFL iBT), which incorporates online technology in test delivery and administration, specifically in the speaking section (Alderson, 2009).

Although these previous studies reported on the use of Internet technology in a number of skill areas, its use in assessing speaking ability is limited, as claimed by Garcia Laborda (2007a), particularly in the LSP context. In addition, the inclusion of multimedia in the test task presentation requires a particular framework to avoid "...the threat of interface-related construct-irrelevant variance in test scores" (Fulcher, 2003a, p. 384). For this reason, the interface design framework in the technology-based language test as proposed by Fulcher (2003a) was utilized in the development of the test for this study.

The effect of task types on speaking ability

One of the key factors affecting language performance is test tasks as referred to by Bachman and Palmer (1996). The effect of test tasks on language performance has attracted increasing interest from researchers in the investigation of testing context, including speaking tests. This topic has been investigated by a number of research studies (Kim, 2009; Teng,

2008; Turner & Upsher, 1995); however, their findings have been inconclusive.

Teng (2008) empirically explored the effect of three task types on EFL speaking performance, with 30 subjects from Taiwanese universities. The three task types (answering questions, describing pictures, and the presentation of information) were investigated to determine their effect on spoken discourse regarding accuracy, complexity, and fluency. The author found that there was no difference in performance across the three task types. However, a significant main effect was found in complexity and fluency in different task types, particularly those focusing on the answering questions task. In addition, task types and context effects via computerized test on second language speaking ability were investigated by Kim (2009). The participants were 162 adult learners of English as a second language at Teacher College, Columbia University. The test-takers' performances were investigated on grammatical competence, discourse, sociolinguistic competence, intelligibility, meaningfulness, and task completion. In terms of data analysis, multivariate generalizability theory (G-theory) and confirmatory factor analysis (CFA) were used. The findings indicated that the test-takers' performances were likely to change according to the context and task types; however, the effects of the two factors on the performances varied. To be precise, the mean differences across tasks were not large. The means of grammatical competence and intelligibility remained more or less stable across different domains and task types. The small effect of task types was found on some speaking components, sociolinguistic competence, and task completion.

In contrast, a significant effect of task types was found in the study of Turner and Upshur (1995). The authors investigated the effect of task types on the relation of communicative efficiency (CE) and grammatical accuracy (GA) in a direct-

speaking test. The two task types (single-sentence creation task and the story retell task) were employed with 130 subjects from elementary schools in Montreal, Canada. The finding revealed significant differences on the relation of CE and GA were found in the two tasks. For the short utterance task, the relation was linear, while a nonlinear relationship existed in the story retell task in which CE exceeded GA. The amount of speech produced was considered the main difference between the two tasks, and had implications on the comprehensibility of participants' speech.

Due to the practical need for a good-quality assessment tool and inconclusive findings of the effects of proficiency levels and task types on the LSP speaking test performances, this present study aims to answer the following research questions.

1. Can the WBST-EFT assess the students' LSP speaking abilities in the English for Tourism course at the University?

2. What are the similarities and differences in English for Tourism speaking abilities of high- and low-ability students in terms of their language knowledge and content knowledge performances in doing the three task types of the WBST-EFT?

Methodology

The population was 230 third year students at Nakhon Ratchasima Rajabhat University (NRRU) who took the English for Tourism II course in the second semester of 2010 academic year. From this population, 120 students were randomly selected to participate in this study. The students' course grades for English for Tourism I (EFT I) were utilized for classification of the students into two ability groups. The mean score of the EFT I course grade was 67.50, and the standard deviation was 10.16. The Z value was calculated to categorize the participants into the high- and low-ability groups. The 60 participants who had the highest Z scores were assigned to the high-ability group (Z = 0.3

to 1.5); and the 60 participants with the lowest Z scores were categorized as the low- ability group (Z = - 0.5 to -1.7). These two ability groups were further divided into three sub-groups. Each sub-group consisted of 20 participants and they were randomly assigned to three task types. There were six sub-groups in total. The present study employed the use of both quantitative and qualitative analysis techniques of the data. The research design was a 2*3 factorial design with stratified sampling. Content analysis was employed in the assessment of speech performances to investigate the similarities and differences of each LSP speaking component between the two ability groups and the three task types. The following figure illustrates the group assignment and research design.

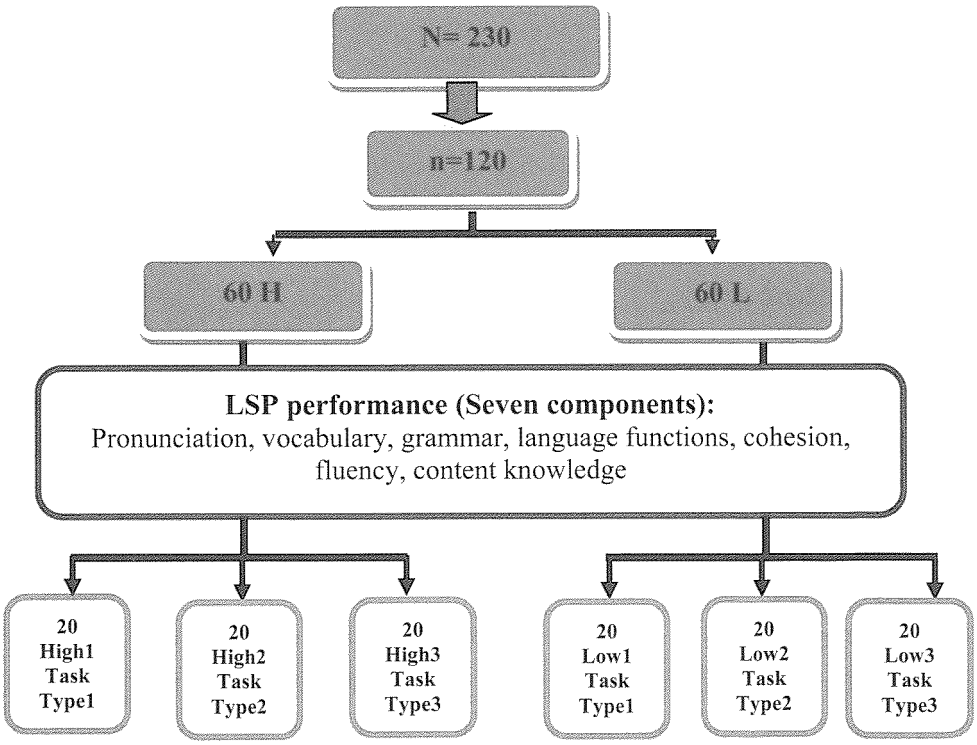

Figure 1: Group assignment with 2*3 factorial design using stratified sampling

Research instruments

The research instruments were a needs analysis questionnaire and a web-based speaking test in English for Tourism (WBST-EFT).

Needs analysis questionnaire

A needs analysis questionnaire was developed and administered to subject specialists to investigate the LSP target language use in tasks and situations, language knowledge required for professional tour guides, and the criteria for assessing language knowledge. The questionnaire comprised four sections. Douglas (2000) suggested that a needs analysis questionnaire allows the researcher to gain insight from experts in the field to include all the specific characteristics of LSP in terms of tasks, situations, and features in language production. The questionnaire was given to 15 subject specialists who each had a minimum of seven years of experience in the field of tourism and had been an instructor of English for Tourism. Results from the questionnaire were used in the selection and classification of test tasks.

The web-based speaking test in English for Tourism (WBST-EFT) and the rating scale

The WBST-EFT was developed against the theoretical framework of the LSP test development proposed by Douglas (2000) which had been modified from the framework of Bachman and Palmer (1996). It was also in line with the interface design framework for Fulcher' s (2003a) technology-integrated test. The WBST-EFT is the LSP semi-direct final achievement speaking test for the English for Tourism II course at Nakhon Ratchasima Rajabhat University (NRRU). The test task contents were based on the results of the needs analysis questionnaire. Subject specialists suggested including all the specific features of an LSP

situation in the test tasks so that these test tasks authentically share significant features of the real world tasks (Douglas, 2000, p.2). Consequently, this representation of the test performance is likely to be similar to that of the actual performances in the target situation. Some components of the task contents were based on the analysis of the English for Tourism II course syllabus.

After the classification and selection of test tasks, the blueprint of the test was drafted. There were three sections in the WBST-EFT, which were categorized by task types based on the situations that were encountered, and language functions that were performed in the professional tour guides' career. In **Task type one**, describing of attractions tasks were pertinent to the language function in expressing tourism content knowledge. For **Task type two**, giving polite suggestions to the tourists' tasks were related to the language functions used for making requests and suggestions, greeting people, and apologizing. In this study, this task type focused mainly on making suggestions to the tourists, particularly on Thai etiquettes. For **Task type three**, the language functions were dealing with tourists' enquiries and complaints tasks. Each task type was made up of two sub-tasks and there were six sub-tasks in total. The details and objectives of the test tasks are presented below.

Section 1 (**Task type one**), presenting tourism-related information, aimed to elicit the students' ability in presenting national tourist attractions and explaining the tour program. This was accomplished through the utilization of two tasks. For Task One, presenting tourist attractions, the students were required to present two of the most famous national attractions in Thailand: the Emerald Buddha Temple and the Grand Palace. They were provided with seven pictures from 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. For Task Two,

describing a one-day tour program in the central region of Thailand, the students were asked to first read the one-day tour itinerary; they were then required to present the information to tourists by providing the additional details of the specified attractions.

In Section 2 (**Task type two**), giving polite suggestions to tourists, the objective was to assess the students' ability in giving polite suggestions to tourists in two different situations. Task Three involved the Summer Palace, in which the students were asked to watch a video clip containing a monologue of the tour guide at the Summer Palace. Then, there were six pictures included in the clip which required the students 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 Four which was similar to Task Three, required the students to watch a video clip of Jatujak Market, and they were then asked to respond to the six pictures containing different scenes by giving polite suggestions regarding what the tourists should do while at the crowded shopping center.

In Section 3 (**Task type three**), dealing with enquiries and complaints, this task type emphasizes the students' ability to deal with tourists' enquiries and complaints on a variety of topics. In Task Five, dealing with enquiries, the students first watched a 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 students were asked to politely and appropriately respond to the enquiry. Task Six, dealing with complaints, incorporated three complaints: an incomplete tour program, an unrequested hotel room, and prolonged wait for a bus. The students attempted this task in a similar way as they did in Task Five by first watching video clips containing different complaints, and then, they were required to

politely and appropriately respond to each complaint (See Appendix A for the WBST-EFT).

Then, all tasks were posted on the Moodle Version 1.9.5 which is a freeware online template and program that is currently used at NRRU. The test was administered in the computer lab during the second semester of the 2010 academic year. At this stage, before the rating scale was developed, the construct to be measured was obtained from the course objectives analysis using the specific purpose language ability framework of Douglas (2000) and speaking ability from Fulcher (2003b).

An analytic rating scale was used in this study due to its appropriateness for the purpose of the test, and it allowed for assessing specific components of language ability. In terms of criteria for correctness, accuracy in linguistic elements, range, complexity, and appropriateness of speech production were used. The rating scale consisted of seven components: knowledge of pronunciation, knowledge of vocabulary, knowledge of grammar, knowledge of language function, knowledge of cohesion, fluency, and content knowledge. Detailed descriptions and specific assessment criteria for each component are presented in the content analysis part. There were five ability bands for each component ranging from band level 0 (a very poor user), 1 (beginner), 2 (a fair user), 3 (a good user) to 4 (a very good user). The ability band level was obtained from the summation of the averaged scores from the two raters in each test task (See Appendix B for the construct definition in the rating scale).

To address scoring validity and reliability concerns, rater training was arranged before the pilot study to assure consistency in the rating. *A priori* validity evidence was obtained from the three experts in the field using the index of item-objective congruence. The results indicated that each test task received the maximum score of one, indicating high content and construct validity. Then, the instruments were pilot-tested with

30 similar subjects and revisions were made before being used in the main study. After the pilot study, the *posteriori* validity evidence on scoring validity and reliability was also investigated in the following steps. First, two raters tried out the rating scale with ten sample speeches. In the case of any discrepancy in the band score in both the pilot and main study for more than one band level, a further discussion was held to arrive at mutual agreement. Then, the inter-rater reliability was calculated with 30 test performances. A Pearson correlation coefficient was applied to assess the inter-rater reliability; the value was .99, indicating that the raters were highly consistent in their rating. Moreover, the *posteriori* evidence on the item discrimination index showed that the values ranged from .58 to .63 for the six tasks, showing that the test could effectively classify the mastery levels of the students' LSP speaking ability. The difficulty values of the test tasks ranged from .28 to .35, indicating that the test was quite difficult. To assure the reliability of the test scores in the main study, the inter-rater reliability was calculated with 120 test performances and replicated in similar steps, as in the pilot study. The results indicated that the correlation coefficient value was .85. Thus, the statistical evidence reflected a high reliability of the rating scores, and it could be claimed that raters were highly consistent in their rating procedures.

Procedures

After the pilot study, the data collection procedure in the main study was conducted and it replicated the stages of the pilot study. The revised versions of the WBST-EFT and the rating scale were administered to 120 third-year students from Nakhon Ratchasima Rajabhat University in the second semester of the 2010 academic year. The students were classified into six sub-groups, and each group was assigned to do different task types. Their performances were audio recorded and stored in the

database to be rated later by two experienced raters. The two raters first used the rating scale with ten sample speeches before they started rating the actual test responses. When there was any discrepancy of the band score between the two raters, a discussion was held to come to a similar agreement based on the scale. Some elements of the descriptors were revised before being used in the main study.

Data Analysis

To answer the first research question on whether the WBST-EFT can assess LSP speaking abilities of the students in the English for Tourism course at the University, *a priori* and *posteriori* validation procedures were performed. To obtain the *a priori* validity evidence on content and construct, three experts in the fields were consulted, and the results indicated high content and construct validity of this instrument. After the test administration, the *posteriori* validity evidence on the scoring validity and reliability of the test was established. To ensure the reliability of the scoring method, rater training was performed. Then, the scores from 120 test performances were analyzed with the Pearson correlation coefficient to measure the inter-rater reliability. The statistical results showed high reliability of the two raters.

To answer the second research question on the similarities and differences in the test performances between the two ability groups and the three task types with the Internet assessment, the scores from the speech performances were computed by two-way ANOVA to check for significant differences among the mean scores of the two ability groups on the three task types. This was done by using the Statistical Package for Social Sciences (SPSS), version 13.5. If the test value was significant, then there exists at least one significant difference between group means. Then, a post-hoc Scheffé test was performed to indicate the significance

of the particular contrast. In addition, content analysis of the 120 speaking performances was conducted to further investigate whether each LSP component would be different between the proficiency levels and among the three task types. The audio-recorded responses were transcribed to find the similarities and differences from each speaking component. Then, these features were categorized by the proficiency levels of the students and the task types.

Findings

The results of the students' LSP test performances measured by the WBST-EFT

Table 1 shows the means and standard deviations of the total band scores of the WBST-EFT.

Table 1: Descriptive statistics of the WBST-EFT total band scores

	Total band scores	Mean	SD
High (n=60)	5	2.63	.50
Low (n=60)	5	1.97	.47

The above table shows that the mean score of the high ability group ($\bar{x}_H$ = 2.63, SD = .50) is greater than that of the low ability group ($\bar{x}_L$ = 1.97, SD = .47) although the variation in the scores within each group is not large, ranging from .47 to .50. From these statistical results, it can be seen that proficiency levels affected the difference in LSP performances of the two ability groups in that the high ability group performed better than the low ability group.

More specifically, the results from the two-way ANOVA indicated that only proficiency levels had a significant effect on

the two different ability groups' total scores, F (1,114) = 55.02, p < .05. However, the statistical result did not show any significant effect of the three task types on the performances of the two ability groups, F (2, 114) = 2.53, p > .05. Similarly, there was no significant interaction effect between the two proficiency levels and the three task types on the total scores, F (2, 114) = .12, p > .05. In other words, the students from different proficiency levels show significantly different LSP speaking performances. However, their performances are stable across the three task types in that the high ability group's scores are constantly high whereas the low ability group's scores are stably low.

Concerning the LSP individual components, the statistical results showed the main effect of the proficiency levels on all of the performances of the LSP individual components of the two ability groups. This means that the LSP individual components' performances of the two ability groups are significantly different: pronunciation [F (1, 114) = 31.42, p < .05], vocabulary [F (1, 114) = 46.66, p < .05], grammar [F (1, 114) = 46.70, p < .05], language functions [F (1, 114) = 42.90, p < .05], cohesion [F (1,114) = 32.36, p < .05], fluency [F (1, 114) = 43.79, p < .05] and content knowledge [F (1,114) = 75.99, p < .05]. The following table shows the means and standard deviations of LSP individual components of the two ability groups.

Table 2: Descriptive statistics of the LSP individual components of the two ability groups

LSP individual components	**Proficiency levels**			
	High (n=60)		**Low (n=60)**	
	Mean	SD	Mean	SD
Content knowledge	2.54	.59	1.75	.57
Vocabulary	2.66	.56	1.96	.59
Grammar	2.66	.57	1.97	.54
Language functions	2.77	.55	2.10	.59
Fluency	2.60	.55	1.96	.52
Pronunciation	2.57	.51	2.01	.56
Cohesion	2.58	.55	2.03	.49
Total	**2.63**	**.50**	**1.97**	**.47**

From Table 2, the most to the least mean differences between the individual components of the two ability groups are as follows: content knowledge ($\bar{x}_H$ = 2.54, SD = .59, $\bar{x}_L$ = 1.75, SD = .57), vocabulary ($\bar{x}_H$ = 2.66, SD = .56, $\bar{x}_L$ = 1.96, SD = .59), grammar ($\bar{x}_H$ = 2.66, SD = .57, $\bar{x}_L$ = 1.97, SD = .54), language functions ($\bar{x}_H$ = 2.77, SD = .55, $\bar{x}_L$ = 2.10, SD = .59), fluency ($\bar{x}_H$ = 2.60, SD = .55, $\bar{x}_L$ = 1.96, SD = .52), pronunciation ($\bar{x}_H$ = 2.57, SD = .51, $\bar{x}_L$ = 2.01, SD = .56) and cohesion ($\bar{x}_H$ = 2.58, SD = .55, $\bar{x}_L$ = 2.03, SD = .49), respectively. Their standard deviations do not differ much and range from .49 to .59. Among the seven components, the mean scores differing most between the two ability groups occur in the content knowledge, the most prominent feature of an LSP. The mean scores of the high ability group's content knowledge are almost twice as many as those of the low ability group whereas the least mean difference is found in cohesion scores, indicating that their cohesion performances do not differ much.

Content analysis on LSP speech performances

Since there was a significant difference between the mean scores of the high and low ability groups, content analysis was conducted to investigate both the similarities and differences in each LSP speaking component of the test performances. The analysis also showed the in-depth information and the prominent features in some of the LSP speaking components associated with the different ability groups and a particular task type. The information in the brackets was added to clarify the responses.

Pronunciation is the first linguistic component of language knowledge. The investigation was concerned with accuracy in pronouncing words and the use of stress and intonation in the performances. The two ability groups similarly mispronounced whole words, endings of words, and consonant clusters in their responses. They also used the wrong emphasis with certain words. These errors were found across the three task types; however, these errors were mainly made by the low ability group. Additionally, errors regarding intonation in their responses were exclusively found in the low ability group. Concerning intonation error, the low ability students responded to the test tasks in monotonous speech as if they were reading the scripts. The following excerpts illustrates the errors made by the low ability group in Task type one. The italicized word represents the error; the correct pronunciation is in the brackets, and the grammatical errors are not corrected.

Incorrect pronunciation of word and wrong stress:

> ...the top floor of the eastern wing are kept *'reallycious* [religious /rɪ'lɪdʒ.əs/] objects...

Incorrect pronunciation of ending and wrong stress:

> Emerald Buddha *i'made* [image /'ɪm.ɪdʒ/], the ordination hall and the gallery...

Incorrect pronunciation of consonant cluster:

It is *co'ntruct* [construct /kə'nstrʌkt/] in 1782...

The investigated features concerning vocabulary were two types of words: generic and tourism technical terms. This component was measured by the accuracy and range in the responses. Range of vocabulary was measured by the number of words per response. The analysis showed that the two ability groups similarly employed both technical and generic terms in their speech. Tourism-related technical terms, particularly about Thai history and architectural structures, were mostly found in Task type one and some of them were limitedly used in Task type two. In Task type one, the students were required to explain about the attractions of Thai architecture, arts, history, and Buddhism. For Task type two, they were required to give suggestions to the tourists on *do's* and *don'ts*, which were related to Thai etiquettes. This information was associated with technical terms. However, most of the generic terms were found in Task type three where the students were asked to resolve problems in an organized trip which was related to general information. The following excerpts demonstrate the technical terms in italics used by the two ability groups in Task type one.

High ability group: This is the *Emerald Buddha Temple*. It was built in the *reign* of *King Rama* the first in seventeen-eighty two. The *Emerald Buddha Temple* was very important because Thai people believed that it was the most *sacred* place[s] in Thailand and the *repository* of *spirits* for all Thai people.

Low ability group: This is the *Emerade [Emerald] Buddha Temple*. The *temple* is very important because there are many interesting things to see inside.

The two ability groups also made similar errors on the use of near synonym words that made it hard to understand the students, and this error was associated with Task types two and three. However, it was also mainly found in the low ability group's responses. Another difference was on the range of vocabulary utilized by each group. The high ability group used a wider range of vocabulary in their responses than the low ability group did across the three task types, and the most salient different was in Task type one. The reason may be that this task type required the largest amount of information among the three task types. The differences in range of the vocabulary between the two ability groups are illustrated in the above examples.

Grammar competence was investigated on the accuracy, range, and complexity of the structures in the responses, particularly concerning the use of the tenses and types of sentences. The range of the structures was measured based on the number of types of sentences per response. The two ability groups similarly used present simple and future tenses across the three task types. However, past simple tense, particularly the passive voice, was mainly found in Task type one in the responses of the two ability groups. For types of sentences, simple and compound constructions were primarily found in Task types one and three, whereas complex was mainly used in Task type two.

On the contrary, the salient grammatical difference in the responses of the two ability groups was seen in the range and complexity of the structures. The high ability group used more types and more complex sentences than the low ability group did

across the three task types, particularly in compound-complex sentences, which was only found in the high ability group's responses. Among the three task types, these differences were noted in the last task type in that four sentence types were found in the high ability group's responses whereas two types were used by the low ability group. The following excerpts from the two ability groups illustrate the example of each type of sentence in Task type three.

High ability group:

Simple sentence: "I will send someone for giving some medicine for your son right now."

Compound sentence: "Okey uh I will introduce to, uh, tour program for today and [pause] we will see the sunset on the, uhm, behind the temple.

Complex sentence: And if your son is not better, can you call me back?"

Compound-complex sentence: "Uhm, Bangkok also offer[s] the best kind of food on the planet [pause] and when you travel in Bangkok you must see some activity in Bangkok, uhm, such as Khao San Road, Paragon, uh, movie at cinema, theater, floating market."

Low ability group:

Simple sentence: "I will shenk[change] your new room for you."

Compound sentence: "I will call the driver right now and it will never happen again."

Another difference, which was similar to the previous LSP component, was regarding the errors that were mainly found in the low ability group's performances across the three task types. Both ability groups made errors of no verb and wrong use of verb form in the sentences across the three task types. Additionally, errors that were prominent in particular task types were found; and among the three task types, Task type one contained the greatest variation in grammatical errors. These errors included wrong use of preposition, wrong pronoun, and no noun in the sentences, and this may be related to the amount of information required in this task type when compared with the other two task types. In Task type two, errors made concerned the wrong use of an adjective instead of an adverb and the wrong use of a preposition, while the inaccurate use of pronoun and the infinitive 'be' were noted in Task type three.

Language functions were investigated on the appropriate use of the three types of language functions: to present tourism-related information, to give polite suggestions to the tourists, and to deal with tourists' enquiries and complaints. The two ability groups similarly explained to the tourists the attractions in Bangkok and the tour itinerary in Task type one. For Task type two, they gave polite suggestions to the tourists; and in Task type three, they responded to tourists' enquiries and complaints. The examples of the use of language functions are presented in the following excerpts from the high ability group.

Task type one: To explain the tourists' attractions in Bangkok

This is the Grand Palace. It was built in seventeen-eighty two by King Rama the first. [The] original living quarters were temporally made of wood and thatch...

Task type two: To give polite suggestions to the tourists at the ritual site
Please do not climb the Buddha image because we should pay respect [to] the Buddha image.

Task type three: To deal with tourists' complaints about the changed tour program
Certainly, that's no problem. If you want we will go to visit and shopping at the floating market.

In contrast, the difference in this LSP component was the inappropriate use of language function which was mainly found in the low ability group in Task type two. This mistake was related to the task requirement that aimed to measure students' ability to give polite suggestion to the tourists. The two ability groups made this mistake by using the direct command 'Do not' with the tourists, who were both the audience and the customers. As part of the construct on using the polite language, 'Do not' would be considered impolite and inappropriate.

Cohesion was investigated on the types (connectors, relative pronouns and time sequence markers) and number of cohesive markers per response, and was measured by the accuracy and range of cohesive markers. From the three task types, the high ability group and the low ability group employed similar types of cohesive markers: connectors, relative pronouns, and time sequence makers. It should be noted that time sequence markers were only found in Task type one to explain about the sequence of the tour program. In Task type two, the connector 'because' was primarily used by the two ability groups to explain Thai etiquettes at the religious site and what the tourists should do at the crowded attractions. In Task type three,

'and' and 'because' were mainly used by the two ability groups to respond to the tourists' enquiries and complaints in Task type three. However, cohesive markers 'if' and 'that' were only found in the responses of the high ability group.

On the contrary, the salient difference between the two ability groups was in the range of the cohesive devices of the responses that the high ability group used. The high ability group employed twice as many connectors as the low ability group in Task types one and three. The examples from the two ability groups in Task type one are presented as follows. The bold font represents cohesive markers.

High ability group: **First**, at eight o'clock we will depart from the Grand Hotel Bangkok. **Next,** at nine-thirty we will arrive at Nakhon Pathom and visit the Golden Pagoda **and** pay respect to the sacred Buddha image. **After that**, have lunch at Ban Ruen Thai restaurant. **Then,** we will visit Sanam [pause] chandra Palace at thirteen-forty five. And **last** [lastly] we will depart from Nakhon Pathom at [pause] fifteen o'clock.

Low ability group: **The first** in [at] egg [eight] o'clock depart from the Grand Hotel Bangkok. [pause] Nine [pause] thirty arrive at Nakhon Pathom [pause] visit [pause] the Golden Pagoda **and** pay respect to the secard [sacred] Buddha imade [image].

From the two excerpts, it was noted that the high ability group used six cohesive markers in the response whereas only two cohesive markers were found in the low ability group. Additionally, the errors on limited or no use of cohesive markers that resulted in unconnected speech were mainly found in the low group's responses, as shown in the above example.

Fluency was investigated on the appropriate use of tempo and pauses in speech. The two ability groups similarly made prominent errors on the inappropriate use of tempo and pauses that obstructed the flow of the responses, and caused the responses to be hard to understand. However, all of the errors were mainly found in the low ability group's responses across the three task types. In each task type, the low ability students used very slow tempo in the responses with short pauses between words. They also inappropriately used long pauses in the middle and at the beginning of the responses; this caused unconnected ideas and directly affected the clarity of the responses. Another error was that the low ability group repeated the same words and phrases, which also obstructed the flow and understanding of the responses. The examples of the responses of the low ability group are illustrated in the following excerpts. The bold font represents errors in pauses and repetition, and the line represents slow tempo as the students were reading word by word.

Task type one: There _ are _one _ hundred_ seventy_egg [eight] _ the mural painting[s] at_the_ gallery. There [they]_are_locase [located] **oppo_opposite**_the_ ordination_ hall.

Task type two: Please _ you _ take _ off _ your _ shoes **[pause] becaud** [because]_**not it is to _it_ is_** ruse[rule].

Task type three: Uh_OK_will_ tow[tell]_ you_ now.
Jack_[Just] moment_ please. I[**pause**]
call [**pause**] to **[pause]** the driver now."

Content knowledge was investigated on types of content knowledge found in the responses and measured by the accuracy and completion of the information in responding to the test tasks. As part of the specific feature of an LSP test in the form of test content, the analysis showed that the two ability groups similarly reported a specific type of content knowledge associated with a particular task type. In Task type one, the content knowledge related to Thai architectural structures, particularly temples and palaces, Thai arts, Thai history of the temples and palaces, and Buddhism was utilized. In Task type two, Thai cultural knowledge on *do's* and *don'ts* at the religious site was noted. This task type also included *do's* and *don'ts* at another tourist attraction. In Task type three, content knowledge was mostly related to the problem-solving in tourism-oriented situations. The following excerpts from the high ability group illustrate a specific type of content knowledge in each task type.

Task type one: Thai history about the temple
This is the highlight of our trip today. It is called the ordination hall. It was built in the reign of King Rama the first [pause].

Task type two: Thai cultural knowledge about the etiquette at the religious site
Please don't sit on the floor of the temple with your feet pointing at the Buddha imesh[image] because it is consider[ed] as highly impolite in Thai culture.

Task type three: Dealing with tourist enquiries

> Uhm, Bangkok also offer the best kind of food on the planet [pause] and when you travel in Bangkok you must see some activity [activities] in Bangkok, uhm, such as Khao San Road, Paragon, uh, movie[s] at cinema , theater, floating market.

In contrast, the difference in errors relating to content knowledge was mostly found in the low ability group across the three task types. Incorrect information was found in Task type one, particularly with the numerical information concerning the size of the Emerald Buddha image and the year in which the attraction was constructed. In Task type two, wrong information was related to the Thai etiquette at the ritual site in that the students could not articulate the correct reason concerning Thai beliefs towards Buddhism. One of the low ability students said, "You should not take a photo because it's peach copyright". The correct reason should be "You should not take a photo because it is considered to be disrespectful in Buddhism beliefs". In Task type three, the wrong information was given on the recommendation of an attraction when dealing with tourists' enquiries; the group was supposed to recommend an attraction in Bangkok, but they incorrectly suggested the attractions in their hometown. The incomplete information in the low ability group was found in relation to the amount of the responses and the task requirement. They were unable to fulfill the task requirement and did not show their ability in providing the content knowledge. This error was salient in Task type two because one low ability student said, "Take the bag on the bud [bus]"; and in Task type three, "I'm sorry very much [very sorry]. I will do it batter [better]".

Discussion

The findings indicated that the WBST-EFT had high scoring regarding validity and reliability based on the statistical evidence. The significant difference in the total means of the two ability groups showed that this assessment tool could effectively differentiate the mastery level of the LSP speaking performances between the two ability groups. The results of all of the individual components' means between the two ability groups were significantly different, reconfirming that the test is a good assessment tool that can discriminate between the ability levels in each component. Therefore, this innovative test should be used in Nakhon Ratchasima Rajabhat University to assess LSP speaking performance in the tourism context with a large number of students. Additionally, the WBST-EFT was constructed based on a particular theoretical framework from the concepts of the LSP test and the WBT using technological advantages, as mentioned by Garcia Laborda (2007a), Hamilton, Klein and Lorie (2000), and Roever (2001). For this reason, this framework can be employed in other LSP speaking tests to assess language production with a vast number of test takers.

Interestingly, the analysis of the data showed prominent characteristics of this LSP test. The similarities were found in most of the LSP components from the responses of the two ability groups, and these features were prominent in a particular task type. Key features were the specific types of words, tenses, sentence constructions, language functions, cohesion and content knowledge. These distinctive features in the responses of the two ability groups may come from the specific purpose input in test task characteristics in the form of test content and task requirement that resulted from target language use analysis and the views of subject specialists (Douglas, 2000). For these reasons, the findings of this study correspond to Douglas's (2000) theory on the specific characteristics of LSP tests. Therefore,

these prominent features should be considered in the construction of other LSP speaking tests with the integration of Internet technology.

Additionally, the analyses reveal an insight into the similarities and differences of the LSP speaking performances among two different ability groups in attempting the three task types. The findings indicated that the students with different proficiency levels had significant and different LSP speaking ability, particularly in the context of the specific test tasks with the technology-integrated assessment. Nevertheless, the performances of the two ability groups across the three task types were not significantly differed. The high ability students had significantly higher scores than their counterparts across the different task types. The findings correspond to the study of Teng (2008), who investigated the effect of three task types on EFL speaking performance in terms of accuracy, complexity, and fluency and found no variation in the test performances across task types. The indifferent performances across tasks may be that the students mainly relied on their target language knowledge to attempt the LSP test tasks as the high ability students' total mean scores of all the LSP components were relatively high across the three task types whereas those of the low ability students were consistently low. More specifically, the greatest means difference of the individual components lies in content knowledge, the integral feature of an LSP test, indicating that the students manipulated their content knowledge ability in relation to their proficiency levels. Thus, these findings provide important information for teachers in both the LSP instructional approach design and LSP curriculum development. The teachers should know how to teach language in specific content classes - that is, they should be able to identify the specific language features and functions used in particular content topics. In addition, they should emphasize specific language objectives and

functions in specific content topics. They should also know how to present these language features in class as part of the course content. Moreover, they should systematically and directly teach these LSP components used in particular content topics and functions to the students.

Finally, the content analysis showed that the difference in proficiency levels strongly affected the LSP production. The high ability group used a wider range of vocabulary and cohesive markers, more types of and complex grammatical structures and more appropriate language functions than the low ability group did. Nonetheless, most of the errors in all components were mainly made by the low ability group and some were exclusively found in the responses of this group of students. These findings on the similarities, differences, and prominent features of the LSP individual speaking components among the two ability groups offer significant information that should be considered in this LSP course content and instructional approach design, and future LSP curriculum development in the Thai tourism domain.

Limitations of the study

The limitations of the present study concerned the small sample size and the limited coverage of the task types that were used by tour guide professionals in Thailand. The task types were selected in relation to the final achievement test content in the English for Tourism course; hence, the test may not have covered all of the actual tasks in the tour guide context. For these reasons, the inference from the test scores must be applied with caution to actual performances in the tourism context; and the findings may not be generalized to other tourism domains beyond Thailand.

Conclusion

This study examines the quality of an Internet-integrated test in terms of validity and reliability in assessing the LSP speaking abilities of the students at a Thai university, and investigates whether this test can discriminate between differences in students' abilities in three task types in the Thai tourism context. The findings showed that the WBST-EFT was an effective assessment tool, constructed upon a particular theoretical framework, and that it had high content and construct validity. The statistical evidence indicated the high reliability of this instrument, and it also had high practicality. Moreover, the findings revealed that students from different proficiency levels had significantly different English for Tourism speaking production in terms of range, accuracy, complexity and appropriateness. From the statistical results and qualitative findings, it can be concluded that the WBST-EFT can discriminate between students' LSP speaking performances. Thus, this instrument should be employed in the University to assess students' LSP speaking abilities. Additionally, due to the high quality and practicality of the WBST-EFT, the framework of this instrument should be used in the development of other LSP tests to assess the speaking ability of a large number of test-takers.

Acknowledgements

The authors would like to thank the Office of the Higher Education Commission, Thailand according to the Strategic Network Project for Production and Development of High Educational Institution Instructors in the Domestic Doctoral Degree Program for the financial support in conducting this study.

The Authors

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

Kanchana Prapphal is a Professor Emeritus at the Chulalongkorn University Language Institute. Her wide range of publications is in the area of language teaching and testing.

References

Alderson, J. C. (2009). Test review: Test of English as a Foreign Language TM: Internet-based test (TOEFL iBT). *Language Testing, 26*(4), 621-631.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology.* Cambridge: Cambridge University Press.

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

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

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). On the net: Introducing standardized EFL/ESL exams. *Language Learning and Technology, 11*(2), 3-9.

Garcia Laborda, J. (2007b). From Fulcher to PLEVALEX: Issues in interface design, validity and reliability in Internet-based language testing. *CALL-EJ Online* 9(1). Retrieve from:http://www.tell.is.ritsumei.ac.jp/callejonline/journal/9-1/laborda.html

Hamilton, L. S., Klein, S. P., & Lorie, W. (2000). *Using web-based testing for large-scale assessment.* Retrieve from: http://www.rand.org/content/dam/rand/pubs/issue_papers/2005/IP196.pdf

Kim, H-J. (2009). *Investigating the effects of context and task type on second language speaking ability.* Unpublished PhD thesis, Columbia University.

Krekeler, C. (2006). Language for special 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.

Teng, H-C. (2008). A study of task type for L2 speaking assessment. In M. Mantero, P. C. Miller, & J. L. Watzke (Eds.), *ISLS readings in language studies* (pp. 433-446). St. Lois, MO: International Society for Language Studies.

Thailand Tourist Arrivals. (2012, March 2). *Thailand Tourist Arrivals from 2007 to 2011 Per Quarter.* Retrieve from: http://www.thaiwebsite.com/tourism.asp

Turner, C. E., & Upsher, J. A. (1995). Some effects of task types on the relation between communicative effectiveness and grammatical accuracy in intensive ESL classes. *TESL Canada Journal, 12*(2), 18-31.

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

Appendix A

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

Task type two (Task Four): Giving polite suggestions to the tourists at a crowded attraction

Task 4: Listen to the tour guide's talk from the video clip. There will be 6 pop up pictures in order. Give appropriate suggestions with reasons to each picture. You will have 3 MINUTES to work on this task. For each picture there will be 10 seconds for organizing your idea and the rest 20 seconds to speak.

NEXT

Appendix B
The rating scale

Language for Specific Purposes (LSP) speaking components	**Bands**				
	4	**3**	**2**	**1**	**0**
Knowledge of pronunciation is the ability to use sound, stress, and intonation to convey the intended meaning of the responses.					
Knowledge of vocabulary is the ability to use both generic and tourism related technical terms to respond to the test tasks.					
Knowledge of grammar is the ability to use standard English grammatical structures and rules to produce comprehensible responses. It includes the use of specific language patterns to construct the test responses.					
Knowledge of language functions means the ability to formulate appropriate responses with the consideration of the situations and social status of the audiences. It includes the ability to present tourism-related information, to give polite suggestions to tourists, and to respond to tourists' enquiries and complaints as presented in the test tasks.					
Knowledge of cohesion is the ability to combine phrases and sentences in a meaningful way which can be seen from the use of cohesive devices in the responses.					
Fluency is the ability to appropriately use tempo and pauses in the language production to maintain a pace of the responses.					
Content knowledge is the ability to present tourism related content knowledge taught in English for Tourism II.					